Don't Compress Away the Evidence: Preserving LLM Ownership Signals in Compressed KV Caches

Anonymous authors
Paper under double-blind review

Abstract

Efficient KV cache compression is essential for deploying long-context large language models, but it introduces an unexamined conflict with model ownership auditing. Existing fingerprinting methods embed ownership evidence through trigger patterns that are deliberately positioned away from routine task semantics—precisely the kind of patterns that cache compression discards as "unimportant". Therefore, we propose Fingerprint-Preserving Causal KV Cache Compression (FinCa), a framework that treats ownership evidence and task utility as joint constraints. Our key insight is to replace static trigger patterns with a low-rank perturbation in a KV projection that generates an input-conditioned structured signal in the runtime cache. Complementing this generation mechanism, we reserve fingerprint-bearing coordinates during compression and apply quantization-index modulation to protect them, ensuring that the ownership signal survives cache reconstruction. As a result, FinCa embeds a robust fingerprint signal from only 20 documents and achieves 100% signal recovery in most evaluated settings.

1 Introduction

The rapid proliferation of open-source large language models (LLMs) has highlighted the need for reliable model ownership verification, as publicly circulated model checkpoints may be further fine-tuned, distilled, or repackaged by unauthorized third parties. This concern is not merely hypothetical. In March 2023, just one week after Meta opened research-access applications for LLaMA-1, the complete model weights were posted on 4chan and spread rapidly (Cox, 2023). In addition, controversies surrounding similar incidents have gradually triggered extensive discussions on community platforms such as X, particularly concerning Claude, GPT, DeepSeek, and Kimi. If the weights of a proprietary model are stolen and subsequently released as an open-source model, the original developer may lose its exclusive advantage, weakening incentives to continue investing in costly training and safety engineering. Model fingerprinting can address this challenge not only as a mechanism for detecting model misuse, but also as a tool for verifying model provenance and ownership. Existing approaches embed a secret and verifiable signal into the model through parameter-space modifications (Uchida et al., 2017), output watermarks that bias generated text (Kirchenbauer et al., 2023), or trigger-based behavioral fingerprints activated by specific inputs (Xu et al., 2025c). To preserve uniqueness while minimizing

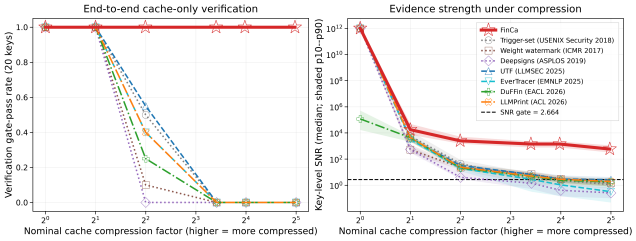


Figure 1: KV-cache compression survival of FinCa, classical controls, and four advanced LLM fingerprinting methods.

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.

interference with normal tasks, these methods commonly employ rare and semantically anomalous triggers (Zhang et al., 2026; Fu et al., 2026).

However, existing approaches generally treat the model as a terminal object whose weights and inference path remain unchanged after deployment—an assumption increasingly unrealistic for long-context LLM serving, where production systems rely on input-dependent and lossy KV-cache compression, including quantization (Liu et al., 2024; Hooper et al., 2024; He et al., 2024), low-rank error compensation (Kang et al., 2024), and importance-based token eviction (Liu et al., 2023a; Zhang et al., 2023; Xiao et al., 2024; Li et al., 2024). Because the rarity and statistical abnormality that make trigger patterns distinctive may also render them vulnerable to attention- or magnitude-based pruning, as shown in Fig 1, a fundamental question arises: how can a pre-embedded fingerprint survive aggressive KV-cache compression?

Our method, Fingerprint-Preserving Causal KV Cache Compression (FinCa), starts from a simple observation: a fingerprint does not have to be observed in the model's output. It can instead appear as an input-dependent signal in the KV cache. We add a small low-rank perturbation to a selected KV projection using secret carrier matrices and a binary symbol vector. During inference, the input activations convert this weight perturbation into a corresponding cache change. An auditor observes the fingerprint by comparing the KV caches produced by the fingerprinted and clean models on the same probe documents, and uses the private carriers to recover the symbol vector. Our probe documents do not need to contain trigger phrases or elicit target responses, so detection does not depend on output text or on specially controlled inputs. In summary, our contributions are as follows:

- We identify and validate an overlooked vulnerability of model fingerprints to inference-time KV-cache compression, and formulate fingerprint ownership preservation as a joint embedding–compression problem.
- We introduce FinCa, comprising a keyed, input-conditioned low-rank perturbation for implanting fingerprints—whose cache manifestation is predictable from ordinary probe activations and a clean reference checkpoint—and a causal, detection-constrained KV-cache codec that preserves fingerprint coordinates.
- We conduct extensive experiments on 10 open-source LLMs, and the results show that FinCa achieves 100% fingerprint signal recovery in the vast majority of severe cache-compression scenarios.

2 Audit protocol and trust boundary

The auditor holds the fingerprint key, the clean reference checkpoint, the private decoder state, and the authenticated cache traces. We use H_0 to denote clean weights and H_1 to denote weights containing the owner's fingerprint. Given a sequence of audit documents, the auditor evaluates whether the observed cache responses agree with H_0 or H_1 and reports ownership only when the pre-declared sequential test reaches its acceptance threshold. All detection parameters and stopping thresholds are fixed before the audited observations are decoded. This protocol assumes gray-box access through the trusted monitor; access through a text-only black-box API is outside the considered setting.

3 Method

Figure 2 provides an overview of FinCa. We first embed a keyed low-rank fingerprint in a selected Value projection, then preserve its cache response through detection-constrained compression. Ownership is subsequently audited by comparing reconstructed caches with a clean reference and applying a calibrated sequential test.

3.1 Embedding the Fingerprint in Model Weights

We embed a fingerprint in the Value projection of a preselected Transformer layer $\ell^\star$ and KV head $p^\star$. Let the weight slice be $W \in \mathbb{R}^{d_v \times d_h}$, where d_h is the input hidden dimension

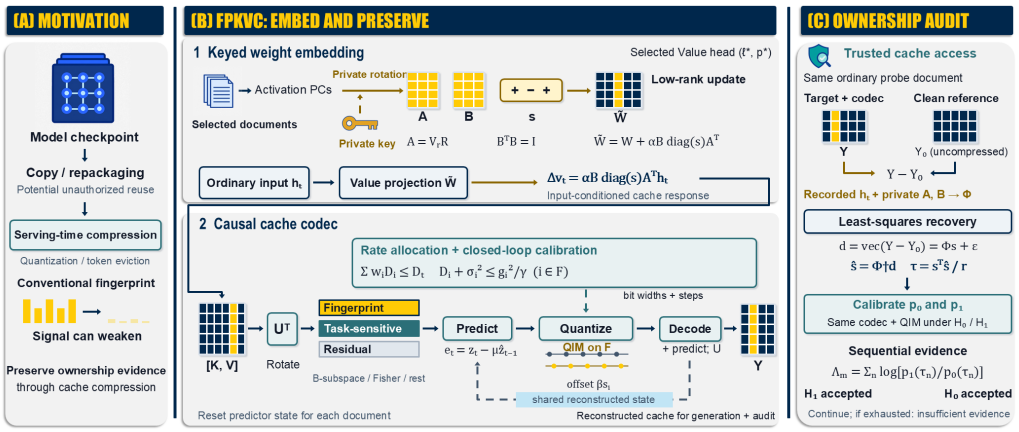


Figure 2: Overview of FinCa: keyed fingerprint embedding, compression-aware preservation, and post-deployment ownership auditing.

and d_v is the Value-head width. The input carrier $A \in \mathbb{R}^{d_h \times r}$ is obtained from the principal directions of the hidden states in selected documents and then subjected to a private-key-controlled orthogonal rotation. The same private key generates a column-orthonormal output carrier $B \in \mathbb{R}^{d_v \times r}$ and a key-generated binary fingerprint vector $s \in \{-1, +1\}^r$, where r is the number of symbols. Given a strength $\alpha > 0$, the embedded weight is

$$\widetilde{W} = W + \alpha B \, \mathrm{diag}(s) A^\top, \tag{1}$$

where $\mathrm{diag}(s)$ is the diagonal matrix of signs. The quantity $A^\top h_t$ gives the projection of the current input onto the selected directions. These values are sign-modulated by the fingerprint components s_j, and B maps the result to a perturbation of the Value vector. Carrier construction and activation responses are given in Appendix C.1. The embedding location, number of symbols, and strength are fixed before auditing; the subsequent compression process separately protects the directions containing these responses.

3.2 Fingerprint Preservation During Compression

We compress the concatenated Key and Value rows for every token, layer, and KV head. In the target head, an orthogonal transform separates the fingerprint directions spanned by the output carrier B from the remaining cache content. Within the remaining directions, we select q task-sensitive directions according to their Fisher sensitivities and complete the basis with the residual directions. The encoder predicts the current coordinates from the reconstructed coordinates at the previous time step and quantizes only the prediction residuals; the scalar causal predictor coefficient is calibrated on the calibration traces and clipped to $[-\mu_{\max}, \mu_{\max}]$. The decoder reconstructs the cache using the same predictor state. Both sides reset their state at the beginning of each document to maintain causal consistency. The transform construction and prediction equations are given in Appendix C.2.

To balance generation quality and detection, we jointly allocate quantization mean-squared errors D_i over all transformed coordinate indices $\mathcal{I}$, where i is a coordinate index and $\mathcal{F} \subset \mathcal{I}$ is the set of fingerprint coordinates:

$$\sum_{i \in \mathcal{I}} w_i D_i \le D_{\mathrm{T}}, \qquad D_{\mathrm{T}} = \eta_{\mathrm{T}} \sum_{i \in \mathcal{I}} w_i \lambda_i, \qquad D_i + \sigma_i^2 \le g_i^2/\gamma, \quad i \in \mathcal{F}. \tag{2}$$

Here, w_i is the Fisher-sensitivity weight, D_{T} is the task-distortion budget, and g_i^2, σ_i^2, and γ are respectively the fingerprint-response power, the variance of errors allowed outside quantization, and the minimum detection signal-to-noise ratio. We parameterize the global budget by the normalized task-distortion ratio η_{T}, λ_i is the calibrated prediction-residual variance of coordinate i. We reduce the total code rate under these constraints and then

use closed-loop calibration to select the smallest available bit width and step size that meet the distortion target; see Appendix C.3.

The precision constraints above control reconstruction error in the fingerprint directions. We further use QIM so that the available reconstruction grid points for these coordinates depend on the fingerprint vector; its induced error is included in the same closed-loop calibration process:

$$q_{t,i} = \mathrm{round}(e_{t,i}/\Delta_i - \beta s_i), \qquad \widehat{e}_{t,i} = \Delta_i(q_{t,i} + \beta s_i). \tag{3}$$

Here, t is the token position, $e_{t,i}$ is the prediction residual, Δ_i is the step size, s_i is the corresponding fingerprint component, and $0 \le \beta < 1/2$ is the offset ratio; round selects the nearest integer to obtain the index $q_{t,i}$, and $\widehat{e}_{t,i}$ is the reconstructed residual. Other coordinates use zero offset. The residuals are reconstructed into the logical cache for inference and also serve as the observations for the audit below.

3.3 Fingerprint Recovery and Sequential Ownership Testing

For the same document, a trusted observer obtains the reconstructed cache Y of the target head and records the actual inputs to the target projection. The owner computes the uncompressed reference Y_0 from the clean model. The cache difference is unfolded in token order as $d = \mathrm{vec}(Y - Y_0)$. Using the recorded inputs and the private carriers, we construct a fingerprint-response matrix Φ corresponding to the observed cache. Each column represents the expected response contribution over the entire document when the corresponding fingerprint component is set to ± 1. When Φ has full column rank, the least-squares estimate is

$$d = \Phi s + \varepsilon, \qquad \widehat{s} = (\Phi^\top \Phi)^{-1} \Phi^\top d, \tag{4}$$

where ε denotes compression and other unexplained deviations, and $\widehat{s}$ is the sign estimate. We recover the signs by thresholding at zero and compute the document score $\tau = r^{-1} s^\top \widehat{s}$.

QIM may also cause a sign-dependent bias in clean caches. Therefore, on independent calibration documents, we use the same encoding and decoding parameters, QIM, and clean reference under the clean-weight hypothesis H_0 and the fingerprinted-weight hypothesis H_1, and fit score densities p_0 and p_1. The decision is based on the relative support of the observed score under these two distributions; agreement with the fingerprint vector alone is insufficient to establish ownership. After the calibration parameters are fixed, we accumulate the log-likelihood ratio over independent audit documents:

$$\Lambda_m = \sum_{n=1}^{m} \log \frac{p_1(\tau_n)}{p_0(\tau_n)}, \tag{5}$$

where m is the number of documents and τ_n is the score for document n. When the cumulative value Λ_m reaches the upper or lower threshold determined by the prescribed error probabilities, we accept H_1 or H_0, respectively; otherwise, we continue sampling. If the documents are exhausted, we report insufficient evidence. The fingerprint-response matrix, score fitting, and thresholds are detailed in Appendix C.4.

4 Experiments

4.1 Experimental Setup

Models. We evaluate FinCa on ten released language-model checkpoints spanning Qwen3, Llama 3.x, Gemma 2/3, Mistral, Ministral, GLM-4, GLM-Z1, and DeepSeek-V2-Lite (Qwen Team, 2025; Dubey et al., 2024; Meta AI, 2024; Gemma Team, 2024; 2025; Jiang et al., 2023; Mistral AI, 2024; Zeng et al., 2024; GLM Team, 2025; DeepSeek-AI, 2024).

Datasets. We use WikiText-2 as the common evaluation corpus (Merity et al., 2017); we also use alpaca-gpt4-data (English and Chinese) (AI-ModelScope, 2023) to simulate cross-dataset and cross-lingual fine-tuning during downstream distribution. The datasets are obtained from public repositories (Merity et al., 2017), and model execution uses the open-source Transformers implementation (Wolf et al., 2020).

Table 1: Matched conditions for ownership auditing. The letter denotes the checkpoint type, and the subscript denotes whether cache compression is disabled (0) or enabled (1).

Condition	Definition and role
C_0	Clean checkpoint with an uncompressed cache. It provides the common reference used to form both audit observations.
C_1	Clean checkpoint processed by the frozen fingerprint-aware codec. It defines the matched null condition H_0 and controls for codec and QIM effects.
F_0	Fingerprinted checkpoint with an uncompressed cache. It is used only to diagnose the effect of weight fingerprinting before compression.
F_1	Fingerprinted checkpoint processed by the same frozen codec. It defines the positive condition H_1 used in the ownership test.

Table 2: Primary evaluation metrics. Arrows indicate the preferred direction; detailed definitions and explanations are provided in Appendix D.

Metric	Interpretation
BER ↓	Fraction of incorrectly decoded payload bits, reported as a percentage.
Exact Rec. ↑	Fraction of documents for which all payload bits are recovered correctly; the optimum is 100%.
Min. SNR ↑	Detectability of the weakest protected fingerprint symbol vs. noise. A stronger weakest symbol means a more reliable system and, with other metrics, forms an evaluation criterion.
Gate	Passes only when every payload symbol satisfies $\widehat{\mathrm{SNR}}_i \geq \gamma$, with $\gamma = 2.6641$, see App. D.
$\Delta \log \mathrm{PPL}$ ↓	Change in log-perplexity of the fingerprinted route relative to its clean counterpart.
Transfer ↑	Ratio of the original FP16 KV-cache size to the complete FinCa representation size, including the cache stream and shared decoder state.
Active ↑	Ratio of the peak clean logical KV-cache size to the codec state retained during generation, including the shared decoder state.

Documents. We use three mutually disjoint document sets, splited from WikiText-2. Selection documents(12) are used to construct the input carriers; calibration documents(16) determine the cache representation, quantization parameters, and score distributions for clean and fingerprinted weights; audit documents(20) are used exclusively for final symbol recovery and ownership decisions. We skip the first 100 valid documents in train/validation and 80 in test, draw candidates from a 20-fold pool, and use a fixed seed. **Setup.** Each frozen configuration is evaluated once in BF16. Our workstation equipped with a A100 40GB GPU and Intel 14900kf CPU. **Evaluation Protocol and Metrics.** We report six primary metrics in Tab. 2. Unless otherwise stated, intervals are computed over documents using 10,000 bootstrap resamples, and clean baselines are repeated three times. The primary audit follows the four aligned routes in Tab. 1. Ownership is tested between the matched-codec clean route (C_1; H_0) and fingerprinted route (F_1; H_1); both use identical documents, token positions, codec/QIM state, and clean uncompressed-cache subtraction reference. Detailed definitions and calculations are deferred to Appendix B.

4.2 Selected Parameter Configurations and Validation across Ten Models

We test FinCa transferability with the same protocol on ten language models. For each model, WikiText-2 selection, calibration, and test documents are disjoint. We run 29 sequential trials over the fingerprint layer, perturbation amplitude α, task budget/rank, predictor clipping, and QIM offset; the selected configuration is then frozen and evaluated on 20 held-out documents. The final runs share the Value projection of head 0, 2 physical symbols, and 1 codec structure. Tab. 3 reports the primary held-out outcomes.

Table 3: Key final-validation results on 20 held-out documents. Transfer is the clean FP16 logical-cache size divided by the complete FinCa deployment representation size.

Model(Layer)	BER(%)↓	Rec.(%)↑	Gate	Transfer↑	Δlog-PPL↓
Qwen3 0.6B (L0)	0%	100%	pass	3.267	$+1.40 \times 10^{-2}$
Llama 3.2 1B (L0)	0%	100%	pass	3.198	$+2.95 \times 10^{-2}$
Gemma 3 1B IT (L0)	0%	100%	pass	3.263	-3.85×10^{-2}
Gemma 2 2B (L12)	0%	100%	pass	3.460	-1.17×10^{-2}
Mistral 7B (L0)	0%	100%	pass	3.429	$+1.83 \times 10^{-1}$
Ministral 8B (L9)	0%	100%	pass	3.220	$+1.25 \times 10^{-2}$
Llama 3.1 8B (L0)	0%	100%	pass	3.392	$+1.13 \times 10^{-1}$
GLM-4 9B (L10)	0%	100%	pass	3.064	$+1.03 \times 10^{-2}$
GLM-Z1 9B (L20)	0%	100%	pass	3.062	$+1.67 \times 10^{-2}$
DeepSeek V2 Lite (L13)	0%	100%	pass	3.817	$+1.02 \times 10^{-1}$

The main result is reliable recovery on all ten models. Each model has BER = 0, complete recovery rate on all 20 documents, and a passing SNR gate. The layer search probes how far the fingerprint insertion can move toward the output while remaining detectable under the same evaluation conditions. A later insertion leaves fewer downstream layers for propagation, making successful detection at that depth a stronger test of signal retention. The resulting Transfer ranges from 3.062 to 3.817 across models; its variation reflects cache statistics and prediction residuals. The change in language-modeling quality varies across models: Δlog-PPL is negative for Gemma 3 1B IT and Gemma 2 2B, but reaches +0.1827 for Mistral 7B and +0.1129 for Llama 3.1 8B.

To ensure a controlled and consistent comparison across models, we vary only the fingerprint insertion layer while keeping all other parameters fixed. Specifically, we set the task-budget ratio to $\eta_{\mathrm{T}} = 4 \times 10^{-3}$, the perturbation amplitude to $\alpha = 0.01$, the task rank to $q = 4$, and the predictor clipping threshold to $\mu_{\max} = 0.9$. The detailed parameter-selection process, together with an analysis of how these parameters affect the performance of FinCa, is provided in Appendix E.2.

4.3 Fingerprint Strength Ablation

Fig. 3 shows that α provides a controllable trade-off among detection strength, recovery reliability, and model quality. Increasing α generally changes the H1–H0 ownership gap and can reduce BER for models with weaker fingerprint responses. However, the effect is model-dependent and is not strictly monotonic across all models. The total Δ log-PPL also varies with α, whereas the compression ratio remains nearly unchanged. These results confirm that α primarily controls the fingerprint signal rather than the codec rate. We therefore select the best-performing α for each model under the same search protocol to obtain a relatively fair model-specific operating point for all subsequent experiments. The complete frozen search parameters for each model are reported in Tab. 8.

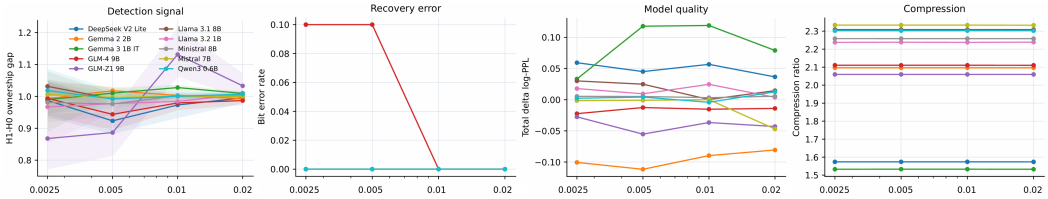


Figure 3: The four panels report the H1–H0 ownership-score gap, bit error rate, total Δ log-PPL, and compression ratio as a function of α with [0.0025, 0.005, 0.01, 0.02]. The ownership-gap shaded regions denote document-level bootstrap confidence intervals.

4.4 Component Ablation Study

We conduct ablation experiments on the key designs of FinCa to observe which design choices allow FinCa to remain detectable after cache coding. We select three mainstream open-weight models, Qwen3-0.6B, Llama-3.2-1B, and Gemma-3-1B—with a two-symbol fingerprint inserted into one Value-projection KV head. each using its frozen final configuration, with all other settings consistent with those in Section 4.1.

For each document, the fingerprinted stream (H_1) is compared with a clean stream (H_0) under the same codec state. We report bit–error–rate(BER), minimum per-symbol SNR, token-weighted $\Delta \log \mathrm{PPL}$, and transfer/active logical-cache compression. Detection uses the propagation-aware least-squares matched statistic; the SNR gate requires every symbol to exceed the threshold for a 0.1 payload-failure rate.

The full method recovers all 40 held-out symbols (BER = 0) for all models. The matched statistic is essential: replacing it with carrier correlation yields BERs of 75.0% (Qwen), 100.0% (Llama), and 2.5% (Gemma); Llama nevertheless passes SNR because this metric discards the sign of the mean gap. Bypassing the codec raises SNR but removes the deployable stream. Removing key-derived carriers preserves BER on these held-out documents but changes SNR and PPL in a model-dependent way. We retain it as part of the full design for

Table 4: Ablations on three open-weight model families. "Alt.layer" replaces only the Value-carrier layer. "Deployable" records whether a public cache stream exists. A variant is eligible as a complete FinCa configuration only when BER = 0, the SNR gate passes, and Deployable is "✓". "−" denotes a metric that is undefined when the codec is bypassed.

Variant	Model	BER(%) ↓	Min.SNR ↑	Gate	Deployable	Δ log PPL ↓	Transfer / Active ↑
Full FinCa	Qwen3-0.6B	0.0%	318.6	✓	✓	$+2.78e^{-4}$	3.26× / 2.50×
	Llama-3.2-1B	0.0%	5993	✓	✓	$+4.03e^{-2}$	3.20× / 2.48×
	Gemma-3-1B	0.0%	182.5	✓	✓	$+2.39e^{-3}$	3.30× / 1.30×
No matched statistic	Qwen3-0.6B	75.0%	0.8212	✗	✓	$+2.78e^{-4}$	3.26× / 2.50×
	Llama-3.2-1B	100.0%	15.78	✓	✓	$+4.03e^{-2}$	3.20× / 2.48×
	Gemma-3-1B	2.5%	1.612	✗	✓	$+2.39e^{-3}$	3.30× / 1.30×
No codec	Qwen3-0.6B	0.0%	615.9	✓	✗	$-3.74e^{-4}$	− / −
	Llama-3.2-1B	0.0%	1.307×10^{6}	✓	✗	$+2.78e^{-3}$	− / −
	Gemma-3-1B	0.0%	667.7	✓	✗	$+1.09e^{-2}$	− / −
No key-derived carriers	Qwen3-0.6B	0.0%	185.9	✓	✓	$-2.23e^{-3}$	3.26× / 2.50×
	Llama-3.2-1B	0.0%	907.8	✓	✓	$+3.25e^{-2}$	3.15× / 2.45×
	Gemma-3-1B	0.0%	708.3	✓	✓	$+2.56e^{-2}$	3.30× / 1.30×
Alt.layer	Qwen3-0.6B	0.0%	103.8	✓	✓	$-5.25e^{-3}$	3.26× / 2.50×
	Llama-3.2-1B	0.0%	375.6	✓	✓	$+3.13e^{-2}$	3.15× / 2.45×
	Gemma-3-1B	0.0%	14.99	✓	✓	$-2.79e^{-3}$	3.30× / 1.30×

private carrier construction. In contrast, changing only the carrier layer lowers minimum SNR to 103.8, 375.6, and 14.99 for Qwen, Llama, and Gemma, respectively, showing that carrier placement materially affects detection margin.

4.5 Comparison with recent LLM fingerprinting methods

To test whether FinCa's motivation persists against both classic DNN watermarking/fingerprinting methods and more recent LLM fingerprinting designs, we select several representative methods for the experiment on Qwen3-0.6B: Trigger-set (Adi et al., 2018), Weight watermark (Uchida et al., 2017), Deepsigns (Rouhani et al., 2019), UTF (Cai et al., 2025), EverTracer (Xu et al., 2025a), DuFFin (Yan et al., 2026), and LLMPrint (Hu et al., 2026b).

Method Adaptation. Following the original papers, we construct the fingerprints before applying KV-cache compression. UTF uses under-trained-token screening on the selection split, secret input–output pairs, and supervised fine-tuning. EverTracer uses victim-model fingerprint injection, a matched reference model, semantic-neighborhood queries, and calibrated probability variation. DuFFin uses its trigger-level and knowledge-level fingerprint queries together with the corresponding joint verification procedure. LLMPrint uses GCG-optimized prompts that induce pairwise first-token preferences near the decision boundary.

After fingerprint construction, each method is passed through the same KV-cache deployment pipeline wherever possible, using the same model checkpoint, compression operator, compression schedule, token-retention setting, and held-out evaluation protocol. We rerun the corresponding verification procedure on the reconstructed cache and convert its outcome into the common comparison metrics used here: BER, exact recovery, Pass, and SNR. This protocol measures how much verifiable evidence survives when fingerprints originally designed around parameters, model outputs, probability variation, or prompt-induced token preferences are transmitted through a lossy KV-cache channel. The comparison therefore isolates the question of interest to FinCa: after the corresponding evidence reaches a deployed KV cache, how much verifiable evidence remains under cache compression?

Protocol. We use the layer-0/head-0 KV rows of Qwen3-0.6B, with all other settings consistent with those in Section 4.4. KIVI-style affine quantization uses per-channel Keys and per-token Values, while token retention keeps four sink tokens and the most recent rows. The stress path includes FP16/full, INT8/full, INT4/full, INT3/50%, INT2/50%, and INT2/25%, where the percentage denotes the fraction of original KV-cache tokens retained after token selection. FinCa protects its two carrier axes with a six-bit QIM lattice

Table 5: Survival of fingerprint evidence under KV-cache compression. FinCa is compared with fair-support adaptations of recent LLM fingerprinting methods. Pass is the fraction of keys satisfying both the SNR and exact-recovery gates.

Method	FP16/full (≈ 1×, baseline)					INT8/full (≈ 2×)				
	Med. SNR	SNR (p10–p90)	BER	Pass	Exact	Med. SNR	SNR (p10–p90)	BER	Pass	Exact
FinCa	–	–	**0.0%**	**100.0%**	**100.0%**	**17660**	**7260–48553**	**0.0%**	**100.0%**	**100.0%**
Trigger-set	–	–	0.0%	100.0%	100.0%	5727	3298–13504	0.0%	100.0%	100.0%
Weight watermark	–	–	0.0%	100.0%	100.0%	530	306–1237	0.0%	100.0%	100.0%
Deepsigns	–	–	0.0%	100.0%	100.0%	548	305–1018	0.0%	100.0%	100.0%
UTF	–	–	0.0%	100.0%	100.0%	3738	2336–5966	0.0%	100.0%	100.0%
EverTracer	–	–	0.0%	100.0%	100.0%	3177	2399–6353	0.0%	100.0%	100.0%
DuFFin	–	–	0.0%	100.0%	100.0%	4292	2117–6348	0.0%	100.0%	100.0%
LLMPrint	–	–	0.0%	100.0%	100.0%	4351	2128–5134	0.0%	100.0%	100.0%
Method	**INT4/full** (≈ 4×)					**INT3/50%** (≈ 10.6×)				
	Med. SNR	SNR (p10–p90)	BER	Pass	Exact	Med. SNR	SNR (p10–p90)	BER	Pass	Exact
FinCa	**2473**	**1504–5204**	**0.0%**	**100.0%**	**100.0%**	**1409**	**770–2264**	**0.0%**	**100.0%**	**100.0%**
Trigger-set	44.3	14.3–84.5	0.0%	50.0%	78.25%	3.03	1.12–8.46	16.38%	0.0%	29.75%
Weight watermark	22.5	14.7–28.5	0.125%	10.0%	75.25%	6.99	4.53–9.54	2.625%	0.0%	40.50%
Deepsigns	3.99	2.99–12.4	3.875%	0.0%	24.75%	1.40	0.515–2.64	31.0%	0.0%	2.25%
UTF	32.5	18.8–43.4	0.0%	55.0%	88.50%	6.62	4.55–11.1	12.375%	0.0%	37.75%
EverTracer	19.3	12.0–29.3	0.0%	40.0%	85.25%	2.78	0.320–6.38	15.25%	0.0%	29.0%
DuFFin	20.9	16.1–29.2	0.25%	25.0%	82.75%	4.63	3.10–7.44	14.125%	0.0%	34.25%
LLMPrint	23.9	15.1–40.3	0.0%	40.0%	82.50%	4.94	2.36–11.2	12.75%	0.0%	35.50%
Method	**INT2/50%** (≈ 16×)					**INT2/25%** (≈ 32×)				
	Med. SNR	SNR (p10–p90)	BER	Pass	Exact	Med. SNR	SNR (p10–p90)	BER	Pass	Exact
FinCa	**1409**	**770–2264**	**0.0%**	**100.0%**	**100.0%**	**580**	**323–1121**	**0.0%**	**100.0%**	**100.0%**
Trigger-set	2.31	1.05–3.23	17.75%	0.0%	110.0%	1.11	0.563–3.62	66.625%	0.0%	5.50%
Weight watermark	3.03	1.55–5.26	7.625%	0.0%	21.0%	1.51	0.815–2.43	13.625%	0.0%	10.0%
Deepsigns	0.406	0.139–1.13	47.125%	0.0%	0.25%	0.258	0.102–0.638	66.875%	0.0%	0.0%
UTF	2.95	1.56–5.61	16.25%	0.0%	21.50%	2.64	1.20–4.21	33.875%	0.0%	11.75%
EverTracer	1.06	0.147–2.33	21.125%	0.0%	13.75%	0.318	0.0477–1.52	39.125%	0.0%	7.0%
DuFFin	2.37	1.50–4.02	19.50%	0.0%	18.75%	1.79	0.796–2.61	36.125%	0.0%	12.75%
LLMPrint	2.83	1.16–4.59	16.0%	0.0%	19.25%	1.93	0.884–3.07	32.75%	0.0%	11.0%

at the lower nominal bit widths; its exact mean bit rate, including this allocation, is used when computing the nominal compression factor. The primary endpoint is the percentage of keys satisfying both the Min. SNR criterion and exact recovery on at least 90% of the test documents. We also report the median SNR and its 10th–90th percentile across keys. Fingerprint construction and quantizer calibration use only the selection split.

Result. In our experiments, all methods achieve 100% Pass and Exact with zero BER under FP16/full and INT8/full. Under INT4/full, FinCa retains 100% on both measures while the baselines reach only 0–55% Pass; under INT3/50%, INT2/50%, and INT2/25%, FinCa remains at 100% on both measures, whereas every baseline has 0% Pass and lower exact recovery. These results show that protecting the carrier directions preserves fingerprint recovery under aggressive compression; Appendix G.2 extends the comparison to a range of representative, diverse, and demanding KV-cache compression schemes.

4.6 Fine-Tuned FinCa Detection

We conduct multi-epoch $(1, 3, 5, 7)$ fine-tuning on the fingerprinted checkpoints (Qwen3 0.6B, Llama 3.2 1B, Gemma 3 1B IT) using the alpaca-gpt4-data (en&zh) datasets to investigate the survival of FinCa under cross-dataset and cross-lingual fine-tuning intensities. Each checkpoint is paired with a clean run using identical training and evaluation settings. Detection uses the frozen WikiText-2 carrier, with 20 fingerprinted and 20 clean test documents per condition. For counts (TP, FN, FP, TN), we report TPR, FNR, FPR, and TNR.

Tab. 6 reports TPR trajectories and the corresponding FPR trajectory. The detector is perfectly stable for Llama and Qwen in both languages. Gemma 3 retains 90% TPR at initialization, but declines with training, more sharply on Chinese data (90/65/40/10%). Therefore, fine-tuning can preserve or erase the carrier depending on the model and corpus; persistence should be measured empirically rather than assumed from the pre-fine-tuning checkpoint.

Table 6: Fine-tuned FinCa detection results. Entries are reported as point estimate [95% CI], with all values in percentage points.

Model	Epoch(s)	English		Chinese	
		TPR(%)	**FNR**(%)	**TPR**(%)	**FNR**(%)
Gemma 3 1B IT	1	90.0 [68.3, 98.8]	10.0 [1.2, 31.7]	90.0 [68.3, 98.8]	10.0 [1.2, 31.7]
	3	90.0 [68.3, 98.8]	10.0 [1.2, 31.7]	65.0 [40.8, 84.6]	35.0 [15.4, 59.2]
	5	70.0 [45.7, 88.1]	30.0 [11.9, 54.3]	40.0 [19.1, 63.9]	60.0 [36.1, 80.9]
	7	65.0 [40.8, 84.6]	35.0 [15.4, 59.2]	10.0 [1.2, 31.7]	90.0 [68.3, 98.8]
Llama 3.2 1B	1,3,5,7	100.0 [83.2, 100.0]	0.0 [0.0, 16.8]	100.0 [83.2, 100.0]	0.0 [0.0, 16.8]
Qwen3 0.6B	1,3,5,7	100.0 [83.2, 100.0]	0.0 [0.0, 16.8]	100.0 [83.2, 100.0]	0.0 [0.0, 16.8]

Note. FNR is the complement of TPR, i.e., FNR = 100% − TPR.

4.7 Tracing Fingerprint Propagation Across Layers and KV Heads

This section we ask whether a fingerprint-induced change remains localized at the injection layer or propagates to subsequent Transformer blocks. Figure 4 traces this response across layers and physical KV-cache heads.

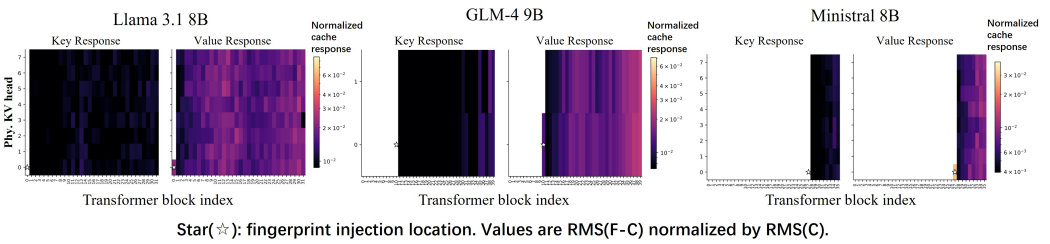


Figure 4: Layer–head propagation of fingerprint-induced key- and value-cache perturbations. The star marks the fingerprint injection layer, and each cell reports the normalized response.

For $x \in \{K, V\}$, let $C_{\ell,h}^{(x)}$ and $F_{\ell,h}^{(x)}$ denote the uncompressed cache tensors produced by the clean and fingerprinted models, respectively, using the same input sequence and cache positions. We define

$$R_{\ell,h}^{(x)} = \mathrm{RMS}\Big(F_{\ell,h}^{(x)} - C_{\ell,h}^{(x)}\Big) \Big/ \Big(\mathrm{RMS}\Big(C_{\ell,h}^{(x)}\Big) + \epsilon\Big), \qquad x \in \{K, V\},$$

where $\mathrm{RMS}(\cdot)$ is computed over token positions and cache coordinates, and $\epsilon > 0$ prevents division by zero. Thus, $R_{\ell,h}^{(x)}$ measures the fingerprint-induced change relative to the corresponding clean cache. The response is strongest near the injection layer but remains visible in subsequent layers, showing that the fingerprint signal is carried forward through the cache. Value responses often extend across more downstream blocks, whereas key responses are more localized. Fingerprints injected near the final layers produce more localized patterns because fewer downstream blocks remain. The corresponding layer-placement sweep is reported in Appendix F.

5 Conclusion

FinCa addresses ownership verification under KV-cache compression. It embeds a secret low-rank fingerprint in a physical KV-head projection and allocates cache distortion to preserve both task-sensitive and fingerprint-bearing directions. Matched auditing against an identically compressed clean control separates weight-derived evidence from codec bias. This formulation makes explicit that cache compression and ownership verification must be designed jointly, rather than evaluated independently. Our results support FinCa as a reproducible framework for studying this trade-off under trusted logical-cache observation.

AI Use Acknowledgments

During the preparation of this work, the authors used generative AI tools, including ChatGPT (OpenAI), as limited auxiliary tools in several aspects of the research and manuscript preparation. In particular, these tools were used to assist with brainstorming and organizing experimental procedures, clarifying implementation details, suggesting possible debugging strategies, and improving the readability, grammar, and overall presentation of the manuscript.

For the implementation, generative AI was used to provide limited assistance with drafting, modifying, explaining, and debugging code, as well as with writing auxiliary scripts for data preprocessing, experiment management, and result visualization. All AI-assisted code was subsequently inspected, adapted, executed, and tested by the authors. The authors independently verified that the code used to produce the reported results was consistent with the methods described in the paper and that the experimental results could be reproduced from the final implementation. Generative AI was not used as a substitute for the authors' responsibility for experimental design, implementation decisions, or interpretation of the results.

With respect to data, generative AI was used, where applicable, only for limited assistance with data-processing scripts, file-format conversion, documentation, and organizational tasks. The tools were not relied upon to invent, fabricate, selectively alter, or otherwise manipulate experimental data, labels, measurements, or reported results. The authors independently collected, curated, processed, and checked the data used in the experiments, and verified that all quantitative results and figures reported in the paper were supported by the underlying data and experimental records.

For mathematical analysis and proofs, generative AI was used only as an auxiliary tool for checking notation, improving exposition, identifying possible gaps in explanations, and suggesting alternative ways to present technical arguments. No AI-generated mathematical statement, derivation, or proof was accepted without independent verification. The authors are solely responsible for the validity of all definitions, lemmas, propositions, theorems, proofs, assumptions, and theoretical claims presented in this work.

Generative AI was also used for language polishing, including grammar correction, sentence restructuring, terminology consistency, and improvement of clarity and readability. The use of these tools did not replace the authors' intellectual contributions, scientific judgment, or responsibility for the manuscript. The authors reviewed and revised all AI-assisted text, verified the accuracy of the cited claims and technical descriptions, and ensured that the final manuscript accurately reflects the authors' own work and conclusions.

All final content, including the experimental methodology, code, data processing procedures, theoretical analysis, proofs, figures, tables, and conclusions, was reviewed and approved by the authors. The authors take full responsibility for the accuracy, integrity, reproducibility, originality, and appropriate attribution of the final publication, including any content that was produced or refined with the assistance of generative AI.

Reproducibility Statement

We provide an anonymized supplementary package that fully reproduces the hyperparameter search experiments and the component ablation experiments with Qwen3-0.6B, Llama-3.2-1B, and Gemma-3-1B. It includes all code, configurations, and instructions. All experiments in the main text and appendix can be reproduced with AI assistance using the supplementary code. Models and datasets are not redistributed; users must download them from official sources and configure the paths as described in the README. The package has been anonymized and contains no identifying information.

References

Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX Security Symposium*, pp. 1615–1631, 2018.

AI-ModelScope. alpaca-gpt4-data, 2023. URL https://www.modelscope.cn/datasets/AI-ModelScope/alpaca-gpt4-data. English and Chinese instruction-following dataset.

Jiacheng Cai, Jiahao Yu, Yangguang Shao, Yuhang Wu, and Xinyu Xing. UTF: Undertrained tokens as fingerprints —— a novel approach to LLM identification. In Leon Derczynski, Jekaterina Novikova, and Muhao Chen (eds.), *Proceedings of the The First Workshop on LLM Security (LLMSEC)*, pp. 1–6, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-279-4. URL https://aclanthology.org/2025.llmsec-1.1/.

Joseph Cox. Facebook's powerful large language model leaks online. VICE, March 2023. URL https://www.vice.com/en/article/facebooks-powerful-large-language-model-leaks-online-4chan-llama/. Accessed: 2026-09-24.

DeepSeek-AI. Deepseek-v2: A strong mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.

Abhimanyu Dubey et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Hang Fu, Wanli Peng, Yinghan Zhou, Jiaxuan Wu, Juan Wen, and Xue Yiming. Inhibitory attacks on backdoor-based fingerprinting for large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26246–26266, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1207. URL https://aclanthology.org/2026.acl-long.1207/.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

GLM Team. Glm-z1-9b-0414. Official model release, 2025. URL https://huggingface.co/THUDM/GLM-Z1-9B-0414.

Yifei He, Luozhi Zhang, Wenhao Wu, and Jingyu Zhou. ZipCache: Accurate and efficient KV cache quantization with salient token identification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 1270–1303. Curran Associates, Inc., 2024. doi: 10.52202/079017-0040. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/028fcbcf85435d39a40c4d61b42c99a4-Paper-Conference.pdf.

Yuepeng Hu, Zhengyuan Jiang, Mengyuan Li, Osama Ahmed, Zhicong Huang, Cheng Hong, and Neil Zhenqiang Gong. Fingerprinting LLMs via prompt injection. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 11795–11810, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.541. URL https://aclanthology.org/2026.acl-long.541/.

Yuepeng Hu, Zhengyuan Jiang, Mengyuan Li, Osama Ahmed, Zhicong Huang, Cheng Hong, and Neil Zhenqiang Gong. Fingerprinting LLMs via prompt injection. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11795–11810, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.541. URL https://aclanthology.org/2026.acl-long.541/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lelio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothee Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Hao Kang, Qingru Zhang, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, and Tuo Zhao. GEAR: An efficient KV cache compression recipe for near-lossless generative inference of LLM. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023.

Junyan Li, Tianle Cai, Yang Zhang, Muhammad Yusuf Hassan, Talha Chafekar, Colorado Reed, Zhile Ren, Pengsheng Guo, Binazir Karimzadeh, Chong Wang, and Chuang Gan. Commvq: Commutative vector quantization for kv cache compression. In *ICML*, 2025a. URL https://arxiv.org/abs/2506.18879.

Tianyang Li et al. Seedprints: Fingerprints can even tell which seed your large language model was trained from. *ArXiv*, 2025b. Uses initialization-born biases as persistent seed-dependent identifiers.

Y. Li et al. Sraf: Stealthy and robust adversarial fingerprint for copyright verification of large language models. *ArXiv*, 2025c. Adversarial fingerprinting framework for LLM ownership verification.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR, 2024.

Zichang Liu, Aditya Desai, Fang Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023a.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 52342–52364. Curran Associates, Inc., 2023b. doi: 10.52202/075280-2279. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a452a7c6c463e4ae8fbdc614c6e983e6-Paper-Conference.pdf.

Yiyang Luo, Ke Lin, Chao Gu, Jiahui Hou, Lijie Wen, and Luo Ping. Lost in overlap: Exploring logit-based watermark collision in LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 620–637, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.37. URL https://aclanthology.org/2025.findings-naacl.37/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Meta AI. Llama 3.2. Official model release, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-1B.

Mistral AI. Ministral 8b instruct 2410. Official model release, 2024. URL https://huggingface.co/mistralai/Ministral-8B-Instruct-2410.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. DeepSigns: An end-to-end watermarking framework for ownership verification of deep neural networks. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, 2019.

Zunhai Su, Hanyu Wei, Zhe Chen, Wang Shen, Linge Li, Huangqi Yu, and Kehong Yuan. Rotatekv: accurate and robust 2-bit kv cache quantization for llms via outlier-aware adaptive rotations. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/690. URL https://doi.org/10.24963/ijcai.2025/690.

Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM International Conference on Multimedia Retrieval (ICMR)*. ACM, 2017.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Zhenhua Xu, Meng Han, and Wenpeng Xing. EverTracer: Hunting stolen large language models via stealthy and robust probabilistic fingerprint. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7008–7031, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.358. URL https://aclanthology.org/2025.emnlp-main.358/.

Zhenhua Xu, Xubin Yue, Zhebo Wang, Qichen Liu, Xixiang Zhao, Jingxuan Zhang, Wenjun Zeng, Wengpeng Xing, Dezhang Kong, Changting Lin, and Meng Han. Copyright protection for large language models: A survey of methods, challenges, and trends. *ArXiv*, abs/2508.11548, 2025b. Survey paper; includes unified framework distinguishing invasive vs. intrinsic fingerprinting.

Zhenhua Xu, Xixiang Zhao, Xubin Yue, Shengwei Tian, Changting Lin, and Meng Han. CTCC: A robust and stealthy fingerprinting framework for large language models via cross-turn contextual correlation backdoor. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6967–6989, Suzhou, China, November 2025c. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.356. URL https://aclanthology.org/2025.emnlp-main.356/.

Yuliang Yan, Haochun Tang, Shuo Yan, and Enyan Dai. DuFFin: A dual-level fingerprinting framework for LLMs IP protection. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5168–5184, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.273. URL https://aclanthology.org/2026.findings-eacl.273/.

Dogyung Yoon, Minsoo Chun, Thomas Allen, H. Muller, Min Wang, and Rajesh Sharma. Intrinsic fingerprint of llms: Continue training is not all you need to steal a model! *ArXiv*, abs/2507.03014, 2025. Introduces attention parameter standard deviation as intrinsic fingerprint.

Aohan Zeng et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. URL https://arxiv.org/abs/2406.12793.

Haobo Zhang, Zhenhua Xu, Junxian Li, Shangfeng Sheng, Dezhang Kong, and Meng Han. Attndiff: Attention-based differential fingerprinting for large language models, 2026. URL https://arxiv.org/abs/2604.05502.

Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph. Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the ACM Asia Conference on Computer and Communications Security*, pp. 159–172, 2018. doi: 10.1145/3196494.3196550.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. Reef: Representation encoding fingerprints for large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 48092–48117, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/77e830bd802f4b63b682038d090a97b6-Paper-Conference.pdf.

Ruisi Zhang and Farinaz Koushanfar. Emmark: Robust watermarks for IP protection of embedded quantized large language models. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pp. 1–6. ACM, 2024. doi: 10.1145/3649329.3655674. URL https://doi.org/10.1145/3649329.3655674.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang "Atlas" Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34661–34710. Curran Associates, Inc., 2023. doi: 10.52202/075280-1506. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6ceefa7b15572587b78ecfcebb2827f8-Paper-Conference.pdf.

A Related Work

Model fingerprinting for large language models. Recent work has unified text watermarking, model watermarking, and model fingerprinting under a common framework, distinguishing *intrinsic fingerprinting*—which extracts non-invasive signatures without modifying training—from *invasive fingerprinting*, which actively embeds verifiable information (Xu et al., 2025b).

Intrinsic fingerprinting. Parameter-distribution signatures distinguish independently trained models via attention-matrix standard deviations that remain stable across continued training (Yoon et al., 2025). Initialization-born next-token biases persist throughout training as seed-dependent identifiers (Li et al., 2025b). Behavioral fingerprints, such as safety-aligned refusal patterns, also asked to remain stable across fine-tuning, quantization, and model merging. These intrinsic approaches exploit natural statistical patterns without altering training.

REEF constructs training-free representation-level fingerprints by comparing the centered kernel alignment (CKA) of internal representations produced by a victim and a suspect model on the same samples (Zhang et al., 2025). It is robust to fine-tuning, pruning, model merging, and parameter permutations, but assumes access to internal activations or representations. FinCa instead fingerprints runtime KV-cache signals and studies their survival under cache compression, representing a different signal location and access model.

LLMPrint provides a complementary black-box and gray-box approach by optimizing prompt-injection fingerprints that induce model-specific token preferences (Hu et al., 2026a). Its fingerprints are designed to remain identifiable after post-training and model quantization. However, the tech community is overwhelmingly skeptical, pointing out that paraphrasing or round-trip translation trivially destroys the watermark(Luo et al., 2025), whereas FinCa does not rely on output-only prompt queries and targets the cache-compression channel.

Invasive fingerprinting. Invasive fingerprinting actively embeds verifiable evidence into models during or after training. Weight-based methods encode signatures via parameter regularizers (Uchida et al., 2017; Fernandez et al., 2023), while backdoor-based methods associate trigger sets with predefined responses for verification through the prediction interface (Zhang et al., 2018; Adi et al., 2018). For LLMs, triggers range from unnatural token sequences to natural-language sentences (Xu et al., 2025b). Recent adversarial frameworks enhance robustness against fine-tuning, pruning, and model merging (Li et al., 2025c).

EmMark is particularly relevant to quantized deployment: it is a post-training weight-watermarking method explicitly designed to preserve watermark recoverability after quantization (Zhang & Koushanfar, 2024). Its protected object is the model-weight watermark, whereas FinCa protects a runtime cache-difference signal against KV-cache compression; therefore, the two methods address different distortion channels.

KV-cache compression. KV-cache compression reduces memory and communication costs via complementary mechanisms. Low-bit methods quantize keys and values, accounting for channel-wise outliers or token-dependent sensitivity (Liu et al., 2024; Hooper et al., 2024; He et al., 2024). Token-reduction methods retain only states deemed important for future attention (Liu et al., 2023a; Zhang et al., 2023; Xiao et al., 2024; Li et al., 2024). These methods optimize for reconstruction error or language-model utility. A fingerprint signal can, however, occupy a low-energy direction largely irrelevant to these objectives and be aggressively distorted. FinCa explicitly augments compression with a detection constraint, separating fingerprint-bearing and task-sensitive coordinates before allocating quantization distortion.

B Experimental Configuration Details

Frozen search protocol. Every model uses the same 5 sequential stages and 29 trials: (i) 5 layer candidates (layer 0, one-quarter, one-half, three-quarters, and the last layer); (ii) $\alpha \in \{0.0025, 0.005, 0.01, 0.02\}$; (iii) task-budget ratio $\in \{0.0005, 0.001, 0.002, 0.004\}$ crossed with task rank $\in \{4, 8, 16\}$; (iv) predictor clip $\in \{0.90, 0.95, 0.98, 0.995\}$; and (v) QIM offset fraction $\in \{0, 0.125, 0.25, 0.375\}$. The common codec settings are 2–16 quantization bits, 24 step-search points over an 8.0 span, distortion tolerance 0.001, and 32-bit basis storage. The selected budget ratio is 0.004 for all 10 models; layer, α, rank, predictor clip, and QIM offset remain model-specific (Table 8).

Data and runtime details. For final testing, we used 12 selection documents, 16 calibration documents, and 20 test documents, with a 20-fold candidate pool, 96-token sequences, and a 16-token minimum document length. We skipped the first 100 valid documents in the training and validation splits and the first 80 in the test split. The fingerprint targeted the Value projection of head 0, with a two-bit payload, two physical symbols, and ECC minimum distance 1. Runs used CUDA, bfloat16 weights, fixed seed 2027, and per-model key derivation from a master key; fingerprint persistence was switched to a dense representation for the frozen final test. Search parallelism followed per-model resource profiles, with at most eight concurrent trials and three CPU threads per trial.

Hardware and software environment. All experiments are conducted on a single local workstation running 64-bit Windows 10. The workstation is equipped with an NVIDIA A100 GPU with 40 GB of VRAM, an Intel processor with 24 physical cores and 32 logical threads, and 32 GB of system memory. GPU execution uses NVIDIA driver 595.95 (CUDA runtime reported by the driver: 13.2). The experiments run in the project-managed Python 3.11.15 environment with PyTorch 2.7.1 compiled for CUDA 12.8 and Transformers 4.53.3. CUDA inference is enabled for all reported runs.

Secret token. We set "ICLR2027-fpkvc-controlled-key-20260807" as secret token for all exp. Repeat friendly. Changing the key has only a negligible impact on the experimental results presented in this paper.

C Supplementary Method Details

C.1 Construction of the Input and Output Carriers

This section supplements the construction of the input carrier A, output carrier B, and fingerprint vector s in the main text. Suppose that the selected documents provide N tokens, and let the input hidden state to the target projection at token n be $h_n \in \mathbb{R}^{d_h}$. Arrange these states row-wise as $H = (h_1, \ldots, h_N)^\top \in \mathbb{R}^{N \times d_h}$. Let $\bar{h} = N^{-1} \sum_{n=1}^{N} h_n$ be the mean hidden state, and let $\mathbf{1}_N$ be the column vector containing N ones. The centered matrix H_c and its singular value decomposition are

$$H_c = H - \mathbf{1}_N \bar{h}^\top = P \Sigma V^\top, \tag{6}$$

where P and V contain the left and right singular vectors, respectively, and Σ is the diagonal matrix of singular values sorted in descending order. The first r columns of V form $V_r \in \mathbb{R}^{d_h \times r}$.

Let $R \in \mathbb{R}^{r \times r}$ be a private-key-generated orthogonal matrix satisfying $R^\top R = I_r$, where I_r is the r-th-order identity matrix. The input carrier is $A = V_r R$. Selecting principal activation directions helps the carriers respond to natural-text inputs, while the private-key-controlled rotation determines the specific directions. The same private key also generates an output carrier $B \in \mathbb{R}^{d_v \times r}$ satisfying $B^\top B = I_r$, together with a fingerprint vector $s \in \{-1, +1\}^r$. These quantities are used for the weight perturbation in Eq. 1 of the main text.

Let t denote a token position, and let $h_t \in \mathbb{R}^{d_h}$ be the actual projection input at that position. The Value response Δv_t produced by the weight perturbation for this input is

$$\Delta v_t = \alpha B \, \mathrm{diag}(s) A^\top h_t. \tag{7}$$

Here, $A^\top h_t$ determines the activation magnitude of each symbol, B determines the directions into which the response enters the Value cache, and s determines the sign of each carrier component.

C.2 Cache Orthogonal Transform and Causal Prediction

This section supplements the cache transform and causal prediction in the main text. We process the logical cache token by token for every layer and KV head. Let b index cache blocks, and let $x_t^{(b)} \in \mathbb{R}^{d_b}$ denote the cache row of block b at position t, where d_b is its dimension. For explicit Key and Value caches, this row is formed by concatenating the Key and Value coordinates of the corresponding head. All cache blocks participate in compression; the target block $b^\star = (\ell^\star, p^\star)$ has a separate set of fingerprint coordinates.

In the target block, we embed the output carrier B into the Value coordinates and orthogonalize it to obtain the fingerprint directions U_f. Based on the Fisher sensitivities computed from calibration documents, we select q language-task directions U_u in its orthogonal complement, where q is the prespecified number of task directions; the remaining directions form $U_\perp$. The complete orthogonal transform for the target block is

$$U^{(b^\star)} = [U_f, U_u, U_\perp]. \tag{8}$$

The transforms for other blocks consist of the task directions and their orthogonal complements. Here, $U_f \in \mathbb{R}^{d_{b^\star} \times r}$, $U_u \in \mathbb{R}^{d_{b^\star} \times q}$, and $U_\perp \in \mathbb{R}^{d_{b^\star} \times (d_{b^\star} - r - q)}$. The complete transform satisfies ${U^{(b^\star)}}^\top U^{(b^\star)} = I_{d_{b^\star}}$, where $I_{d_{b^\star}}$ is the identity matrix of the corresponding order.

We omit the cache-block superscript below and use U for the orthogonal transform of the current block and x_t for the current cache row. The transformed coordinates are $z_t = U^\top x_t$. Let μ denote the prediction coefficient fitted on calibration documents and constrained to a stable range, and let $\widehat{z}_{t-1}$ be the reconstructed coordinates at the previous time step. The current prediction residual e_t, quantized residual $\widehat{e}_t$, reconstructed coordinates $\widehat{z}_t$, and reconstructed cache row $\widehat{x}_t$ satisfy

$$e_t = z_t - \mu \widehat{z}_{t-1}, \qquad \widehat{z}_t = \mu \widehat{z}_{t-1} + \widehat{e}_t, \qquad \widehat{x}_t = U \widehat{z}_t. \tag{9}$$

At the beginning of each document, both encoder and decoder set the prediction state to zero and always use the reconstructed coordinates at the previous time step to predict the current coordinates, thereby keeping the prediction states consistent.

The next section gives the distortion allocation and quantizer calibration for each coordinate.

C.3 Detection-Constrained Precision Allocation

After fixing the cache transform and causal predictor, let $\lambda_i > 0$ be the variance of the prediction residual on transformed coordinate $i \in \mathcal{I}$. Let $D_i > 0$ denote the quantization mean-squared error allocated to that coordinate, let $w_i > 0$ denote the Fisher-sensitivity weight of the language-modeling loss to the error in that coordinate, and let $D_{\mathrm{T}} > 0$ denote the task-distortion budget shared by all cache blocks. For a fingerprint coordinate $i \in \mathcal{F}$, g_i^2 denotes the cache-response power between clean and fingerprinted weights. Specifically, we run the two weights separately on the same set of calibration documents, pair the corresponding cache rows document by document in the target cache block, and take their difference. We project the difference onto the fixed cache transform, and define g_i^2 as the mean square of fingerprint coordinate i over all valid tokens in these calibration documents. Selection and audit documents are not used in this estimate. Let σ_i^2 denote the error variance outside quantization, and let $\gamma > 0$ denote the minimum signal-to-noise ratio required for sign recovery.

The maximum permitted quantization distortion for a fingerprint coordinate is

$$c_i = \begin{cases} \min\left\{\lambda_i, \dfrac{g_i^2}{\gamma} - \sigma_i^2\right\}, & i \in \mathcal{F}, \\ \lambda_i, & i \in \mathcal{I} \setminus \mathcal{F}. \end{cases} \tag{10}$$

For any $i \in \mathcal{F}$, if $g_i^2/\gamma - \sigma_i^2 \le 0$, the error outside quantization already exceeds the recovery range for that coordinate, so the current configuration does not satisfy the detection requirement.

Under a Gaussian scalar approximation, our method jointly allocates continuous distortion across all cache blocks by solving

$$\begin{aligned} \min_{\{D_i\}_{i\in\mathcal{I}}} \quad & \sum_{i\in\mathcal{I}} \frac{1}{2} \log_2 \frac{\lambda_i}{D_i} \\ \text{s.t.} \quad & \sum_{i\in\mathcal{I}} w_i D_i \le D_{\mathrm{T}}, \qquad 0 < D_i \le c_i, \quad i \in \mathcal{I}. \end{aligned} \tag{11}$$

The objective approximates the sum of the rates required by the individual coordinates. The global constraint limits the effect of cache-reconstruction error on language-modeling quality, while the coordinate-wise caps c_i impose both the prediction-residual variance and the fingerprint-recovery requirements.

When $\sum_{i\in\mathcal{I}} w_i c_i \le D_{\mathrm{T}}$, the global task-distortion constraint is inactive and the optimal allocation is $D_i^\star = c_i$. Otherwise, there is a unique $\nu > 0$ such that

$$D_i^\star = \min\left\{c_i, \frac{1}{2\nu w_i \ln 2}\right\}, \qquad \sum_{i\in\mathcal{I}} w_i D_i^\star = D_{\mathrm{T}}. \tag{12}$$

This result follows from the optimality conditions of Eq. 11. The left-hand side is monotonically decreasing in ν, so ν can be determined by one-dimensional bisection. Larger w_i produce smaller distortion targets. Fingerprint coordinates are also constrained by Eq. 10 and may therefore receive higher quantization precision even when their task sensitivity is low.

The continuous distortion targets $D_i^\star$ must next be converted into executable integer bit widths and quantization step sizes. Let b_i and Δ_i denote the bit width and quantization step size for coordinate i, respectively. We search candidate configurations from low to high bit widths and search candidate quantization step sizes at each bit width. For each

candidate, calibration documents are encoded and reconstructed token by token according to Eq. 9; fingerprint coordinates additionally use the QIM offset in Eq. 3. Throughout this process, the reconstructed coordinates at the previous time step are used as the prediction state for the next time step, so that the actual reconstruction error of the full prediction and quantization process is measured.

Let $\widehat{D}_i$ be the closed-loop mean-squared error measured for a candidate configuration on calibration documents, and let $\rho \geq 0$ be the allowed numerical tolerance. We select the smallest available bit width and its corresponding quantization step size satisfying

$$\widehat{D}_i \leq (1+\rho) D_i^\star, \qquad i \in \mathcal{I}, \tag{13}$$

and additionally verify

$$\sum_{i \in \mathcal{I}} w_i \widehat{D}_i \leq (1+\rho) D_{\mathrm{T}}. \tag{14}$$

If no candidate quantizer within the given bit-width range satisfies Eq. 13, we reject the current configuration. Thus, the continuous optimization provides the distortion target for each coordinate, and closed-loop calibration determines the bit width and quantization step size used in practice.

C.4 Fingerprint Recovery, Bias Calibration, and Sequential Ownership Audit

This section supplements the specific arrangement in the main text, the calibration of document scores, and the computation of sequential-test thresholds.

Construction of the propagation matrix. Let j index a fingerprint symbol, and let a_j and b_j denote the j-th columns of the input carrier A and output carrier B, respectively. For the actual projection input h_t at token position t, the Value response of a unit symbol is $\phi_{t,j} = \alpha(a_j^\top h_t) b_j$. For explicit Key and Value caches, we construct a response vector with the same dimension as the target logical cache row, set the Key coordinates to zero, and set the Value coordinates to $\phi_{t,j}$. We concatenate these vectors in the same token order as $d = \mathrm{vec}(Y - Y_0)$ to obtain column j of the propagation matrix Φ.

When Φ has full column rank, we obtain the sign estimate $\widehat{s}$ from Eq. 4 in the main text. Its j-th component $\widehat{s}_j$ is recovered as $+1$ when nonnegative and as -1 otherwise. The ownership score is $\tau = r^{-1} s^\top \widehat{s}$, and subsequent decisions use this continuous score.

QIM bias calibration. Because QIM selects shifted quantization grid points according to the fingerprint vector, a compressed cache produced with clean weights may also exhibit a sign-dependent bias. On independent calibration documents, we compute document scores separately under the clean-weight hypothesis H_0 and the fingerprinted-weight hypothesis H_1. Both use the same calibration documents, cache representation, causal predictor, bit widths, quantization step sizes, and QIM offsets, as well as the same clean uncompressed cache as the recovery reference. Thus, the score under H_0 includes the bias introduced by cache processing and QIM, whereas the score under H_1 additionally includes the response produced by the weight fingerprint.

Let $k \in \{0, 1\}$ index these two hypotheses. We fit one-dimensional Gaussian score distributions:

$$\tau \mid H_k \sim \mathcal{N}(\mu_k, \sigma_k^2), \qquad k \in \{0, 1\}, \tag{15}$$

where μ_k and σ_k^2 are the mean and variance of the corresponding calibration scores, respectively, and $\mathcal{N}$ denotes a Gaussian distribution. The probability densities under the two hypotheses are denoted by p_0 and p_1, respectively. To avoid degenerate densities caused by a finite number of calibration samples, the variance estimates use a prespecified positive lower bound. All distribution parameters are fixed before audit documents are enabled.

Sequential test and thresholds. For independently arriving audit documents, we accumulate the log-likelihood ratio Λ_m according to Eq. 5 in the main text, where m is the

number of observed documents. Each document provides one statistical observation: a positive log-likelihood-ratio increment supports H_1, and a negative increment supports H_0.

Let $\epsilon_{\mathrm{FP}} \in (0,1)$ and $\epsilon_{\mathrm{FN}} \in (0,1)$ denote the prescribed false-positive and false-negative probabilities, respectively. They determine the Wald upper and lower thresholds T_{upper} and T_{lower}:

$$T_{\mathrm{upper}} = \log \frac{1-\epsilon_{\mathrm{FN}}}{\epsilon_{\mathrm{FP}}}, \qquad T_{\mathrm{lower}} = \log \frac{\epsilon_{\mathrm{FN}}}{1-\epsilon_{\mathrm{FP}}}. \tag{16}$$

When $\Lambda_m \geq T_{\mathrm{upper}}$, we accept H_1; when $\Lambda_m \leq T_{\mathrm{lower}}$, we accept H_0. When it lies between the two thresholds, we continue collecting documents. If a prespecified maximum number of documents is reached without crossing either threshold, we report insufficient evidence.

If a single audit tests multiple candidate models, we first allocate the overall false-positive probability among the candidate models and then determine each model's ϵ_{FP} and T_{upper}. Calibration documents remain separate from audit documents. Sequential decisions use only the fixed distribution parameters and subsequently arriving audit scores. The following two examples illustrate the distinction between binary sign recovery and continuous scores, and how same-configuration calibration affects the evidence provided by a single document.

Example: binary sign recovery. Suppose the fingerprint vector is

$$s = (+1, -1, +1, -1)^\top,$$

and an audit document yields the least-squares estimate

$$\widehat{s} = (0.84, -0.71, -0.18, -0.63)^\top.$$

With zero as the decision threshold, nonnegative estimates are recovered as $+1$ and negative estimates as -1, so the recovered result for this document is

$$\widetilde{s} = (+1, -1, -1, -1)^\top.$$

The third symbol disagrees with the fingerprint vector. If the symbol error rate is reported for this document, its BER is $1/4 = 0.25$. The ownership score for the same document is

$$\tau = \frac{1}{4} s^\top \widehat{s} = \frac{0.84 + 0.71 - 0.18 + 0.63}{4} = 0.50.$$

The signs of the symbol estimates determine the recovered result, while their magnitudes are reflected by τ as the strength of the evidence provided by the document. The subsequent sequential test uses the continuous score τ and combines multiple independent documents to make the ownership decision.

Example: QIM bias calibration. Suppose same-configuration calibration gives the following document-score distributions for clean weights H_0 and fingerprinted weights H_1:

$$\tau \mid H_0 \sim \mathcal{N}(0.20, 0.20^2), \qquad \tau \mid H_1 \sim \mathcal{N}(0.80, 0.20^2).$$

Under H_0, the positive mean 0.20 indicates that QIM has already introduced a sign-dependent bias into the compressed cache of the clean weights. Same-configuration calibration incorporates this bias into the score distribution for H_0.

If an audit document produces the score $\tau = 0.55$, its log-likelihood-ratio increment is

$$\log \frac{p_1(0.55)}{p_0(0.55)} = \frac{(0.55-0.20)^2 - (0.55-0.80)^2}{2 \times 0.20^2} = 0.75.$$

This document provides positive evidence supporting H_1. If another document produces the score $\tau = 0.25$, then

$$\log \frac{p_1(0.25)}{p_0(0.25)} = -3.75,$$

Although this score is positive, it is closer to the calibrated distribution under H_0 and therefore provides evidence supporting the clean weights. The sequential test accumulates the log-likelihood ratios from all audit documents and makes a decision when the cumulative result reaches a prespecified threshold.

D Metric Definitions and Audit Details

This section defines the metrics used in codec selection, held-out evaluation, and sequential ownership auditing. The notation follows the current method. The fingerprint contains r physical symbols, $s \in \{-1, +1\}^r$ is the keyed target vector, H_0 denotes the clean-weight route, and H_1 denotes the fingerprinted-weight route. A held-out document is processed with the same matched codec configuration in the two routes, and the resulting scores are compared at the document level.

Bit-error rate and exact recovery. For held-out document d and physical fingerprint symbol $j \in [r]$, let $S_{d,j}^{(1)}$ denote the real-valued per-symbol score obtained under the fingerprinted hypothesis H_1. In the least-squares implementation, this score is the j-th component of the estimate $\widehat{s}_d$ obtained from the document-specific response matrix Φ_d. The decoded symbol is obtained by thresholding this score at zero:

$$\widehat{s}_{d,j} = \mathrm{sign}_+(S_{d,j}^{(1)}), \qquad \mathrm{sign}_+(x) = \begin{cases} +1, & x \geq 0, \\ -1, & x < 0. \end{cases}$$

For N_{test} held-out documents, the bit-error rate is

$$\mathrm{BER} = \frac{\sum_{d=1}^{N_{\text{test}}} \sum_{j=1}^{r} \mathbf{1}[\widehat{s}_{d,j} \neq s_j]}{N_{\text{test}} r}. \tag{17}$$

BER is therefore computed over all document–symbol decoding events. It is different from the fraction of documents that are recovered completely. The exact-recovery rate, reported when applicable, is

$$\mathrm{ExactRecovery} = \frac{1}{N_{\text{test}}} \sum_{d=1}^{N_{\text{test}}} \mathbf{1}[\widehat{s}_{d,j} = s_j \text{ for every } j \in [r]]\,. \tag{18}$$

Document-level ownership score. The sequential ownership test uses a continuous document-level score rather than the binary recovery vector alone. Let $S_{d,j}^{(h)}$ denote the matched per-symbol score for document d, symbol j, and hypothesis H_h, where $h \in \{0, 1\}$. The signed document score is

$$\tau_d^{(h)} = \frac{1}{r} \sum_{j=1}^{r} S_{d,j}^{(h)} s_j. \tag{19}$$

For the least-squares recovery in the main text, this reduces under the fingerprinted route to

$$\tau_d^{(1)} = \frac{1}{r} s^\top \widehat{s}_d.$$

The signs of the components determine the recovered symbols, whereas their magnitudes contribute to $\tau_d^{(h)}$ and therefore determine how much evidence the document contributes to the sequential test. The calibration stage fits score distributions under both H_0 and H_1, and the audit accumulates the log-likelihood ratios of the resulting document scores.

Detection scores and empirical SNR. For each held-out document d and protected symbol j, let $S_{d,j}^{(0)}$ and $S_{d,j}^{(1)}$ be the matched per-symbol scores from the clean and fingerprinted routes, respectively. In the reported evaluation, the two hypotheses use the same number of aligned held-out documents. The document-level empirical mean for hypothesis H_h is

$$\overline{S}_{h,j} = \frac{1}{N_{\text{test}}} \sum_{d=1}^{N_{\text{test}}} S_{d,j}^{(h)}, \qquad h \in \{0, 1\}, \quad j \in [r]. \tag{20}$$

The empirical linear SNR for physical symbol j is

$$\widehat{\mathrm{SNR}}_j = \frac{\left(\overline{S}_{1,j} - \overline{S}_{0,j}\right)^2}{\widehat{\mathrm{Var}}_{H_1}(S_j) + \sigma^2_{\text{audit},j}}, \tag{21}$$

where the variance is computed from the individual document scores under H_1:

$$\widehat{\mathrm{Var}}_{H_1}(S_j) = \frac{1}{N_{\mathrm{test}} - 1} \sum_{d=1}^{N_{\mathrm{test}}} \left(S_{d,j}^{(1)} - \overline{S}_{1,j} \right)^2 . \tag{22}$$

The implementation uses the unbiased sample variance of the document-level scores, with denominator $N_{\mathrm{test}} - 1$. This variance normalizes the squared difference between the mean score under H_1 and the mean score under H_0. The resulting SNR measures the separation between the fingerprinted and clean score populations relative to the observed document-level variability.

The term $\sigma^2_{\mathrm{audit},j}$ denotes additional score-domain variance reserved for a specified audit nuisance or attack. It is distinct from the empirical document-score variance and from the allocation noise σ_i^2 used in the coordinate-wise distortion constraint. The final configurations set $\sigma^2_{\mathrm{audit},j} = 0$, because no additional attack or nuisance variance is registered in the reported evaluation. In the implementation, the denominator is clamped to 10^{-12} before division; this numerical floor prevents a zero denominator when the empirical variance is zero. We report

$$\text{Min. SNR} = \min_{j \in [r]} \widehat{\mathrm{SNR}}_j . \tag{23}$$

The numerator uses the difference between the fingerprinted-route and clean-route means, while the denominator uses the fingerprinted-route variance specified above.

Detection threshold and coordinate constraint. The SNR threshold is fixed before held-out test evaluation from the registered exact-payload failure budget. Let ρ be the allowed probability that at least one of the r independently protected symbols fails. The associated per-symbol failure probability is

$$p_{\mathrm{sym}} = 1 - (1 - \rho)^{1/r}, \tag{24}$$

and the corresponding Gaussian linear SNR threshold is

$$\gamma = \left[F_{\mathrm{N}}^{-1}(1 - p_{\mathrm{sym}}) \right]^2 , \tag{25}$$

where F_{N}^{-1} is the standard-normal quantile function. The same dimensionless gate is applied to every protected coordinate. It is the γ appearing in the coordinate constraint

$$D_i + \sigma_i^2 \le g_i^2 / \gamma, \qquad i \in \mathcal{F},$$

where D_i is quantization MSE, σ_i^2 is the variance of errors outside quantization, and g_i^2 is the calibrated fingerprint-response power. This connects the reported per-symbol SNR metric to the distortion caps used during codec allocation.

For the final configuration with $r = 2$ physical symbols and $\rho = 0.1$,

$$p_{\mathrm{sym}} = 1 - \sqrt{0.9}, \qquad \gamma = \left[F_{\mathrm{N}}^{-1}\left(\sqrt{0.9}\right) \right]^2 = 2.6641381772.$$

Perplexity and log-perplexity change. Let $q \in \{F_1, C_1\}$ denote the fingerprinted compressed route and the matched clean compressed route, respectively. For document d, let $\ell_{q,d,t}$ denote the next-token negative log-likelihood at scored token t, and let T_d be the number of scored tokens in that document. The route-level perplexity is computed from the aggregate token loss:

$$\mathrm{PPL}_q = \exp\left(\frac{\sum_{d=1}^{N_{\mathrm{test}}} \sum_{t=1}^{T_d} \ell_{q,d,t}}{\sum_{d=1}^{N_{\mathrm{test}}} T_d} \right). \tag{26}$$

Equivalently, if L_q is the summed next-token negative log-likelihood and N_{tok} is the total number of scored tokens, then $\mathrm{PPL}_q = \exp(L_q / N_{\mathrm{tok}})$. The paired quality change is

$$\Delta \log \mathrm{PPL} = \log \mathrm{PPL}_{F_1} - \log \mathrm{PPL}_{C_1} . \tag{27}$$

The pairing compares the two routes on the same held-out documents. When an uncertainty interval is reported, documents are resampled with replacement 10,000 times while preserving the clean–fingerprinted pairing within each resample. The reported 95% upper confidence bound is the upper endpoint of the resulting two-sided 95% percentile bootstrap interval, denoted as $95\%UCB$. This is a quality diagnostic for the paired route comparison.

Storage quantities and compression ratios. The codec first selects a precision configuration that satisfies the task and fingerprint-detection constraints. The resulting deployment representation has a serialized cache-code stream and a shared decoder state. The stream contains the quantized cache indices and the metadata needed to parse, reconstruct, and authenticate the record. The shared state contains the cache transform basis, causal predictor, and fingerprint-related parameters stored before serving. Both parts are included because the cache cannot be reconstructed from the request record alone.

For a fixed codec configuration c, let B_{FP16} be the size of the corresponding uncompressed FP16 logical cache over the evaluated horizon, let $B_{\mathrm{stream}}(c)$ be the serialized cache-code stream, and let B_{ζ_c} be the shared decoder state ζ_c. The total deployment size used by codec selection is

$$\mathrm{Bytes}(c) = B_{\mathrm{stream}}(c) + B_{\zeta_c}. \tag{28}$$

The corresponding transfer/storage ratio is

$$R_{\mathrm{transfer}}(c) = \frac{B_{\mathrm{FP16}}}{\mathrm{Bytes}(c)}. \tag{29}$$

Because the FP16 reference is the numerator, $R_{\mathrm{transfer}}(c) > 1$ means $\mathrm{Bytes}(c) < B_{\mathrm{FP16}}$: the complete compressed deployment representation occupies fewer bytes than the corresponding uncompressed logical cache. The ratio is used to compare the total deployment size of the codec configuration with that FP16 reference.

For serving-time storage, let $B^{\max}_{\mathrm{clean}}$ be the peak size of the clean logical cache and let $B^{\max}_{\mathrm{active}}(c)$ be the peak per-request state retained by the active codec. We define

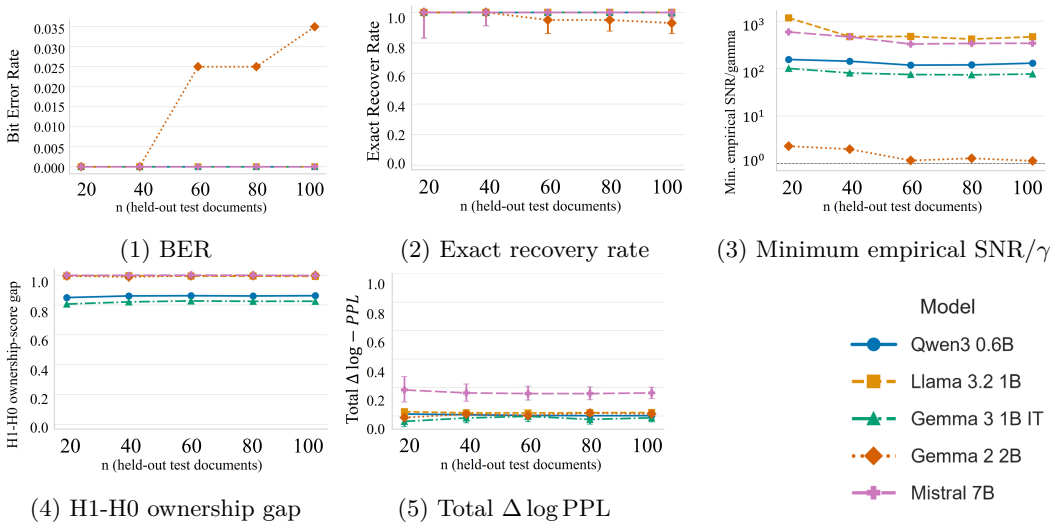
$$R_{\mathrm{active}}(c) = \frac{B^{\max}_{\mathrm{clean}}}{B^{\max}_{\mathrm{active}}(c) + B_{\zeta_c}}. \tag{30}$$

This ratio measures the state retained during serving. Metadata needed to parse and authenticate the request record is already included in $B_{\mathrm{stream}}(c)$ and is counted once.

E Additional Experimental Results

E.1 Sensitivity to Held-Out Test Size

We evaluate whether held-out test size affects fingerprint detection and task quality. Across six models, we fix the selection, calibration, and all remaining FinCa parameters, while increasing the cumulative independent WikiText-2 test set to $n \in \{20, 40, 60, 80, 100\}$. As shown in Fig. 5, all models achieve zero BER and perfect exact recovery at $n \leq 40$. At $n = 60$–100, BER reaches at most 0.035 and exact recovery remains at least 0.93. The H1-H0 ownership gap stays positive and the minimum empirical SNR/γ remains above one throughout. Total $\Delta \log \mathrm{PPL}$ shows no persistent degradation as n increases. Thus, larger test sets reveal limited detection errors while preserving overall ownership evidence and generation quality.



(1) BER (2) Exact recovery rate (3) Minimum empirical SNR/γ

(4) H1-H0 ownership gap (5) Total $\Delta \log \mathrm{PPL}$

Figure 5: Sensitivity of FinCa detection and quality metrics to cumulative held-out test size.

E.2 Details of Best-Parameter Search and Final Validation across Ten Models

All ten models completed their independent final trials with frozen parameters. The parameter search used a sequential staged procedure. The selected Transformer layer is denoted by $\ell^\star$, the embedding strength by α, the normalized task-budget ratio by η_{T}, the number of task-sensitive directions by q, the predictor clipping bound by $\mu_{\max}$, and the QIM offset fraction by β. The fingerprint length was fixed at $r = 2$ for all models.

At the beginning of each model search, the base configuration used $\alpha = 0.01$, $\eta_{\mathrm{T}} = 0.001$, $q = 8$, $\mu_{\max} = 0.98$, and $\beta = 0.25$. The search varied one parameter group at a time, except for the joint budget–rank stage. After each stage, the selected parameters were inherited by all subsequent stages. Parameters not included in the current stage therefore remained at their base values or at the values selected by preceding stages.

The search therefore contains $5 + 4 + (4 \times 3) + 4 + 4 = 29$ validation trials per model. The "Search Trials" column reports completed trials over scheduled trials, and all models completed 29/29 trials.

The following tables report the parameter-sweep trials conducted for each model. The parameter records use the implementation field names *layer*, *alpha*, *task budget ratio*, *task rank*, *predictor clip*, and *QIM offset fraction*; these correspond respectively to $\ell^\star$, α, η_{T}, q,

Table 7: Control structure of the staged parameter search.

Stage	Swept parameters	Settings for the remaining parameters
Layer	$\ell^\star$: five model-dependent candidate layers	$\alpha = 0.01$, $\eta_{\mathrm{T}} = 0.001$, $q = 8$, $\mu_{\max} = 0.98$, and $\beta = 0.25$ use the base configuration.
Embedding strength	$\alpha \in \{0.0025, 0.005, 0.01, 0.02\}$	The layer selected in the preceding stage is retained. $\eta_{\mathrm{T}} = 0.001$, $q = 8$, $\mu_{\max} = 0.98$, and $\beta = 0.25$ remain fixed.
Budget–rank	$\eta_{\mathrm{T}} \in \{0.0005, 0.001, 0.002, 0.004\}$ and $q \in \{4, 8, 16\}$	The selected $\ell^\star$ and α are retained. The 4×3 Cartesian product gives 12 trials. $\mu_{\max} = 0.98$ and $\beta = 0.25$ remain fixed.
Predictor	$\mu_{\max} \in \{0.90, 0.95, 0.98, 0.995\}$	The selected $\ell^\star$, α, η_{T}, and q are retained. The QIM offset remains $\beta = 0.25$.
QIM	$\beta \in \{0, 0.125, 0.25, 0.375\}$	All parameters selected in the preceding stages are retained.

Table 8: Selected frozen parameters.

Model	Layer $\ell^\star$	Strength α	Task-Budget Ratio η_{T}	Task Rank q	Predictor Clip $\mu_{\max}$	QIM Offset β
Qwen3 0.6B	0	0.010	0.004	16	0.900	0.125
Llama 3.2 1B	0	0.020	0.004	4	0.900	0.250
Gemma 3 1B IT	0	0.005	0.004	4	0.950	0.375
Gemma 2 2B	12	0.005	0.004	16	0.980	0.250
Mistral 7B	0	0.0025	0.004	4	0.995	0.250
Ministral 8B	9	0.010	0.004	8	0.995	0.250
Llama 3.1 8B	0	0.0025	0.004	4	0.900	0.375
GLM-4 9B	10	0.010	0.004	8	0.995	0.375
GLM-Z1 9B	20	0.010	0.004	8	0.995	0.250
DeepSeek V2 Lite	13	0.010	0.004	8	0.900	0.125

$\mu_{\max}$, and β. The tables list all trials from the five search stages, followed by the independent final test with the selected parameters.

The reported *transfer* and *active* values are, respectively, the transfer-storage compression ratio and the active logical-cache compression ratio of the FinCa representation.

Our parameter-sweep experiments explore, within finite, empirically motivated ranges, how the principal performance indicators vary with the fingerprint and codec parameters. The final configuration reported in Table.8 is, in our view, a suitable operating point within these ranges; it is not a globally optimal parameter combination, but it provides a favorable trade-off among recovery accuracy, SNR, language-model quality, and compression. The tables include trial-level measurements to make these trends and the basis for the final selection transparent.

Table 9: All successful search trials for Deepseek V2 Lite.

Parameters	BER↓	Min.SNR↑	Gate	Δlog-PPL	95% UCB	Budget	Transfer / active ↑	Tok/s
a=$2.5e^{-3}$	0.00	80.03	pass	$2.2e^{-2}$	$8.7e^{-2}$	$9e^{-1}$	1.57 / $5.2e^{-1}$	4.07
a=$5e^{-3}$	0.00	28.03	pass	$2e^{-2}$	$6.6e^{-2}$	$9e^{-1}$	1.57 / $5.2e^{-1}$	3.60
a=$1e^{-2}$	0.00	38.36	pass	$1e^{-2}$	$8.4e^{-2}$	$9.1e^{-1}$	1.57 / $5.2e^{-1}$	3.59
a=$2e^{-2}$	0.00	152.38	pass	$5.4e^{-2}$	$1.6e^{-1}$	$9e^{-1}$	1.57 / $5.2e^{-1}$	3.64
B=$5e^{-4}$, R=4.00	0.00	38.44	pass	$2.9e^{-2}$	$6.8e^{-2}$	1.20	1.50 / $5.1e^{-1}$	4.90
B=$5e^{-4}$, R=8.00	0.00	35.21	pass	$2.2e^{-2}$	$7.6e^{-2}$	1.20	1.50 / $5.1e^{-1}$	3.66
B=$5e^{-4}$, R=16.00	0.00	34.78	pass	$2.7e^{-2}$	$8.9e^{-2}$	1.21	1.50 / $5.1e^{-1}$	3.58
B=$1e^{-3}$, R=4.00	0.00	37.48	pass	$2.3e^{-2}$	$8.6e^{-2}$	$9e^{-1}$	1.57 / $5.2e^{-1}$	3.66
B=$1e^{-3}$, R=8.00	0.00	38.36	pass	$1e^{-2}$	$8.4e^{-2}$	$9.1e^{-1}$	1.57 / $5.2e^{-1}$	4.30
B=$1e^{-3}$, R=16.00	0.00	34.84	pass	$3.9e^{-2}$	$1.1e^{-1}$	$9.2e^{-1}$	1.57 / $5.2e^{-1}$	3.70
B=$2e^{-3}$, R=4.00	0.00	37.16	pass	$2.4e^{-2}$	$1.3e^{-1}$	$7.4e^{-1}$	1.65 / $5.2e^{-1}$	3.64
B=$2e^{-3}$, R=8.00	0.00	34.81	pass	$6.2e^{-3}$	$9.9e^{-2}$	$7.4e^{-1}$	1.65 / $5.2e^{-1}$	3.62
B=$2e^{-3}$, R=16.00	0.00	35.07	pass	$9.8e^{-3}$	$1.1e^{-1}$	$7.4e^{-1}$	1.65 / $5.2e^{-1}$	3.62
B=$4e^{-3}$, R=4.00	0.00	34.47	pass	$3.1e^{-2}$	$1.3e^{-1}$	$6.1e^{-1}$	1.72 / $5.3e^{-1}$	4.28
B=$4e^{-3}$, R=8.00	0.00	32.84	pass	$1.3e^{-2}$	$1.3e^{-1}$	$6.2e^{-1}$	1.72 / $5.3e^{-1}$	3.61
B=$4e^{-3}$, R=16.00	0.00	35.96	pass	$4.8e^{-2}$	$1.5e^{-1}$	$6.1e^{-1}$	1.72 / $5.3e^{-1}$	3.90
L=0.00	0.00	38.36	pass	$1e^{-2}$	$8.4e^{-2}$	$9.1e^{-1}$	1.57 / $5.2e^{-1}$	3.73
L=6.00	0.00	6.01	pass	$3e^{-2}$	$1.2e^{-1}$	$9e^{-1}$	1.57 / $5.2e^{-1}$	3.63
L=13.00	0.00	11.61	pass	$3.2e^{-2}$	$1.2e^{-1}$	$9e^{-1}$	1.57 / $5.2e^{-1}$	3.63
L=20.00	0.00	36.50	pass	$3.8e^{-2}$	$1.2e^{-1}$	$9e^{-1}$	1.57 / $5.2e^{-1}$	4.01
L=26.00	0.00	54.15	pass	$3.2e^{-2}$	$1.1e^{-1}$	$8.8e^{-1}$	1.57 / $5.2e^{-1}$	3.61
p=$9e^{-1}$	0.00	34.47	pass	$3.1e^{-2}$	$1.3e^{-1}$	$6.1e^{-1}$	1.72 / $5.3e^{-1}$	3.60
p=$9.5e^{-1}$	0.00	34.47	pass	$3.1e^{-2}$	$1.3e^{-1}$	$6.1e^{-1}$	1.72 / $5.3e^{-1}$	3.96
p=$9.8e^{-1}$	0.00	34.47	pass	$3.1e^{-2}$	$1.3e^{-1}$	$6.1e^{-1}$	1.72 / $5.3e^{-1}$	3.66
p=$9.9e^{-1}$	0.00	34.47	pass	$3.1e^{-2}$	$1.3e^{-1}$	$6.1e^{-1}$	1.72 / $5.3e^{-1}$	3.59
f=0.00	0.00	32.82	pass	$2.8e^{-2}$	$1.4e^{-1}$	$6.1e^{-1}$	1.72 / $5.3e^{-1}$	3.65
f=$1.2e^{-1}$	0.00	12.61	pass	$4.1e^{-2}$	$1.4e^{-1}$	$6.2e^{-1}$	1.72 / $5.3e^{-1}$	4.13
f=$2.5e^{-1}$	0.00	34.47	pass	$3.1e^{-2}$	$1.3e^{-1}$	$6.1e^{-1}$	1.72 / $5.3e^{-1}$	4.05
f=$3.8e^{-1}$	0.00	21.16	pass	$2.5e^{-2}$	$1.1e^{-1}$	$6.1e^{-1}$	1.72 / $5.3e^{-1}$	3.60

Table 10: All successful search trials for Gemma2 2B.

Parameters	BER↓	Min.SNR↑	Gate	Δlog-PPL	95% UCB	Budget	Transfer / active ↑	Tok/s
a=$2.5e^{-3}$	0.00	2.09	fail	$-1e^{-1}$	$-6.2e^{-2}$	1.84	2.10 / 1.07	10.43
a=$5e^{-3}$	0.00	9.47	pass	$-1.1e^{-1}$	$-6.9e^{-2}$	1.84	2.10 / 1.07	10.45
a=$1e^{-2}$	0.00	93.41	pass	$-9e^{-2}$	$-3.4e^{-2}$	1.83	2.10 / 1.07	9.67
a=$2e^{-2}$	0.00	658.02	pass	$-8.1e^{-2}$	$-7e^{-2}$	1.79	2.10 / 1.07	9.76
B=$5e^{-4}$, R=4.00	0.00	16.90	pass	$-1e^{-1}$	$-4.6e^{-2}$	2.79	1.96 / 1.04	10.39
B=$5e^{-4}$, R=8.00	0.00	17.47	pass	$-1.7e^{-1}$	$-1.3e^{-1}$	2.88	1.96 / 1.04	10.45
B=$5e^{-4}$, R=16.00	0.00	20.38	pass	$-1.3e^{-1}$	$-9.6e^{-2}$	2.72	1.96 / 1.04	6.24
B=$1e^{-3}$, R=4.00	0.00	14.76	pass	$-8.2e^{-2}$	$-5.8e^{-2}$	1.78	2.10 / 1.07	6.18
B=$1e^{-3}$, R=8.00	0.00	9.47	pass	$-1.1e^{-1}$	$-6.9e^{-2}$	1.84	2.10 / 1.07	5.90
B=$1e^{-3}$, R=16.00	0.00	11.60	pass	$-2.5e^{-2}$	$5e^{-2}$	1.81	2.10 / 1.07	5.63
B=$2e^{-3}$, R=4.00	0.00	7.50	pass	$-1.8e^{-2}$	$5.2e^{-2}$	1.27	2.25 / 1.11	10.29
B=$2e^{-3}$, R=8.00	0.00	10.89	pass	$-8.6e^{-2}$	$-1.9e^{-2}$	1.28	2.25 / 1.11	10.87
B=$2e^{-3}$, R=16.00	0.00	10.20	pass	$-9.6e^{-2}$	$-4.3e^{-2}$	1.28	2.25 / 1.11	10.51
B=$4e^{-3}$, R=4.00	0.00	14.32	pass	$-1.2e^{-1}$	$-2.7e^{-2}$	$9e^{-1}$	2.42 / 1.15	10.44
B=$4e^{-3}$, R=8.00	0.00	17.64	pass	$3.2e^{-3}$	$7.1e^{-2}$	$9.1e^{-1}$	2.42 / 1.15	10.36
B=$4e^{-3}$, R=16.00	0.00	11.14	pass	$-9.3e^{-2}$	$2.6e^{-2}$	$9.4e^{-1}$	2.42 / 1.15	10.42
L=0.00	0.00	5654.83	pass	$-1.9e^{-1}$	$-1.5e^{-1}$	1.76	2.10 / 1.07	10.51
L=6.00	0.00	43.14	pass	$-1.1e^{-1}$	$-4.6e^{-2}$	1.81	2.10 / 1.07	10.68
L=13.00	0.00	61.35	pass	$-9.6e^{-2}$	$-2.8e^{-2}$	1.76	2.10 / 1.07	9.21
L=19.00	0.00	118.97	pass	$-1.7e^{-1}$	$-1.2e^{-1}$	1.74	2.10 / 1.07	9.16
L=25.00	0.00	93.41	pass	$-9e^{-2}$	$-3.4e^{-2}$	1.83	2.10 / 1.07	12.72
p=$9e^{-1}$	0.00	6.91	pass	$-6.8e^{-2}$	$9.9e^{-3}$	$9.7e^{-1}$	2.42 / 1.15	9.72
p=$9.5e^{-1}$	0.00	11.14	pass	$-9.3e^{-2}$	$2.6e^{-2}$	$9.4e^{-1}$	2.42 / 1.15	9.63
p=$9.8e^{-1}$	0.00	11.14	pass	$-9.3e^{-2}$	$2.6e^{-2}$	$9.4e^{-1}$	2.42 / 1.15	10.25
p=$9.9e^{-1}$	0.00	11.14	pass	$-9.3e^{-2}$	$2.6e^{-2}$	$9.4e^{-1}$	2.42 / 1.15	10.25
f=$1.2e^{-1}$	0.00	11.83	pass	$-8.9e^{-2}$	$2.5e^{-2}$	$9.4e^{-1}$	2.42 / 1.15	10.27
f=$2.5e^{-1}$	0.00	11.14	pass	$-9.3e^{-2}$	$2.6e^{-2}$	$9.4e^{-1}$	2.42 / 1.15	10.46
f=$3.8e^{-1}$	0.00	10.96	pass	$-9.2e^{-2}$	$2.6e^{-2}$	$9.4e^{-1}$	2.42 / 1.15	10.44
f=0.00	0.00	11.97	pass	$-9e^{-2}$	$2.8e^{-2}$	$9.4e^{-1}$	2.42 / 1.15	10.29

Table 11: All successful search trials for Gemma3 1B IT.

Parameters	BER↓	Min.SNR↑	Gate	Δlog-PPL	95% UCB	Budget	Transfer / active ↑	Tok/s
a=$2.5e^{-3}$	0.00	71.64	pass	$3.3e^{-2}$	$1.5e^{-1}$	2.00	1.53 / $5.3e^{-1}$	10.81
a=$5e^{-3}$	0.00	829.23	pass	$1.2e^{-1}$	$2.4e^{-1}$	1.98	1.53 / $5.3e^{-1}$	10.76
a=$1e^{-2}$	0.00	1875.18	pass	$1.2e^{-1}$	$2.9e^{-1}$	1.96	1.53 / $5.3e^{-1}$	10.57
a=$2e^{-2}$	0.00	2776.70	pass	$7.9e^{-2}$	$1.5e^{-1}$	1.88	1.53 / $5.3e^{-1}$	10.61
B=$5e^{-4}$, R=4.00	0.00	241.48	pass	$1.3e^{-1}$	$1.9e^{-1}$	3.34	1.46 / $5.3e^{-1}$	9.41
B=$5e^{-4}$, R=8.00	0.00	241.48	pass	$1.9e^{-1}$	$3.3e^{-1}$	3.35	1.46 / $5.3e^{-1}$	9.64
B=$5e^{-4}$, R=16.00	0.00	1058.06	pass	$2.3e^{-2}$	$9.5e^{-2}$	2.67	1.46 / $5.3e^{-1}$	10.80
B=$1e^{-3}$, R=4.00	0.00	829.23	pass	$9.6e^{-2}$	$2.2e^{-1}$	1.98	1.53 / $5.3e^{-1}$	10.70
B=$1e^{-3}$, R=8.00	0.00	829.23	pass	$1.2e^{-1}$	$2.4e^{-1}$	1.98	1.53 / $5.3e^{-1}$	10.64
B=$1e^{-3}$, R=16.00	0.00	276.53	pass	$6.8e^{-2}$	$2.2e^{-1}$	1.94	1.53 / $5.3e^{-1}$	10.71
B=$2e^{-3}$, R=4.00	0.00	344.84	pass	$1.4e^{-1}$	$2.5e^{-1}$	1.33	1.61 / $5.4e^{-1}$	10.69
B=$2e^{-3}$, R=8.00	0.00	344.84	pass	$2.2e^{-1}$	$3.1e^{-1}$	1.33	1.61 / $5.4e^{-1}$	10.75
B=$2e^{-3}$, R=16.00	0.00	365.33	pass	$7.9e^{-2}$	$1.9e^{-1}$	1.20	1.61 / $5.4e^{-1}$	10.81
B=$4e^{-3}$, R=4.00	0.00	384.39	pass	$8.1e^{-2}$	$2.1e^{-1}$	$9.4e^{-1}$	1.70 / $5.5e^{-1}$	10.75
B=$4e^{-3}$, R=8.00	0.00	384.39	pass	$1.5e^{-1}$	$3.6e^{-1}$	$9.3e^{-1}$	1.70 / $5.5e^{-1}$	10.44
B=$4e^{-3}$, R=16.00	0.00	345.73	pass	$9.3e^{-2}$	$2.5e^{-1}$	1.02	1.70 / $5.5e^{-1}$	10.46
L=0.00	0.00	1875.18	pass	$1.2e^{-1}$	$2.9e^{-1}$	1.96	1.53 / $5.3e^{-1}$	10.89
L=6.00	0.00	3.20	pass	$3e^{-2}$	$1.1e^{-1}$	1.93	1.53 / $5.3e^{-1}$	10.91
L=13.00	0.00	1.80	fail	$5.3e^{-2}$	$1.8e^{-1}$	1.96	1.53 / $5.3e^{-1}$	10.88
L=19.00	$1e^{-1}$	3.51	pass	$5.8e^{-2}$	$1.4e^{-1}$	1.94	1.53 / $5.3e^{-1}$	10.90
L=25.00	$2e^{-1}$	1.45	fail	$8.5e^{-2}$	$2.1e^{-1}$	1.95	1.53 / $5.3e^{-1}$	11.34
p=$9e^{-1}$	0.00	370.87	pass	$-1.3e^{-1}$	$-2.4e^{-3}$	$9.4e^{-1}$	1.70 / $5.5e^{-1}$	10.79
p=$9.5e^{-1}$	0.00	384.39	pass	$8.1e^{-2}$	$2.1e^{-1}$	$9.4e^{-1}$	1.70 / $5.5e^{-1}$	10.75
p=$9.8e^{-1}$	0.00	384.39	pass	$8.1e^{-2}$	$2.1e^{-1}$	$9.4e^{-1}$	1.70 / $5.5e^{-1}$	10.62
p=$9.9e^{-1}$	0.00	384.39	pass	$8.1e^{-2}$	$2.1e^{-1}$	$9.4e^{-1}$	1.70 / $5.5e^{-1}$	10.69
f=0.00	0.00	184.05	pass	$9.2e^{-2}$	$2.1e^{-1}$	$9.3e^{-1}$	1.70 / $5.5e^{-1}$	10.77
f=$1.2e^{-1}$	0.00	160.73	pass	$1.8e^{-1}$	$4e^{-1}$	$9.4e^{-1}$	1.70 / $5.5e^{-1}$	10.74
f=$2.5e^{-1}$	0.00	384.39	pass	$8.1e^{-2}$	$2.1e^{-1}$	$9.4e^{-1}$	1.70 / $5.5e^{-1}$	10.65
f=$3.8e^{-1}$	0.00	247.11	pass	$9e^{-2}$	$3.5e^{-1}$	$9.3e^{-1}$	1.70 / $5.5e^{-1}$	10.64

Table 12: All successful search trials for GLM4 9B.

Parameters	BER↓	Min.SNR↑	Gate	Δlog-PPL	95% UCB	Budget	Transfer / active ↑	Tok/s
a=$2.5e^{-3}$	$1e^{-1}$	1.56	fail	$-2.2e^{-2}$	$3.1e^{-3}$	2.85	2.11 / 1.19	8.56
a=$5e^{-3}$	$1e^{-1}$	5.58	pass	$-1.3e^{-2}$	$1.7e^{-2}$	2.86	2.11 / 1.19	8.76
a=$1e^{-2}$	0.00	10.05	pass	$-1.5e^{-2}$	$1.2e^{-2}$	2.86	2.11 / 1.19	7.93
a=$2e^{-2}$	0.00	87.85	pass	$-1.4e^{-2}$	$1.6e^{-2}$	2.87	2.11 / 1.19	8.71
B=$5e^{-4}$, R=4.00	0.00	39.84	pass	$-2.9e^{-2}$	$-4.9e^{-3}$	4.71	1.97 / 1.15	8.74
B=$5e^{-4}$, R=8.00	0.00	17.08	pass	$-1.9e^{-2}$	$1.8e^{-3}$	4.70	1.97 / 1.15	8.67
B=$5e^{-4}$, R=16.00	0.00	31.78	pass	$-1.8e^{-2}$	$-1.1e^{-2}$	4.64	1.97 / 1.15	8.81
B=$1e^{-3}$, R=4.00	0.00	23.24	pass	$-6.9e^{-3}$	$1.6e^{-2}$	2.86	2.11 / 1.19	8.68
B=$1e^{-3}$, R=8.00	0.00	10.05	pass	$-1.5e^{-2}$	$1.2e^{-2}$	2.86	2.11 / 1.19	8.69
B=$1e^{-3}$, R=16.00	0.00	37.31	pass	$-1.1e^{-2}$	$1.4e^{-2}$	2.85	2.11 / 1.19	8.27
B=$2e^{-3}$, R=4.00	0.00	7.29	pass	$-1.5e^{-2}$	$2e^{-2}$	1.80	2.27 / 1.24	8.34
B=$2e^{-3}$, R=8.00	0.00	17.88	pass	$-8.1e^{-3}$	$2e^{-2}$	1.80	2.27 / 1.24	8.30
B=$2e^{-3}$, R=16.00	0.00	6.17	pass	$-4.6e^{-3}$	$3.2e^{-2}$	1.80	2.27 / 1.24	8.34
B=$4e^{-3}$, R=4.00	$1e^{-1}$	3.38	pass	$-7.2e^{-3}$	$4.3e^{-2}$	1.25	2.45 / 1.29	8.32
B=$4e^{-3}$, R=8.00	0.00	5.61	pass	$-3.7e^{-3}$	$4.1e^{-2}$	1.26	2.45 / 1.29	8.32
B=$4e^{-3}$, R=16.00	$1e^{-1}$	3.30	pass	$-7e^{-3}$	$3.5e^{-2}$	1.26	2.45 / 1.29	8.78
L=0.00	0.00	50.81	pass	$-1.2e^{-2}$	$1.2e^{-2}$	2.89	2.11 / 1.19	8.61
L=10.00	0.00	10.05	pass	$-1.5e^{-2}$	$1.2e^{-2}$	2.86	2.11 / 1.19	8.66
L=20.00	0.00	30.11	pass	$-1.9e^{-2}$	$4.5e^{-3}$	2.90	2.11 / 1.19	8.65
L=30.00	0.00	13.48	pass	$5e^{-3}$	$2.7e^{-2}$	2.88	2.11 / 1.19	8.00
L=39.00	0.00	25.23	pass	$-7.1e^{-3}$	$1.3e^{-2}$	2.86	2.11 / 1.19	8.58
p=$9e^{-1}$	0.00	4.58	pass	$1.4e^{-2}$	$9.1e^{-2}$	1.26	2.44 / 1.29	8.71
p=$9.5e^{-1}$	0.00	3.12	pass	$-6.6e^{-3}$	$7.9e^{-2}$	1.25	2.44 / 1.29	6.96
p=$9.8e^{-1}$	0.00	5.61	pass	$-3.7e^{-3}$	$4.1e^{-2}$	1.26	2.45 / 1.29	8.68
p=$9.9e^{-1}$	0.00	3.17	pass	$-1.2e^{-2}$	$3.7e^{-2}$	1.27	2.45 / 1.29	8.69
f=0.00	0.00	3.80	pass	$-5e^{-3}$	$5e^{-2}$	1.26	2.45 / 1.29	8.69
f=$1.2e^{-1}$	0.00	3.78	pass	$9.4e^{-3}$	$7.6e^{-2}$	1.26	2.45 / 1.29	8.66
f=$2.5e^{-1}$	0.00	3.17	pass	$-1.2e^{-2}$	$3.7e^{-2}$	1.27	2.45 / 1.29	8.64
f=$3.8e^{-1}$	0.00	3.35	pass	$2.5e^{-3}$	$6.6e^{-2}$	1.26	2.45 / 1.29	8.73

Table 13: All successful search trials for GLM Z1 9B.

Parameters	BER↓	Min.SNR↑	Gate	Δlog-PPL	95% UCB	Budget	Transfer / active ↑	Tok/s
a=$2.5e^{-3}$	0.00	$5.1e^{-1}$	fail	$-2.8e^{-2}$	$3.9e^{-2}$	2.44	2.06 / 1.18	8.65
a=$5e^{-3}$	0.00	$7.3e^{-1}$	fail	$-5.5e^{-2}$	$9.1e^{-3}$	2.45	2.06 / 1.18	8.67
a=$1e^{-2}$	0.00	16.76	pass	$-3.7e^{-2}$	$1.8e^{-2}$	2.44	2.06 / 1.18	7.10
a=$2e^{-2}$	0.00	22.79	pass	$-4.3e^{-2}$	$1.7e^{-2}$	2.46	2.06 / 1.18	8.68
B=$5e^{-4}$, R=4.00	0.00	14.66	pass	$-3.9e^{-2}$	$-3.5e^{-3}$	3.96	1.93 / 1.13	8.69
B=$5e^{-4}$, R=8.00	0.00	9.82	pass	$-5.8e^{-2}$	$3e^{-3}$	3.98	1.93 / 1.13	8.65
B=$5e^{-4}$, R=16.00	0.00	7.47	pass	$-7.4e^{-2}$	$-1.2e^{-2}$	3.99	1.93 / 1.13	8.72
B=$1e^{-3}$, R=4.00	0.00	5.58	pass	$-6.3e^{-2}$	$-3.5e^{-4}$	2.44	2.06 / 1.18	8.48
B=$1e^{-3}$, R=8.00	0.00	16.76	pass	$-3.7e^{-2}$	$1.8e^{-2}$	2.44	2.06 / 1.18	8.71
B=$1e^{-3}$, R=16.00	0.00	7.25	pass	$-6.8e^{-2}$	$-1.8e^{-2}$	2.46	2.06 / 1.18	8.53
B=$2e^{-3}$, R=4.00	0.00	18.96	pass	$-5.4e^{-2}$	$-1.4e^{-2}$	1.62	2.21 / 1.22	8.68
B=$2e^{-3}$, R=8.00	0.00	3.66	pass	$-4.9e^{-2}$	$-1.2e^{-2}$	1.61	2.21 / 1.22	8.59
B=$2e^{-3}$, R=16.00	0.00	4.68	pass	$-5.7e^{-2}$	$-1.5e^{-2}$	1.61	2.21 / 1.22	8.67
B=$4e^{-3}$, R=4.00	0.00	2.54	fail	$-6.8e^{-2}$	$-1.1e^{-2}$	1.13	2.38 / 1.27	8.65
B=$4e^{-3}$, R=8.00	0.00	9.29	pass	$-4.7e^{-2}$	$1.5e^{-2}$	1.13	2.38 / 1.27	8.66
B=$4e^{-3}$, R=16.00	0.00	4.81	pass	$5.2e^{-3}$	$5.9e^{-2}$	1.14	2.38 / 1.27	8.70
L=0.00	0.00	1270.56	pass	$-1.6e^{-2}$	$5.5e^{-2}$	2.41	2.06 / 1.18	8.61
L=10.00	0.00	7.93	pass	$7.9e^{-3}$	$5.6e^{-2}$	2.43	2.06 / 1.18	8.02
L=20.00	0.00	40.62	pass	$-3.3e^{-2}$	$9.7e^{-3}$	2.42	2.06 / 1.18	8.53
L=30.00	0.00	287.78	pass	$-6.6e^{-2}$	$-3.2e^{-2}$	2.44	2.06 / 1.18	8.67
L=39.00	0.00	16.76	pass	$-3.7e^{-2}$	$1.8e^{-2}$	2.44	2.06 / 1.18	7.41
p=$9e^{-1}$	0.00	8.02	pass	$-1.1e^{-1}$	$-3.1e^{-2}$	1.17	2.37 / 1.27	8.56
p=$9.5e^{-1}$	$1e^{-1}$	3.01	pass	$-6.8e^{-2}$	$-2.3e^{-2}$	1.16	2.38 / 1.27	7.73
p=$9.8e^{-1}$	0.00	9.29	pass	$-4.7e^{-2}$	$1.5e^{-2}$	1.13	2.38 / 1.27	8.48
p=$9.9e^{-1}$	0.00	3.90	pass	$8.2e^{-4}$	$5.6e^{-2}$	1.14	2.38 / 1.27	8.67
f=0.00	0.00	3.10	pass	$3.2e^{-3}$	$5.6e^{-2}$	1.14	2.38 / 1.27	8.71
f=$2.5e^{-1}$	0.00	3.90	pass	$8.2e^{-4}$	$5.6e^{-2}$	1.14	2.38 / 1.27	8.76
f=$1.2e^{-1}$	0.00	3.75	pass	$5.4e^{-3}$	$5.9e^{-2}$	1.14	2.38 / 1.27	8.52
f=$3.8e^{-1}$	0.00	4.16	pass	$5.3e^{-3}$	$5.9e^{-2}$	1.14	2.38 / 1.27	8.04

Table 14: All successful search trials for Llama 3.1 8B.

Parameters	BER↓	Min.SNR↑	Gate	Δlog-PPL	95% UCB	Budget	Transfer / active ↑	Tok/s
a=$2.5e^{-3}$	0.00	476.04	pass	$3e^{-2}$	$4.9e^{-2}$	2.51	2.31 / 1.46	10.83
a=$5e^{-3}$	0.00	449.27	pass	$2.5e^{-2}$	$3.2e^{-2}$	2.53	2.31 / 1.46	10.96
a=$1e^{-2}$	0.00	11687.93	pass	$3.6e^{-4}$	$1.3e^{-2}$	2.51	2.31 / 1.46	10.72
a=$2e^{-2}$	0.00	17749.57	pass	$1.5e^{-2}$	$3e^{-2}$	2.47	2.31 / 1.46	11.03
B=$5e^{-4}$, R=4.00	0.00	730.15	pass	$-2.4e^{-2}$	$-8.2e^{-4}$	4.22	2.15 / 1.39	10.89
B=$5e^{-4}$, R=8.00	0.00	755.59	pass	$-3.8e^{-2}$	$-9.4e^{-3}$	4.22	2.15 / 1.39	11.03
B=$5e^{-4}$, R=16.00	0.00	1130.19	pass	$-4.8e^{-2}$	$-1.6e^{-2}$	4.26	2.15 / 1.39	11.32
B=$1e^{-3}$, R=4.00	0.00	419.81	pass	$7.6e^{-3}$	$2.7e^{-2}$	2.49	2.31 / 1.46	11.20
B=$1e^{-3}$, R=8.00	0.00	476.04	pass	$3e^{-2}$	$4.9e^{-2}$	2.51	2.31 / 1.46	11.21
B=$1e^{-3}$, R=16.00	0.00	340.18	pass	$1.5e^{-2}$	$4.8e^{-2}$	2.54	2.30 / 1.46	11.13
B=$2e^{-3}$, R=4.00	0.00	331.95	pass	$7.7e^{-3}$	$2.4e^{-2}$	1.62	2.49 / 1.53	9.76
B=$2e^{-3}$, R=8.00	0.00	330.71	pass	$5.2e^{-3}$	$3.8e^{-2}$	1.62	2.49 / 1.53	11.09
B=$2e^{-3}$, R=16.00	0.00	338.19	pass	$2.2e^{-2}$	$4.6e^{-2}$	1.65	2.49 / 1.53	11.12
B=$4e^{-3}$, R=4.00	0.00	223.52	pass	$-2.4e^{-3}$	$4.5e^{-2}$	1.14	2.70 / 1.61	4.91
B=$4e^{-3}$, R=8.00	0.00	225.29	pass	$1.1e^{-2}$	$5.3e^{-2}$	1.15	2.70 / 1.61	4.90
B=$4e^{-3}$, R=16.00	0.00	223.51	pass	$3.3e^{-2}$	$4.9e^{-2}$	1.15	2.69 / 1.60	4.89
L=0.00	0.00	11687.93	pass	$3.6e^{-4}$	$1.3e^{-2}$	2.51	2.31 / 1.46	11.17
L=8.00	0.00	64.48	pass	$2e^{-2}$	$3.4e^{-2}$	2.68	2.29 / 1.45	11.08
L=16.00	0.00	22.68	pass	$9.5e^{-3}$	$2.8e^{-2}$	2.67	2.29 / 1.45	11.05
L=24.00	$2e^{-1}$	1.33	fail	$1.2e^{-2}$	$2.5e^{-2}$	2.66	2.29 / 1.45	11.41
L=31.00	0.00	4.48	pass	$1.1e^{-2}$	$3.4e^{-2}$	2.66	2.29 / 1.45	11.18
p=$9e^{-1}$	0.00	223.26	pass	$-1e^{-2}$	$3e^{-2}$	1.14	2.70 / 1.61	4.85
p=$9.5e^{-1}$	0.00	223.52	pass	$-2.4e^{-3}$	$4.5e^{-2}$	1.14	2.70 / 1.61	4.92
p=$9.8e^{-1}$	0.00	223.52	pass	$-2.4e^{-3}$	$4.5e^{-2}$	1.14	2.70 / 1.61	11.16
p=$9.9e^{-1}$	0.00	223.52	pass	$-2.4e^{-3}$	$4.5e^{-2}$	1.14	2.70 / 1.61	10.98
f=0.00	0.00	160.22	pass	$-7.7e^{-3}$	$5.4e^{-2}$	1.14	2.70 / 1.61	9.59
f=$1.2e^{-1}$	0.00	178.90	pass	$-5.8e^{-3}$	$5.4e^{-2}$	1.14	2.70 / 1.61	10.92
f=$2.5e^{-1}$	0.00	223.26	pass	$-1e^{-2}$	$3e^{-2}$	1.14	2.70 / 1.61	5.41
f=$3.8e^{-1}$	0.00	215.14	pass	$-5.2e^{-3}$	$4.1e^{-2}$	1.14	2.70 / 1.61	10.94

Table 15: All successful search trials for Llama 3.2 1B.

Parameters	BER↓	Min.SNR↑	Gate	Δlog-PPL	95% UCB	Budget	Transfer / active ↑	Tok/s
a=$2.5e^{-3}$	0.00	10.59	pass	$1.8e^{-2}$	$3.3e^{-2}$	2.48	2.24 / 1.43	20.61
a=$5e^{-3}$	0.00	215.55	pass	$9.4e^{-3}$	$4.9e^{-2}$	2.47	2.24 / 1.43	20.61
a=$1e^{-2}$	0.00	628.89	pass	$2.5e^{-2}$	$4.7e^{-2}$	2.50	2.24 / 1.43	20.28
a=$2e^{-2}$	0.00	397.91	pass	$3.8e^{-3}$	$2e^{-2}$	2.48	2.24 / 1.43	19.93
B=$5e^{-4}$, R=4.00	0.00	598.79	pass	$2.8e^{-2}$	$4.4e^{-2}$	4.06	2.09 / 1.37	20.36
B=$5e^{-4}$, R=8.00	0.00	884.73	pass	$3.5e^{-2}$	$5.2e^{-2}$	4.12	2.09 / 1.37	20.39
B=$5e^{-4}$, R=16.00	0.00	706.00	pass	$4.1e^{-2}$	$6.3e^{-2}$	4.18	2.08 / 1.37	21.06
B=$1e^{-3}$, R=4.00	0.00	416.64	pass	$4.7e^{-4}$	$1.4e^{-2}$	2.49	2.24 / 1.44	20.84
B=$1e^{-3}$, R=8.00	0.00	397.91	pass	$3.8e^{-3}$	$2e^{-2}$	2.48	2.24 / 1.43	20.20
B=$1e^{-3}$, R=16.00	0.00	508.00	pass	$6.9e^{-3}$	$2.8e^{-2}$	2.50	2.24 / 1.43	20.39
B=$2e^{-3}$, R=4.00	0.00	516.58	pass	$-3.4e^{-2}$	$1.3e^{-2}$	1.58	2.42 / 1.51	20.61
B=$2e^{-3}$, R=8.00	0.00	777.90	pass	$6.9e^{-3}$	$4.7e^{-2}$	1.60	2.42 / 1.51	20.57
B=$2e^{-3}$, R=16.00	0.00	1193.63	pass	$-6.7e^{-3}$	$3.7e^{-2}$	1.61	2.41 / 1.50	19.41
B=$4e^{-3}$, R=4.00	0.00	1396.64	pass	$-2e^{-2}$	$2.2e^{-2}$	1.13	2.62 / 1.58	18.74
B=$4e^{-3}$, R=8.00	0.00	3074.49	pass	$-1.4e^{-2}$	$2.5e^{-2}$	1.13	2.62 / 1.58	20.58
B=$4e^{-3}$, R=16.00	0.00	932.82	pass	$-6.3e^{-3}$	$1.6e^{-2}$	1.14	2.62 / 1.58	20.55
L=0.00	0.00	628.89	pass	$2.5e^{-2}$	$4.7e^{-2}$	2.50	2.24 / 1.43	20.76
L=4.00	0.00	4.70	pass	$-1.4e^{-2}$	$5.5e^{-3}$	2.56	2.23 / 1.43	20.38
L=8.00	0.00	5.24	pass	$3.5e^{-3}$	$3e^{-2}$	2.60	2.23 / 1.43	20.30
L=12.00	0.00	159.29	pass	$-8.6e^{-3}$	$2.2e^{-2}$	2.58	2.23 / 1.43	20.43
L=15.00	0.00	4.35	pass	$1.6e^{-2}$	$3.5e^{-2}$	2.60	2.23 / 1.43	26.43
p=$9e^{-1}$	0.00	1397.35	pass	$2.2e^{-3}$	$4.7e^{-2}$	1.13	2.62 / 1.58	20.41
p=$9.5e^{-1}$	0.00	1396.64	pass	$-2e^{-2}$	$2.2e^{-2}$	1.13	2.62 / 1.58	20.27
p=$9.8e^{-1}$	0.00	1396.64	pass	$-2e^{-2}$	$2.2e^{-2}$	1.13	2.62 / 1.58	20.93
p=$9.9e^{-1}$	0.00	1396.64	pass	$-2e^{-2}$	$2.2e^{-2}$	1.13	2.62 / 1.58	20.81
f=0.00	0.00	635.81	pass	$9.2e^{-3}$	$5e^{-2}$	1.13	2.62 / 1.58	18.11
f=$1.2e^{-1}$	0.00	312.03	pass	$5.3e^{-3}$	$5.1e^{-2}$	1.13	2.62 / 1.58	17.88
f=$2.5e^{-1}$	0.00	1397.35	pass	$2.2e^{-3}$	$4.7e^{-2}$	1.13	2.62 / 1.58	20.42
f=$3.8e^{-1}$	0.00	620.31	pass	$-3.7e^{-3}$	$3.8e^{-2}$	1.13	2.62 / 1.58	20.36

Table 16: All successful search trials for Ministral 8B.

Parameters	BER↓	Min.SNR↑	Gate	Δlog-PPL	95% UCB	Budget	Transfer / active ↑	Tok/s
a=$2.5e^{-3}$	0.00	$3.4e^{-1}$	fail	$5.4e^{-3}$	$1.3e^{-2}$	2.54	2.26 / 1.45	4.40
a=$5e^{-3}$	0.00	1.18	fail	$5.2e^{-3}$	$1.8e^{-2}$	2.55	2.26 / 1.45	4.39
a=$1e^{-2}$	0.00	5.56	pass	$3.3e^{-3}$	$1.3e^{-2}$	2.55	2.26 / 1.45	4.37
a=$2e^{-2}$	0.00	20.09	pass	$5.5e^{-3}$	$2.1e^{-2}$	2.54	2.26 / 1.45	4.44
B=$5e^{-4}$, R=4.00	0.00	5.85	pass	$1.5e^{-3}$	$6.6e^{-3}$	4.24	2.10 / 1.38	4.32
B=$5e^{-4}$, R=8.00	0.00	5.12	pass	$-4.5e^{-3}$	$3.8e^{-3}$	4.24	2.10 / 1.38	4.40
B=$5e^{-4}$, R=16.00	0.00	5.51	pass	$7.3e^{-4}$	$8.6e^{-3}$	4.24	2.10 / 1.38	4.42
B=$1e^{-3}$, R=4.00	0.00	4.65	pass	$9.1e^{-3}$	$1.9e^{-2}$	2.55	2.26 / 1.45	4.37
B=$1e^{-3}$, R=8.00	0.00	5.56	pass	$3.3e^{-3}$	$1.3e^{-2}$	2.55	2.26 / 1.45	4.33
B=$1e^{-3}$, R=16.00	0.00	5.30	pass	$7.3e^{-3}$	$1.7e^{-2}$	2.56	2.26 / 1.45	4.36
B=$2e^{-3}$, R=4.00	0.00	5.07	pass	$2.8e^{-2}$	$4.3e^{-2}$	1.66	2.44 / 1.52	4.41
B=$2e^{-3}$, R=8.00	0.00	3.35	pass	$1.8e^{-2}$	$3.1e^{-2}$	1.67	2.44 / 1.52	4.34
B=$2e^{-3}$, R=16.00	0.00	3.48	pass	$1.6e^{-2}$	$3.2e^{-2}$	1.67	2.44 / 1.52	4.37
B=$4e^{-3}$, R=4.00	0.00	6.23	pass	$-1.4e^{-2}$	$3.7e^{-2}$	1.18	2.64 / 1.60	4.30
B=$4e^{-3}$, R=8.00	0.00	4.10	pass	$-1.6e^{-2}$	$2.8e^{-2}$	1.18	2.64 / 1.60	4.38
B=$4e^{-3}$, R=16.00	0.00	4.87	pass	$-1e^{-2}$	$2.4e^{-2}$	1.18	2.64 / 1.60	4.34
L=0.00	0.00	17519.59	pass	$-6.7e^{-3}$	$1.1e^{-2}$	2.52	2.26 / 1.45	4.03
L=9.00	0.00	201.08	pass	$-9.1e^{-4}$	$1.4e^{-2}$	2.53	2.26 / 1.45	4.41
L=18.00	0.00	95.58	pass	$2.7e^{-3}$	$1.3e^{-2}$	2.54	2.26 / 1.45	4.43
L=27.00	0.00	5.56	pass	$3.3e^{-3}$	$1.3e^{-2}$	2.55	2.26 / 1.45	4.42
L=35.00	0.00	85.32	pass	$1.2e^{-2}$	$2e^{-2}$	2.54	2.26 / 1.45	4.39
p=$9e^{-1}$	0.00	4.25	pass	$-7.9e^{-3}$	$3.5e^{-2}$	1.18	2.64 / 1.60	4.40
p=$9.5e^{-1}$	0.00	4.10	pass	$-1.6e^{-2}$	$2.8e^{-2}$	1.18	2.64 / 1.60	4.32
p=$9.8e^{-1}$	0.00	4.10	pass	$-1.6e^{-2}$	$2.8e^{-2}$	1.18	2.64 / 1.60	4.38
p=$9.9e^{-1}$	0.00	4.10	pass	$-1.6e^{-2}$	$2.8e^{-2}$	1.18	2.64 / 1.60	4.39
f=0.00	0.00	4.02	pass	$-1.6e^{-2}$	$2.9e^{-2}$	1.18	2.64 / 1.60	4.22
f=$1.2e^{-1}$	0.00	4.09	pass	$-1.6e^{-2}$	$3.1e^{-2}$	1.18	2.64 / 1.60	4.40
f=$2.5e^{-1}$	0.00	4.10	pass	$-1.6e^{-2}$	$2.8e^{-2}$	1.18	2.64 / 1.60	4.45
f=$3.8e^{-1}$	0.00	3.87	pass	$-1.9e^{-2}$	$2.4e^{-2}$	1.18	2.64 / 1.60	4.38

Table 17: All successful search trials for Mistral 7B.

Parameters	BER↓	Min.SNR↑	Gate	Δlog-PPL	95% UCB	Budget	Transfer / active ↑	Tok/s
a=$2.5e^{-3}$	0.00	3306.53	pass	$-1.5e^{-3}$	$2.8e^{-2}$	2.95	2.33 / 1.47	11.96
a=$5e^{-3}$	0.00	3534.96	pass	$-9.2e^{-4}$	$3e^{-2}$	2.92	2.33 / 1.47	11.73
a=$1e^{-2}$	0.00	26268.78	pass	$6e^{-5}$	$2.6e^{-2}$	2.95	2.33 / 1.47	10.16
a=$2e^{-2}$	0.00	146826.61	pass	$-4.7e^{-2}$	$-1.5e^{-2}$	2.90	2.33 / 1.47	11.15
B=$5e^{-4}$, R=4.00	0.00	1239.32	pass	$1.2e^{-2}$	$4.3e^{-2}$	5.08	2.17 / 1.40	11.96
B=$5e^{-4}$, R=8.00	0.00	1215.89	pass	$-5.9e^{-4}$	$3e^{-2}$	5.09	2.17 / 1.40	4.77
B=$5e^{-4}$, R=16.00	0.00	1211.73	pass	$-7.4e^{-3}$	$2.7e^{-2}$	5.17	2.17 / 1.40	4.78
B=$1e^{-3}$, R=4.00	0.00	3295.93	pass	$1.1e^{-2}$	$5.9e^{-2}$	2.95	2.33 / 1.47	4.77
B=$1e^{-3}$, R=8.00	0.00	3306.48	pass	$-1.5e^{-3}$	$2.8e^{-2}$	2.95	2.33 / 1.47	4.86
B=$1e^{-3}$, R=16.00	0.00	3356.01	pass	$9.5e^{-3}$	$5e^{-2}$	2.93	2.33 / 1.47	11.98
B=$2e^{-3}$, R=4.00	0.00	1056.05	pass	$-5.8e^{-2}$	$-1.8e^{-2}$	1.83	2.52 / 1.54	10.70
B=$2e^{-3}$, R=8.00	0.00	1068.58	pass	$-4.6e^{-2}$	$3.1e^{-2}$	1.82	2.52 / 1.54	11.89
B=$2e^{-3}$, R=16.00	0.00	1051.14	pass	$-6.6e^{-2}$	$-4.1e^{-3}$	1.83	2.52 / 1.54	8.62
B=$4e^{-3}$, R=4.00	0.00	45.03	pass	$4.9e^{-2}$	$1.1e^{-1}$	1.25	2.72 / 1.61	3.54
B=$4e^{-3}$, R=8.00	0.00	45.02	pass	$5.1e^{-2}$	$1e^{-1}$	1.26	2.72 / 1.61	3.20
B=$4e^{-3}$, R=16.00	0.00	45.13	pass	$4.8e^{-2}$	$1.1e^{-1}$	1.26	2.72 / 1.61	3.19
L=0.00	0.00	26268.78	pass	$6e^{-5}$	$2.6e^{-2}$	2.95	2.33 / 1.47	5.12
L=8.00	0.00	711.95	pass	$-2.9e^{-2}$	$-3.3e^{-3}$	2.96	2.33 / 1.46	4.90
L=16.00	0.00	78.35	pass	$-7.3e^{-3}$	$2.7e^{-2}$	2.96	2.33 / 1.46	5.18
L=24.00	0.00	18.35	pass	$-2.2e^{-2}$	$9.5e^{-3}$	2.97	2.33 / 1.46	5.12
L=31.00	0.00	249.59	pass	$1.1e^{-2}$	$5.7e^{-2}$	2.96	2.33 / 1.46	10.97
p=$9e^{-1}$	0.00	45.08	pass	$5.6e^{-2}$	$1.2e^{-1}$	1.25	2.72 / 1.61	3.59
p=$9.5e^{-1}$	0.00	45.03	pass	$4.9e^{-2}$	$1.1e^{-1}$	1.25	2.72 / 1.61	2.47
p=$9.8e^{-1}$	0.00	45.03	pass	$4.9e^{-2}$	$1.1e^{-1}$	1.25	2.72 / 1.61	3.16
p=$9.9e^{-1}$	0.00	45.03	pass	$4.9e^{-2}$	$1.1e^{-1}$	1.25	2.72 / 1.61	5.03
f=0.00	0.00	439.24	pass	$6.9e^{-2}$	$1.3e^{-1}$	1.25	2.72 / 1.61	4.98
f=$1.2e^{-1}$	0.00	208.44	pass	$5e^{-2}$	$1e^{-1}$	1.25	2.72 / 1.61	5.09
f=$2.5e^{-1}$	0.00	45.03	pass	$4.9e^{-2}$	$1.1e^{-1}$	1.25	2.72 / 1.61	2.89
f=$3.8e^{-1}$	0.00	217.23	pass	$6.6e^{-2}$	$1.2e^{-1}$	1.24	2.72 / 1.61	3.40

Table 18: All successful search trials for Qwen3 0.6b.

Parameters	BER↓	Min.SNR↑	Gate	Δlog-PPL	95% UCB	Budget	Transfer / active ↑	Tok/s
a=$2.5e^{-3}$	0.00	99.35	pass	$2.3e^{-3}$	$4e^{-2}$	1.79	2.30 / 1.44	6.22
a=$5e^{-3}$	0.00	929.85	pass	$4.2e^{-3}$	$3.7e^{-2}$	1.77	2.30 / 1.44	6.19
a=$1e^{-2}$	0.00	781.75	pass	$-4.2e^{-3}$	$3.4e^{-2}$	1.80	2.30 / 1.44	6.25
a=$2e^{-2}$	0.00	23229.87	pass	$1.3e^{-2}$	$4.5e^{-2}$	1.79	2.30 / 1.44	6.03
B=$5e^{-4}$, R=4.00	0.00	2429.97	pass	$1e^{-2}$	$2.5e^{-2}$	2.81	2.15 / 1.38	3.31
B=$5e^{-4}$, R=8.00	0.00	2576.96	pass	$5.2e^{-3}$	$1.4e^{-2}$	2.80	2.15 / 1.38	3.30
B=$5e^{-4}$, R=16.00	0.00	2174.93	pass	$-8.2e^{-3}$	$2.2e^{-3}$	2.82	2.15 / 1.38	3.31
B=$1e^{-3}$, R=4.00	0.00	744.34	pass	$-2.2e^{-3}$	$3.1e^{-2}$	1.79	2.30 / 1.44	3.15
B=$1e^{-3}$, R=8.00	0.00	781.75	pass	$-4.2e^{-3}$	$3.4e^{-2}$	1.80	2.30 / 1.44	3.17
B=$1e^{-3}$, R=16.00	0.00	1416.94	pass	$1.4e^{-2}$	$4.8e^{-2}$	1.79	2.30 / 1.44	3.19
B=$2e^{-3}$, R=4.00	0.00	247.67	pass	$-8.3e^{-3}$	$2.2e^{-2}$	1.25	2.48 / 1.51	3.13
B=$2e^{-3}$, R=8.00	0.00	250.55	pass	$-3.6e^{-2}$	$-2.8e^{-3}$	1.24	2.48 / 1.51	3.18
B=$2e^{-3}$, R=16.00	0.00	273.50	pass	$-2.1e^{-2}$	$1.3e^{-2}$	1.26	2.48 / 1.51	6.80
B=$4e^{-3}$, R=4.00	0.00	681.22	pass	$-1.3e^{-2}$	$2.2e^{-2}$	$9.5e^{-1}$	2.68 / 1.58	6.50
B=$4e^{-3}$, R=8.00	0.00	711.73	pass	$-1.1e^{-2}$	$1.3e^{-2}$	$9.5e^{-1}$	2.68 / 1.58	6.89
B=$4e^{-3}$, R=16.00	0.00	736.36	pass	$-1.5e^{-2}$	$5.9e^{-3}$	$9.5e^{-1}$	2.68 / 1.58	6.86
L=0.00	0.00	781.72	pass	$-4.2e^{-3}$	$3.4e^{-2}$	1.80	2.30 / 1.44	4.94
L=7.00	$1e^{-1}$	3.03	pass	$1.5e^{-2}$	$3.2e^{-2}$	1.79	2.30 / 1.44	4.99
L=14.00	$4e^{-1}$	3.89	pass	$6.5e^{-3}$	$4.6e^{-2}$	1.79	2.30 / 1.44	4.98
L=21.00	0.00	16.32	pass	$-1.1e^{-2}$	$2.5e^{-2}$	1.79	2.30 / 1.44	4.92
L=27.00	0.00	13.67	pass	$-2.8e^{-3}$	$4.3e^{-2}$	1.79	2.30 / 1.44	4.62
p=$9e^{-1}$	0.00	1515.60	pass	$4.5e^{-3}$	$4.6e^{-2}$	$9.5e^{-1}$	2.68 / 1.58	6.13
p=$9.5e^{-1}$	0.00	741.35	pass	$-4.1e^{-3}$	$1.6e^{-2}$	$9.5e^{-1}$	2.68 / 1.58	6.08
p=$9.8e^{-1}$	0.00	736.36	pass	$-1.5e^{-2}$	$5.9e^{-3}$	$9.5e^{-1}$	2.68 / 1.58	6.12
p=$9.9e^{-1}$	0.00	734.43	pass	$-2.3e^{-2}$	$1e^{-2}$	$9.5e^{-1}$	2.68 / 1.58	6.15
f=0.00	0.00	289.71	pass	$-2.1e^{-2}$	$3.2e^{-2}$	$9.4e^{-1}$	2.68 / 1.58	6.11
f=$1.2e^{-1}$	0.00	654.57	pass	$-2e^{-2}$	$3.7e^{-2}$	$9.5e^{-1}$	2.68 / 1.58	6.15
f=$2.5e^{-1}$	0.00	1515.60	pass	$4.5e^{-3}$	$4.6e^{-2}$	$9.5e^{-1}$	2.68 / 1.58	6.01
f=$3.8e^{-1}$	0.00	399.47	pass	$-2.2e^{-2}$	$2.7e^{-2}$	$9.4e^{-1}$	2.68 / 1.58	6.08

Table 19: Transformer layers evaluated in the controlled layer-placement sweep. The same fixed-α fingerprint sidecar is rebound to each listed target layer within a model.

Model	Insertion layers
Qwen3 0.6B	$\{0, 5, 10, 15, 20, 25\}$
Llama 3.2 1B	$\{0, 5, 10, 15\}$
Gemma 3 1B IT	$\{0, 5, 10, 15, 20, 25\}$
Gemma 2 2B	$\{0, 5, 10, 15, 20, 25\}$
Mistral 7B	$\{0, 5, 10, 15, 20, 25, 30\}$
Ministral 8B	$\{0, 5, 10, 15, 20, 25, 30\}$
Llama 3.1 8B	$\{0, 5, 10, 15, 20, 25, 30\}$
GLM-4 9B	$\{0, 5, 10, 15, 20, 25, 30\}$
GLM-Z1 9B	$\{0, 5, 10, 15, 20, 25, 30\}$
DeepSeek V2 Lite	$\{0, 5, 10, 15, 20, 25\}$

F Causal Layer-Placement Sweep of Cache Signal Propagation

We study how the placement of the fingerprint injection affects downstream cache responses. We perform layer-placement sweeps on ten language models. For each model, the model weights, tokenizer, WikiText-2 test split, sequence-selection protocol, sequence length, physical KV head, target projection, perturbation amplitude, and response definition remain unchanged across all placement positions. The fingerprint is inserted into physical KV head 0 of the Value projection with $\alpha = 0.01$.

Table 19 lists the Transformer layers evaluated for each model. We use a five-layer stride throughout the sweep. The sweep covers layers up to 30 for models with sufficient depth. For the shallower models, it ends at the last valid layer sampled by the same stride.

For each placement, we capture clean and fingerprinted caches from the same five held-out sequences and compute the component-wise response at every Transformer layer ℓ and physical KV head h:

$$R^{(c)}_{\ell,h} = \frac{\mathrm{RMS}\Big(F^{(c)}_{\ell,h} - C^{(c)}_{\ell,h}\Big)}{\mathrm{RMS}\Big(C^{(c)}_{\ell,h}\Big) + 10^{-12}}, \tag{31}$$

where $c \in \{\mathrm{Key}, \mathrm{Value}\}$ denotes the cache component. The same fixed $\alpha = 0.01$ projection sidecar is rebound to each target layer, so the sweep changes only the placement while keeping the projection perturbation, physical head, and target projection fixed. This design constitutes a controlled causal placement intervention rather than a set of independently optimized final trials. Each map uses the same model-level logarithmic color limits; exact zero responses are left blank, and the white star marks the direct insertion location. Hence, color indicates response magnitude rather than a signed increase or decrease.

In our experiments, the response maps show a consistent propagation pattern across all ten models. Transformer blocks before the insertion point remain inactive. At the insertion block, the selected Value head exhibits a direct response. The response then propagates to multiple physical KV heads in later blocks. Moving the insertion point deeper shifts the beginning of the response pattern to a later layer and shortens the downstream propagation range. The direct response at the insertion layer remains localized to the selected Value head.

The layer-head response maps from Figure 6 to Figure 15 show this propagation pattern across the evaluated models. Within our experimental scope, these results show that FinCa creates a localized perturbation at the selected cache position and that downstream attention computation propagates this perturbation through subsequent Transformer blocks.

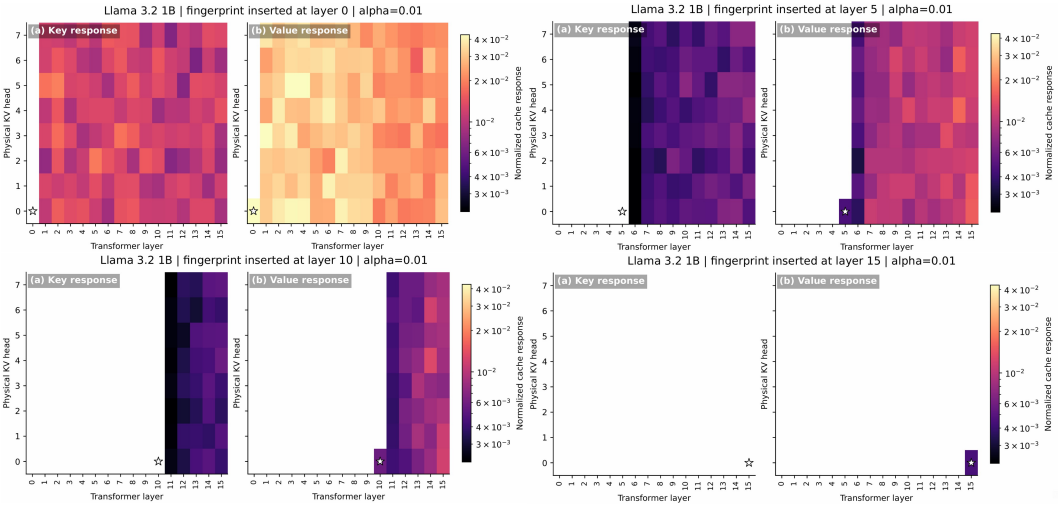


Figure 6: Llama 3.2 1B. Layer-head response after insertion from 0 to 15.

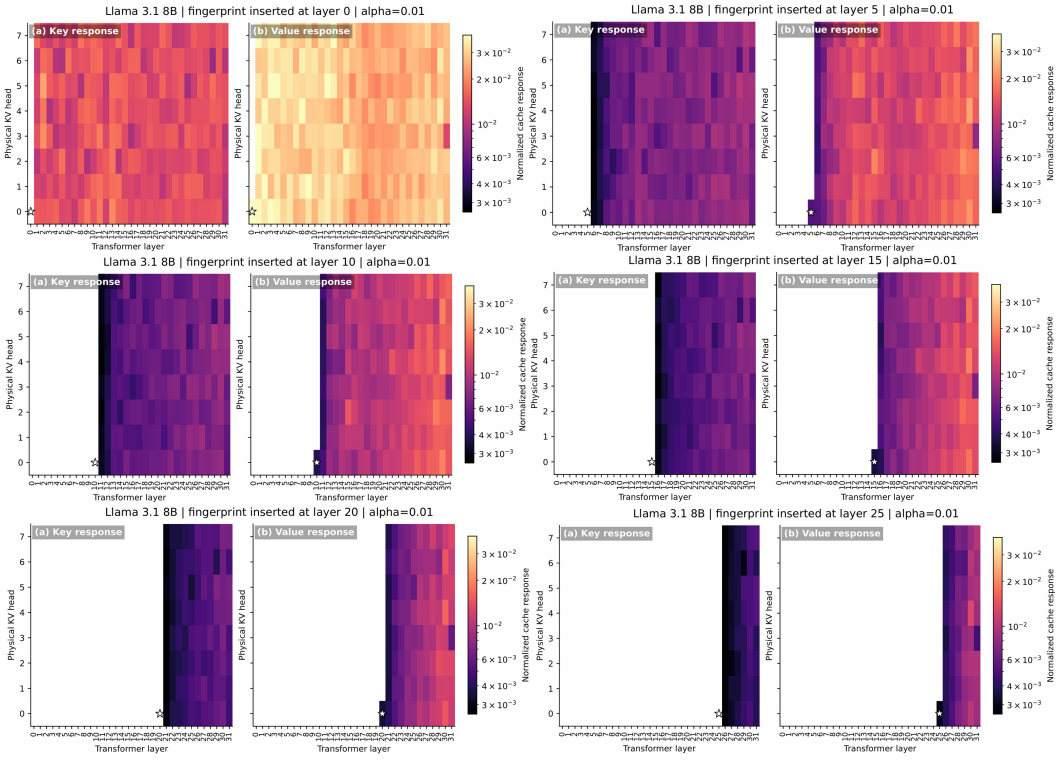


Figure 7: Llama 3.1 8B. Layer-head response after insertion from 0 to 25.

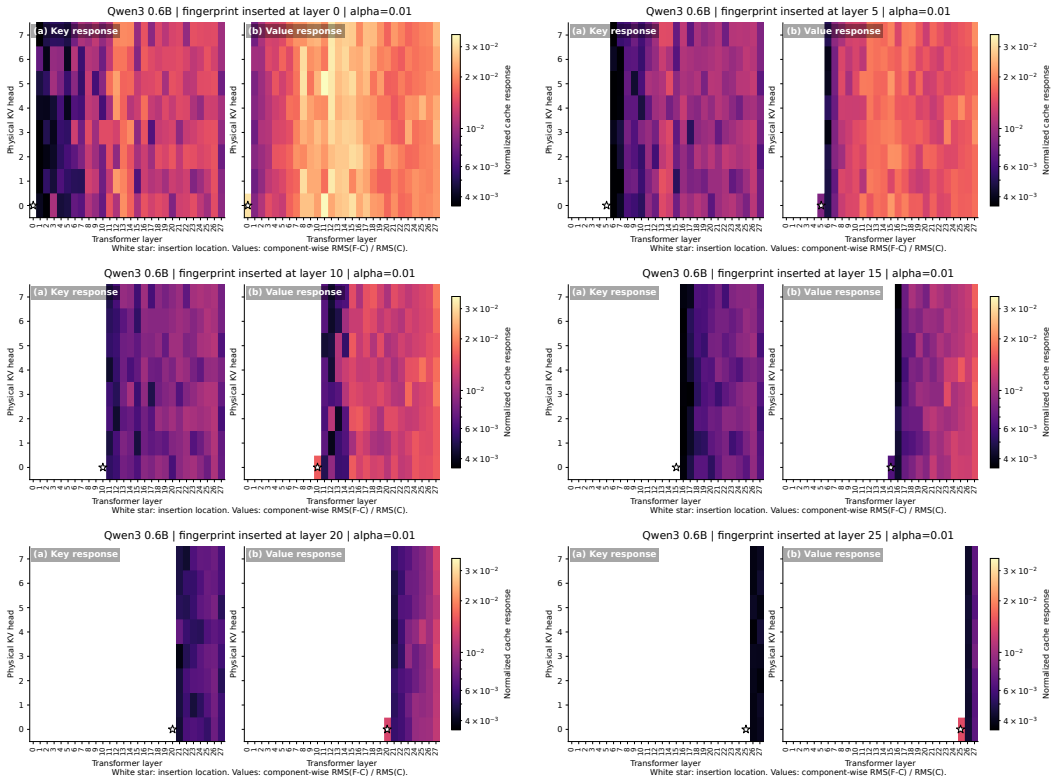


Figure 8: Qwen3 0.6B. Layer-head response after insertion from 0 to 25.

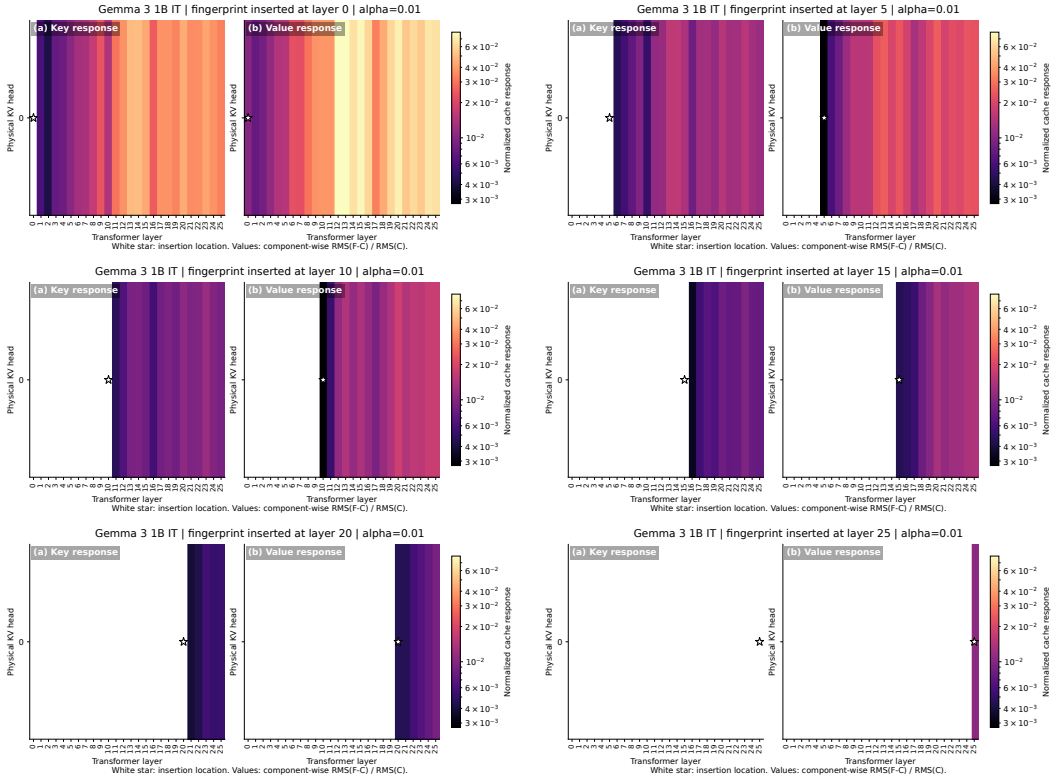


Figure 9: Gemma 3 1B IT. Layer-head response after insertion from 0 to 25.

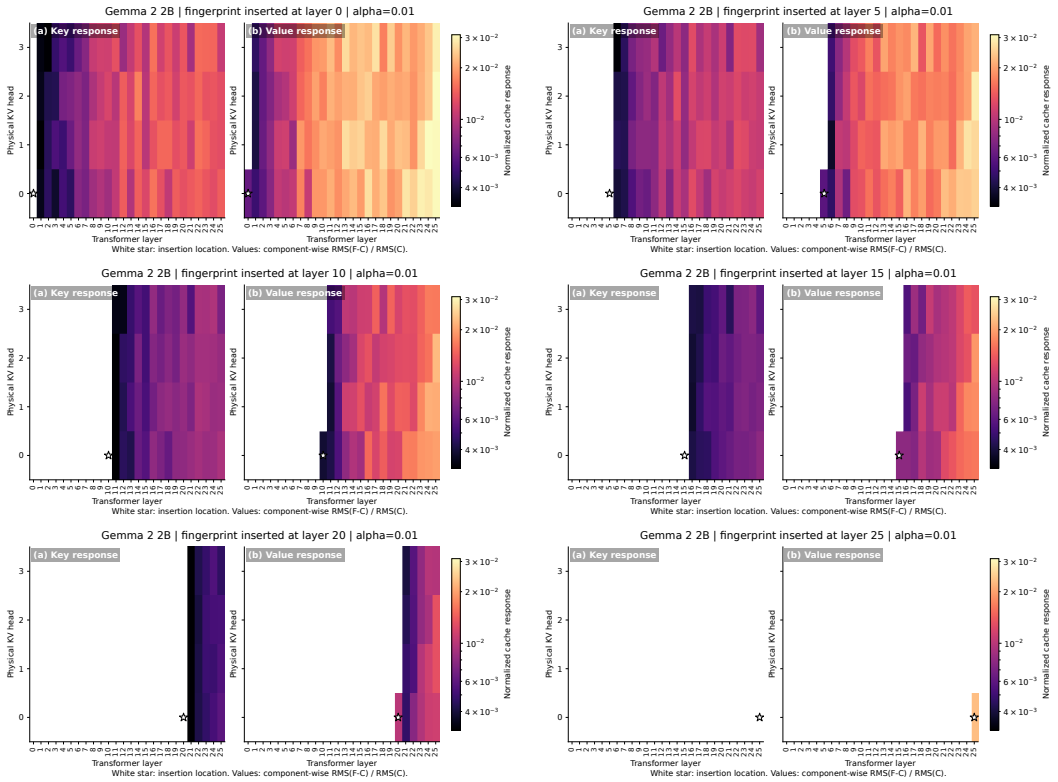


Figure 10: Gemma 2 2B. Layer-head response after insertion from 0 to 25.

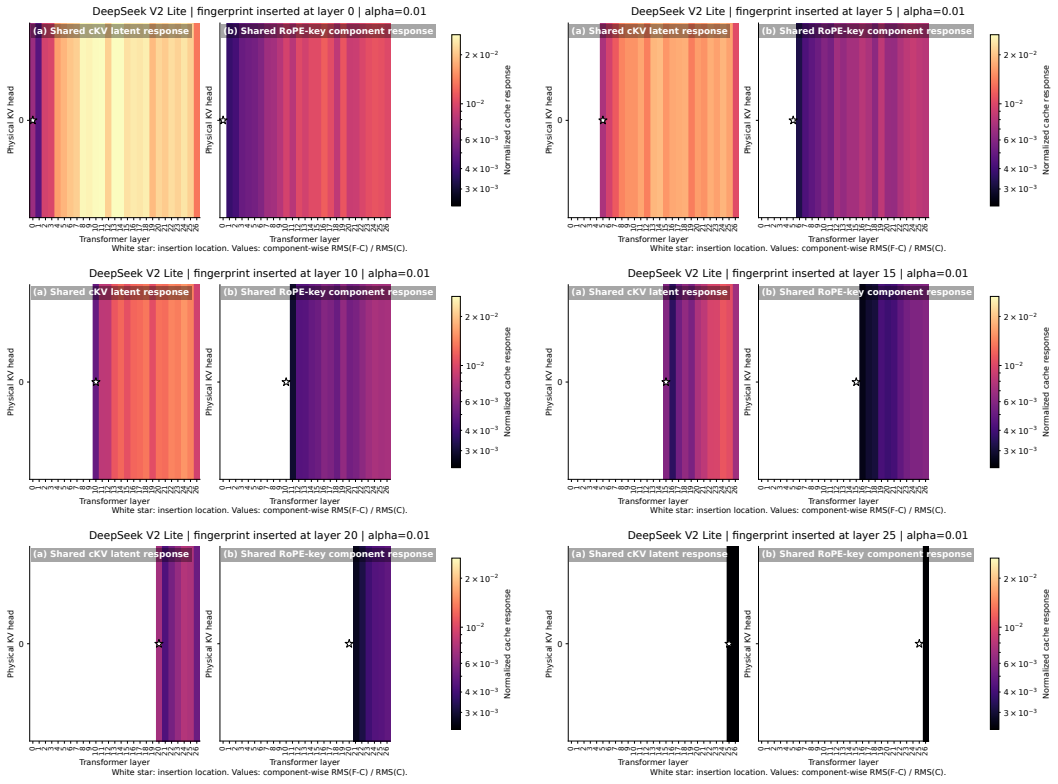


Figure 11: DeepSeek V2 Lite. Layer-head response after insertion from 0 to 25.

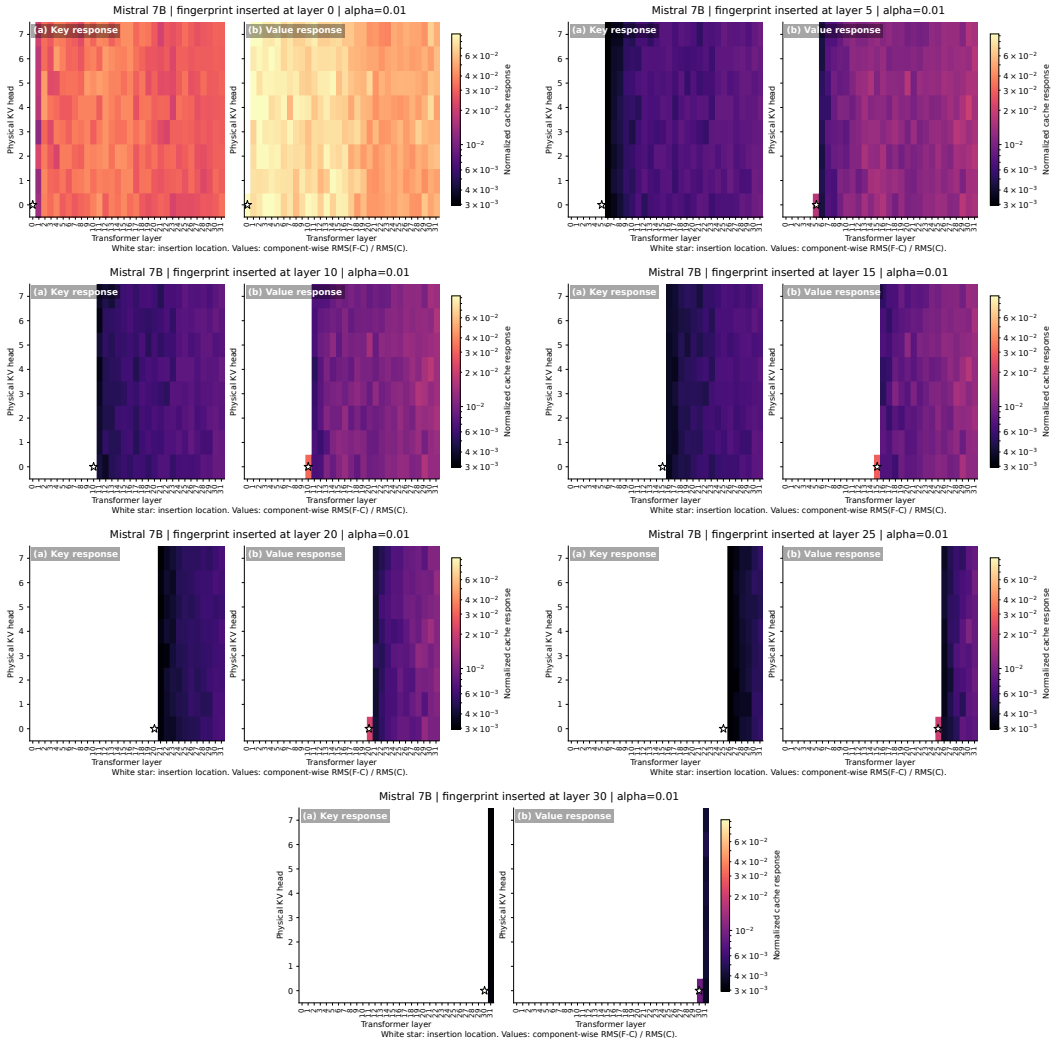


Figure 12: Mistral 7B. Layer-head response after insertion from 0 to 30.

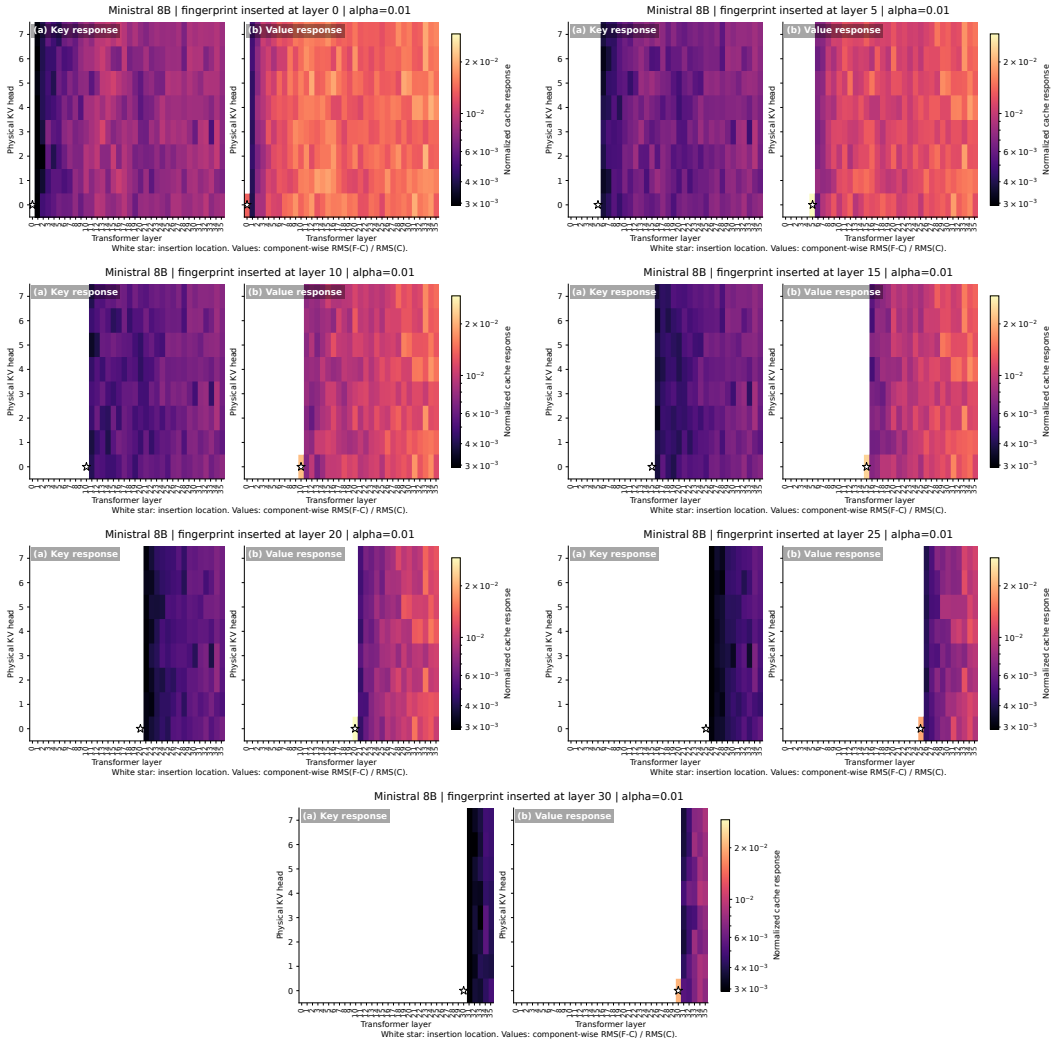


Figure 13: Ministral 8B. Layer-head response after insertion from 0 to 30.

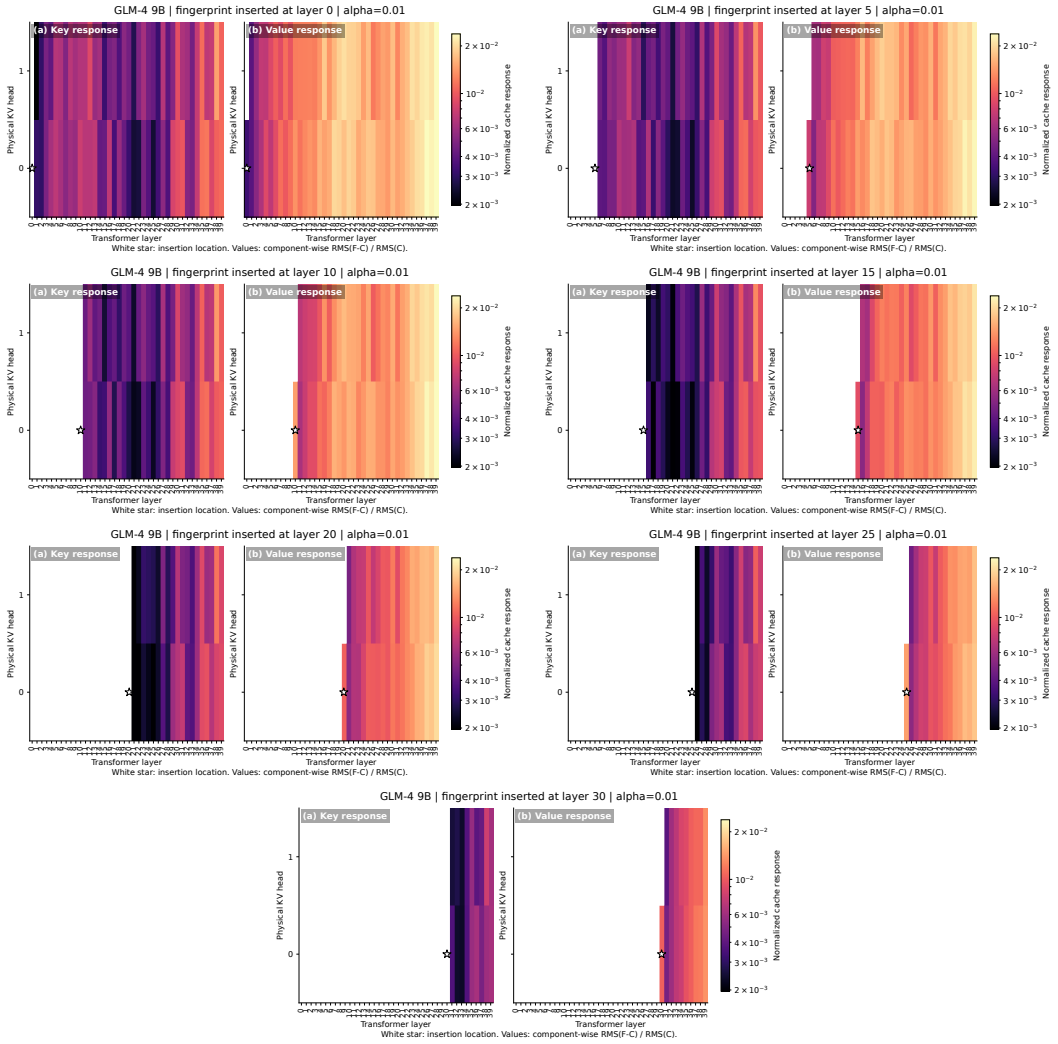


Figure 14: GLM-4 9B. Layer-head response after insertion from 0 to 30.

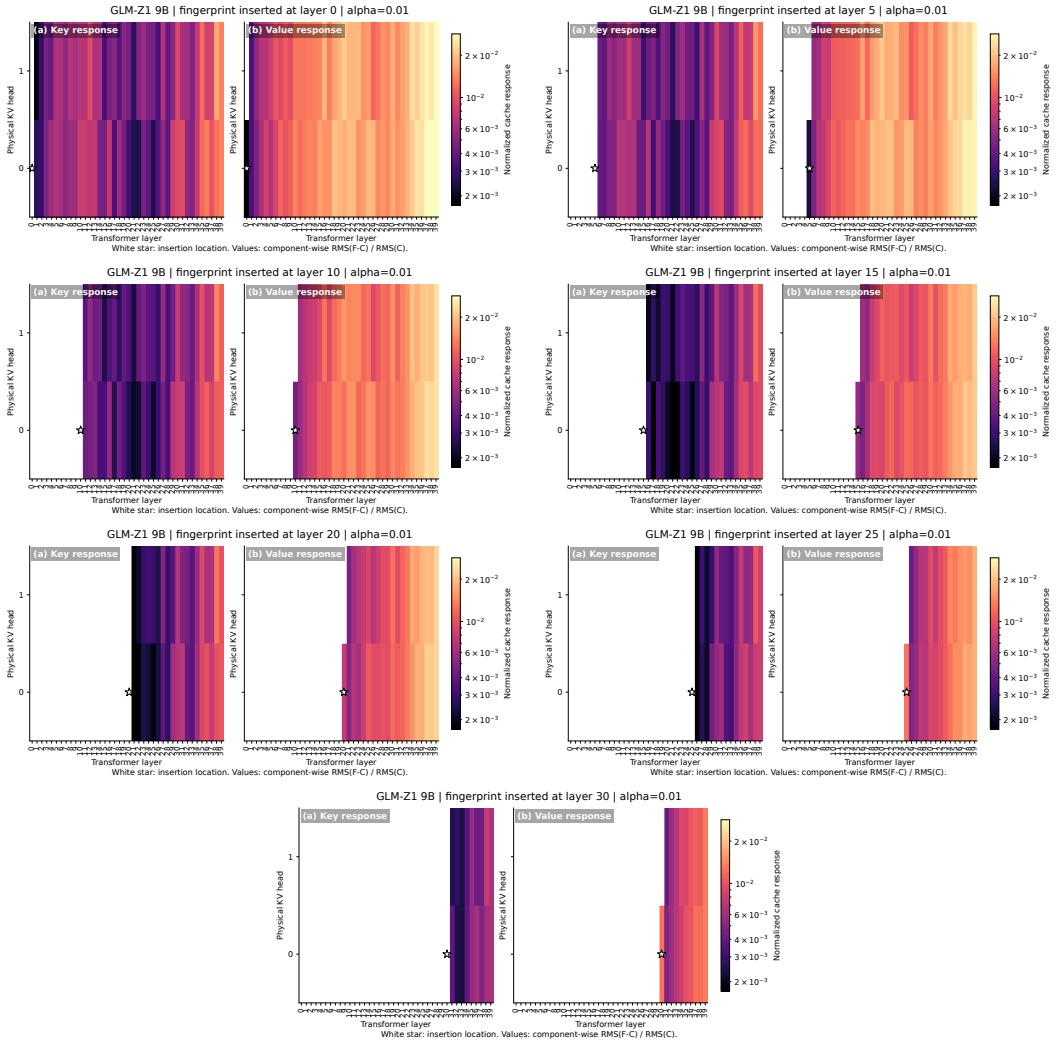


Figure 15: GLM-Z1 9B. Layer-head response after insertion from 0 to 30.

G KV-cache quantization and fingerprint retention

G.1 Fingerprint survival under KV-cache compression

This experiment measures how the ownership signal behaves when the same base checkpoint and its fingerprinted counterpart are executed through different KV-cache routes. Let c denote a cache route. The comparison contains five routes: *No cache* calls the model with `use_cache=False`; it is a model-quality control and has no cache payload to compress. *FP16* keeps the ordinary uncompressed cache and supplies the reference for cache distortion and score retention. *INT8 Reference* and *INT4 Reference* are software reference quantizers used to test ordinary low-bit cache reconstruction; the table abbreviates them as INT8 and INT4. The FinCa row runs the authenticated codec at the configured operating point.

The two reference routes use the same per-token symmetric row quantizer. For one cache row $x = (x_1, \dots, x_w)$ and bit width $\beta \in \{4, 8\}$, define

$$q_{\max} = 2^{\beta-1} - 1, \qquad \alpha_\beta(x) = \frac{\max\{\max_{1\le i\le w} |x_i|, 10^{-8}\}}{q_{\max}},$$

$$\widetilde{x}_i = \mathrm{clip}\left(\mathrm{round}\left(\frac{x_i}{\alpha_\beta(x)}\right), -q_{\max}, q_{\max}\right), \qquad \widehat{x}_i = \alpha_\beta(x)\widetilde{x}_i.$$

INT8 Reference sets $\beta = 8$ and *INT4 Reference* sets $\beta = 4$. The quantized row is stored as an integer tensor together with one FP32 scale, both on the CPU; for $\beta = 4$, the runtime tensor is still an INT8 container, while the logical payload accounting uses four bits per element. The reconstructed row $\widehat{x}$ is materialized only for the transient attention computation. The logical payload therefore contains $w\beta$ bits for the row and 32 bits for its scale. This definition makes the two reference rows reproducible without relying on a vendor-specific integer-attention kernel. The FinCa row uses the calibrated authenticated codec and its private decode state under the same checkpoint pair, fingerprint placement, and document split.

For route c and hypothesis H_h, write $\mathrm{PPL}_{h,c}$ for the perplexity obtained from the aggregate-token definition of PPL_q given in Eq. 26, with $h = 0$ and $h = 1$ corresponding to the clean and fingerprinted routes, respectively. The first two numeric columns are $\mathrm{PPL}_{0,c}$ and $\mathrm{PPL}_{1,c}$. The reported Log PPL Delta is the fingerprinted route PPL relative to the clean no-cache control,

$$\Delta_c^{\mathrm{diag}} = \log\left(\frac{\mathrm{PPL}_{1,c}}{\mathrm{PPL}_{0,\text{No cache}}}\right).$$

This diagnostic combines route-induced changes and the effect of fingerprint insertion. The same-route fingerprint overhead is $\log(\mathrm{PPL}_{1,c} / \mathrm{PPL}_{0,c})$, while the clean-route change relative to the FP16 cache is $\log(\mathrm{PPL}_{0,c} / \mathrm{PPL}_{0,\mathrm{FP16}})$. These are different quantities, so the No-cache and FP16 clean PPL values are allowed to differ even when the model and evaluation documents are unchanged.

Cache MSE compares the fingerprinted cache produced by route c with the fingerprinted FP16 cache at the audited layer and physical KV head. If N_{blk} is the number of aligned cache blocks, then

$$\mathrm{MSE}_c = \frac{1}{N_{\mathrm{blk}}} \sum_{\ell=1}^{N_{\mathrm{blk}}} \mathrm{mean}\left[\left(K_\ell^{(1;c)} - K_\ell^{(1;\mathrm{FP16})}\right)^2\right].$$

Here ℓ indexes aligned cache blocks, and the reported value is an absolute mean squared error. The No-cache route has no cache blocks, so the table uses "-" for cache MSE and the cache-dependent audit columns.

Score Retention measures how much of the FP16 ownership evidence remains. Let $\tau_d^{(1;c)}$ denote the document-level ownership score $\tau_d^{(1)}$ from the canonical definition in this appendix, evaluated for document d under the fingerprinted hypothesis H_1 with cache route c. Thus, its per-symbol scores are the matched scores $S_{d,j}^{(1;c)}$, and the keyed target entries are the

already-defined s_j. Define the route-level ownership gap and retention by

$$G_c = \frac{1}{N_{\text{test}}} \sum_{d=1}^{N_{\text{test}}} \tau_d^{(1;c)}, \qquad R_c = \frac{G_c}{G_{\text{FP16}}}.$$

The ratio is evaluated separately for each model. FP16 is therefore 1 by construction. A value above 1 is also meaningful: it indicates a larger signed carrier-aligned gap than the FP16 run, rather than a probability exceeding its range. No-cache has no score-retention value because it never produces the cache evidence used by this audit.

The cache-method comparison is restricted to models that expose a common explicit K/V-cache abstraction. For these models, the cache can be aligned using the same four coordinates: layer, physical KV head, token position, and within-head K/V coordinate. Consequently, the five routes operate on the same type of cache object, and Cache MSE, the route-level PPL diagnostic, BER, Exact Recovery, and Score Retention retain the same definitions across models. This common representation is essential for a controlled cross-model comparison, because differences in the reported metrics should reflect the cache transformation rather than a change in the cache object being transformed.

Within this evaluation class, and under a finite compute budget, we selected a broad model suite rather than concentrating the comparison on a single model family or scale. The eight checkpoints cover five model families—Gemma, GLM, Llama, Mistral, and Qwen—and parameter scales from 0.6B to 9B: Qwen3 0.6B, Llama 3.2 1B, Gemma 3 1B IT, Gemma 2 2B, Ministral 8B, Llama 3.1 8B, GLM-4 9B, and GLM-Z1 9B. This selection introduces variation in model family, parameter scale, KV-head configuration, and native attention schedule while preserving the common explicit-KV observation coordinates required by the comparison. Table 20 summarizes the five-route measurements for this eight-model suite.

G.2 Fingerprint survival under method-level KV-cache compression

The controlled FP16, INT8, and INT4 quantizers in Section. 4.5 isolate the effect of numerical precision. We next ask whether the conclusion persists under cache policies that combine quantization or eviction with method-specific cache management. Starting from the same clean/fingerprinted checkpoint pair and the same held-out documents as in the preceding experiment, we replace only the cache policy. We evaluate H2O-style (Zhang et al., 2023) and Scissorhands-style eviction (Liu et al., 2023b), KIVI-style (Liu et al., 2024) and KVQuant-style quantization (Hooper et al., 2024), a GEAR-style low-rank approximation (Kang et al., 2024), RotateKV (Su et al., 2025) and CommVQ (Li et al., 2025a) on the eight-model evaluation suite.

Adaptation. Let $X = [K, V] \in \mathbb{R}^{T \times 2d}$ denote the concatenated key–value block of one attention head, where T is the number of cached tokens and d is the head dimension. The KIVI- and KVQuant-style routes use the same affine decode rule,

$$\widehat{x} = s \, \text{clip}(\text{round}((x - z)/s), 0, 2^{n_b} - 1) + z,$$

where n_b is the bit width and (z, s) are the minimum and step size estimated over the specified group. KIVI-style quantization uses a per-channel group over cached tokens for K and a per-token group over channels for V. KVQuant-style quantization uses per-channel groups for both K and V, with (z, s) estimated after clipping each channel to its central 99.9% interval. This controlled distinction preserves the principal grouping and outlier-handling choices of the two methods while keeping the decoder and evaluation interface identical.

RotateKV- and CommVQ-style routes are integrated through the same layer-level explicit-KV interface. The former applies a calibrated grouped rotation and outlier-aware channel reordering before low-bit reconstruction, followed by an inverse transformation that restores the original K/V-cache coordinates. The latter uses calibration-derived diagonal additive codebooks to reconstruct the key and value components separately under a two-bit average representation. In both cases, the transformed cache is returned to the same layer–head–token layout used by the other routes, so the downstream attention computation and the

Table 20: Fingerprint survival under FP16 and integer-quantized KV-cache references, with a no-cache control and the FinCa operating point.

Route	Clean PPL	Finger PPL	Log PPL Delta	Cache MSE vs FP16	BER	Exact Recovery	Min SNR	SNR Gate	Score Retention
Gemma 2 2B									
No Cache	1444.053	1445.055	0.001	-	-	-	-	-	-
FP16	1445.933	1445.537	0.001	0.000	0.0%	100.0%	6.39×10^5	Pass	1.000
INT8	1453.375	1452.499	0.006	0.003	0.0%	100.0%	772.257	Pass	1.002
INT4	1917.152	1911.908	0.281	0.364	0.0%	100.0%	24.099	Pass	0.814
FinCa	1421.308	1423.318	-0.015	0.019	0.0%	100.0%	328.033	Pass	0.991
Gemma 3 1B IT									
No Cache	364.071	363.610	-0.001	-	-	-	-	-	-
FP16	358.573	364.041	0.000	0.000	0.0%	100.0%	2.00×10^4	Pass	1.000
INT8	362.605	360.889	-0.009	0.016	0.0%	100.0%	205.122	Pass	0.916
INT4	1269.291	1211.606	1.202	3.100	0.0%	100.0%	30.841	Pass	0.880
FinCa	343.709	349.479	-0.041	0.071	0.0%	100.0%	306.662	Pass	0.971
GLM-4 9B									
No Cache	22.315	22.328	0.001	-	-	-	-	-	-
FP16	22.291	22.327	0.001	0.000	0.0%	100.0%	3.11×10^5	Pass	1.000
INT8	22.279	22.184	-0.006	0.002	0.0%	100.0%	1036.275	Pass	1.003
INT4	138.926	136.996	1.815	0.236	0.0%	100.0%	23.210	Pass	0.929
FinCa	22.527	22.617	0.013	0.028	0.0%	100.0%	611.922	Pass	1.005
GLM-Z1 9B									
No Cache	42.493	42.489	0.000	-	-	-	-	-	-
FP16	42.274	42.283	-0.005	0.000	0.0%	100.0%	2.40×10^5	Pass	1.000
INT8	41.703	41.728	-0.018	0.023	0.0%	100.0%	424.644	Pass	0.999
INT4	302.620	302.389	1.962	0.965	0.0%	100.0%	17.156	Pass	0.958
FinCa	42.503	42.500	0.000	0.017	0.0%	100.0%	236.930	Pass	1.003
Llama 3.1 8B									
No Cache	13.978	13.993	0.001	-	-	-	-	-	-
FP16	14.018	13.999	0.002	0.000	0.0%	100.0%	1.26×10^5	Pass	1.000
INT8	14.292	14.212	0.017	0.000	0.0%	100.0%	63.139	Pass	1.058
INT4	2353.882	2385.117	5.140	0.032	95.00%	5.00%	0.050	Fail	0.012
FinCa	15.714	15.671	0.114	0.008	0.0%	100.0%	198.696	Pass	0.987
Llama 3.2 1B									
No Cache	22.887	22.837	-0.002	-	-	-	-	-	-
FP16	22.875	22.777	-0.005	0.000	0.0%	100.0%	1.12×10^7	Pass	1.000
INT8	23.235	23.160	0.012	0.000	0.0%	100.0%	409.698	Pass	0.961
INT4	771.029	771.383	3.518	0.040	2.50%	95.00%	0.065	Fail	0.002
FinCa	23.744	23.739	0.037	0.008	0.0%	100.0%	6628.986	Pass	0.995
Ministral 8B									
No Cache	15.177	15.186	0.001	-	-	-	-	-	-
FP16	15.182	15.184	0.001	0.000	0.0%	100.0%	6.59×10^5	Pass	1.000
INT8	15.200	15.207	0.002	0.003	0.0%	100.0%	626.159	Pass	1.001
INT4	1362.360	1357.792	4.494	1.667	0.0%	100.0%	1.812	Fail	0.466
FinCa	15.407	15.406	0.015	0.031	0.0%	100.0%	197.509	Pass	0.995
Qwen3 0.6B									
No Cache	47.758	47.714	-0.001	-	-	-	-	-	-
FP16	47.793	47.729	-0.001	0.000	0.0%	100.0%	1.38×10^6	Pass	1.000
INT8	2140.690	2205.847	3.833	0.085	5.00%	90.00%	1.602	Fail	0.100
INT4	7.99×10^5	7.99×10^5	9.724	1.459	50.00%	0.0%	0.000	Fail	0.000
FinCa	48.028	48.066	0.006	0.302	0.0%	100.0%	206.989	Pass	1.005

fingerprint-retention audit remain unchanged. These implementations preserve the central cache-transformation mechanisms of RotateKV and CommVQ while using the common causal-cache evaluation protocol described above.

For eviction, $\bar{A}_i^{(t)}$ is the attention mass assigned to cached position i at decoding step t. H2O-style eviction ranks non-recent, non-sink positions by cumulative importance $h_i^{(t)} = h_i^{(t-1)} + \bar{A}_i^{(t)}$. Scissorhands-style eviction instead uses persistent importance $h_i^{(t)} = \max\{\rho h_i^{(t-1)}, \bar{A}_i^{(t)}\}$ with $\rho = 0.995$. Both policies retain four sink tokens, the most re-

cent 32 tokens, and the highest-scoring remaining positions within a budget of 64 tokens. For the GEAR-style control, the selected X is reconstructed from its rank-r truncated SVD, $\widehat{X} = U_r \Sigma_r V_r^\top$, with $r = 8$; this route is a low-rank approximation control on the audited layer/head rather than a complete reproduction of GEAR's mixed-precision error-compensation runtime.

For eviction routes, clean and fingerprinted traces are aligned on the intersection of their final retained token positions before computing the carrier-matched statistic. Thus the retention score $R_r = G_r/G_{\mathrm{FP16}}$, defined in the preceding subsection, is evaluated on the cache state that remains available to the observer.

How to read the tables. The tables separate model quality from fingerprint recovery. For model quality, compare the clean and fingerprinted PPL values with the FP16 row of the same model; the compact screening diagnostic $\Delta_{\log \mathrm{PPL}}$ should be close to 0 (smaller $|\Delta|$ is better). For fingerprint recovery, **BER** and **Exact Recovery** are the hard checks (ideal values are 0% and 100%), while **Score Retention** is the most direct continuous measure of ownership evidence: 1.00 means FP16-level evidence, values below 1 indicate attenuation, and values above 1 indicate a larger signed gap rather than a probability. Cache MSE and Min SNR are supporting diagnostics; SNR Gate is their pass/fail summary. The FP16 row is the within-model reference.

Result analysis. In our experiments, Table 20 shows that FinCa preserves the fingerprint signal under low-bit KV-cache reconstruction. Across all eight models, BER remains 0% and Exact Recovery remains 100%, while Score Retention ranges from 0.971 to 1.006, with a mean of 0.994, closely matching the FP16 reference. All FinCa configurations also pass the SNR Gate. INT8 reference quantization still recovers every symbol, although its Score Retention varies more substantially across models, indicating that ordinary 8-bit reconstruction introduces measurable cache distortion. The degradation becomes pronounced at 4 bits: PPL increases sharply for several models, and Llama 3.1 8B and Llama 3.2 1B exhibit BER values of 95% and 2.5%, with Score Retention falling to 0.012 and 0.002, respectively. Within our experimental range, FinCa therefore keeps cache-induced distortion within a level that preserves ownership evidence, whereas aggressive unauthenticated low-bit quantization can make fingerprint recovery unreliable.

Table 21 evaluates eviction, low-rank approximation, and method-level quantization. H2O, Scissorhands, and most GEAR configurations maintain 0% BER and 100% Exact Recovery across the eight models, with Score Retention remaining close to 1. These results show that retaining important tokens or applying a moderate low-rank approximation does not substantially weaken the fingerprint. KIVI, RotateKV, and KVQuant are more sensitive to the model and compression setting: several configurations still achieve complete recovery, while others introduce larger PPL and Cache MSE increases together with wider Score Retention variation. KVQuant produces 2.5% BER, 95% Exact Recovery, and an SNR Gate failure on Gemma 3 1B IT and GLM-4 9B. CommVQ 2-bit retains complete symbol recovery, but its Score Retention is consistently lower, showing that the fingerprint can remain decodable while the strength of the ownership evidence decreases. In our experiments, token-selection-based eviction provides the most stable operating point, while stronger quantization and higher-error reconstruction require careful compatibility checks with the fingerprint mechanism.

These findings provide a practical basis for selecting KV-cache mechanisms when deploying fingerprinted language models in real-world serving environments.

Table 21: Fingerprint survival under method-level KV-cache compression.

Route	Clean PPL	Finger PPL	Log PPL Delta	Cache MSE vs FP16	BER	Exact Recovery	Min SNR	SNR Gate	Score Retention
Gemma 2 2B									
H2O	1489.699	1492.563	0.033	0.009	0.0%	100.0%	3.02×10^5	Pass	1.000
Scissorhands	1499.340	1500.535	0.038	0.009	0.0%	100.0%	3.11×10^5	Pass	1.000
KIVI INT2	1112.153	1112.335	-0.261	0.511	0.0%	100.0%	56.1	Pass	0.999
KVQuant INT2	1266.295	1268.698	-0.130	0.398	0.0%	100.0%	9.0	Pass	1.029
GEAR Rank8	1743.153	1743.591	0.189	0.035	0.0%	100.0%	6.19×10^5	Pass	1.000
RotateKV INT2	1815.284	1816.247	0.229	0.506	0.0%	100.0%	114.9	Pass	1.029
CommVQ 2-bit	1847.475	1851.837	0.249	0.236	0.0%	100.0%	289.4	Pass	0.723
Gemma 3 1B IT									
H2O	377.680	376.283	0.033	0.000	0.0%	100.0%	4387.8	Pass	0.995
Scissorhands	373.687	372.697	0.023	0.000	0.0%	100.0%	6910.4	Pass	0.995
KIVI INT2	889.959	818.149	0.810	28.644	0.0%	100.0%	4.8	Pass	0.575
KVQuant INT2	628.064	637.462	0.560	10.023	2.5%	95.0%	1.9	Fail	1.032
GEAR Rank8	397.466	403.430	0.103	16.820	0.0%	100.0%	45.7	Pass	0.923
RotateKV INT2	1750.158	1758.710	1.575	10.571	0.0%	100.0%	21.4	Pass	1.430
CommVQ 2-bit	446.949	434.179	0.176	5.291	0.0%	100.0%	18.8	Pass	0.824
GLM-4 9B									
H2O	22.538	22.644	0.015	0.009	0.0%	100.0%	2104.2	Pass	1.000
Scissorhands	22.589	22.693	0.017	0.007	0.0%	100.0%	804.6	Pass	1.003
KIVI INT2	30.288	31.226	0.336	0.628	0.0%	100.0%	18.6	Pass	0.967
KVQuant INT2	44.660	44.924	0.700	1.058	2.5%	95.0%	2.3	Fail	1.041
GEAR Rank8	22.307	22.321	0.000	0.005	0.0%	100.0%	2.86×10^5	Pass	1.000
RotateKV INT2	63.026	64.348	1.059	0.619	0.0%	100.0%	34.5	Pass	0.981
CommVQ 2-bit	34.289	34.809	0.445	0.474	0.0%	100.0%	82.5	Pass	0.632
GLM-Z1 9B									
H2O	43.231	43.245	0.018	0.006	0.0%	100.0%	1.85×10^5	Pass	1.000
Scissorhands	43.174	43.178	0.016	0.006	0.0%	100.0%	1.98×10^5	Pass	1.000
KIVI INT2	69.378	69.541	0.493	0.526	0.0%	100.0%	20.4	Pass	1.029
KVQuant INT2	121.725	121.933	1.054	0.838	0.0%	100.0%	4.5	Pass	1.004
GEAR Rank8	42.409	42.482	0.000	0.004	0.0%	100.0%	2.05×10^5	Pass	1.000
RotateKV INT2	166.726	166.644	1.367	0.924	0.0%	100.0%	24.7	Pass	1.020
CommVQ 2-bit	95.644	95.647	0.811	0.416	0.0%	100.0%	69.7	Pass	0.762
Llama 3.1 8B									
H2O	14.078	14.087	0.008	0.000	0.0%	100.0%	1.55×10^5	Pass	1.000
Scissorhands	14.322	14.334	0.025	0.000	0.0%	100.0%	1.39×10^5	Pass	1.000
KIVI INT2	38.445	38.721	1.019	0.186	0.0%	100.0%	13.6	Pass	1.474
KVQuant INT2	4284.907	3836.702	5.615	0.204	0.0%	100.0%	6.8	Pass	0.975
GEAR Rank8	14.311	14.328	0.025	0.329	0.0%	100.0%	63.5	Pass	0.570
RotateKV INT2	34.502	35.057	0.919	0.245	0.0%	100.0%	67.9	Pass	0.909
CommVQ 2-bit	78.414	78.816	1.730	0.052	0.0%	100.0%	79.0	Pass	0.745
Llama 3.2 1B									
H2O	23.038	22.938	0.002	0.000	0.0%	100.0%	5.99×10^6	Pass	1.000
Scissorhands	23.159	23.091	0.009	0.000	0.0%	100.0%	5.09×10^6	Pass	1.000
KIVI INT2	114.347	106.158	1.534	0.326	0.0%	100.0%	33.2	Pass	0.849
KVQuant INT2	2236.019	2392.397	4.650	0.353	0.0%	100.0%	35.7	Pass	0.987
GEAR Rank8	25.298	25.405	0.104	0.468	0.0%	100.0%	19.5	Pass	0.482
RotateKV INT2	1067.558	955.666	3.732	0.442	0.0%	100.0%	168.4	Pass	0.981
CommVQ 2-bit	162.775	155.528	1.916	0.093	0.0%	100.0%	146.7	Pass	0.762
Ministral 8B									
H2O	15.317	15.321	0.010	0.004	0.0%	100.0%	2.91×10^5	Pass	1.000
Scissorhands	15.336	15.342	0.011	0.007	0.0%	100.0%	6870.6	Pass	0.999
KIVI INT2	28.093	28.125	0.617	0.629	0.0%	100.0%	5.7	Pass	1.022
KVQuant INT2	322.594	322.883	3.058	1.517	0.0%	100.0%	6.7	Pass	0.955
GEAR Rank8	15.188	15.195	0.001	0.000	0.0%	100.0%	8.92×10^5	Pass	1.000
RotateKV INT2	23.280	23.319	0.430	0.567	0.0%	100.0%	21.8	Pass	1.044
CommVQ 2-bit	61.021	61.154	1.394	0.459	0.0%	100.0%	49.7	Pass	0.688
Qwen3 0.6B									
H2O	48.698	48.627	0.018	0.000	0.0%	100.0%	4.64×10^5	Pass	1.000
Scissorhands	48.457	48.279	0.011	0.000	0.0%	100.0%	3.39×10^5	Pass	1.000
KIVI INT2	275.740	256.973	1.683	3.566	0.0%	100.0%	27.0	Pass	1.468
KVQuant INT2	2.11×10^4	1.68×10^4	5.863	3.434	0.0%	100.0%	10.4	Pass	1.105
GEAR Rank8	48.874	48.882	0.023	0.457	0.0%	100.0%	13.1	Pass	0.490
RotateKV INT2	2.96×10^4	3.60×10^4	6.625	17.268	0.0%	100.0%	25.7	Pass	1.170
CommVQ 2-bit	2053.456	1945.445	3.707	3.617	0.0%	100.0%	105.4	Pass	0.696

H Ownership Score Distribution

H.1 H0/H1 Ownership Score Distribution

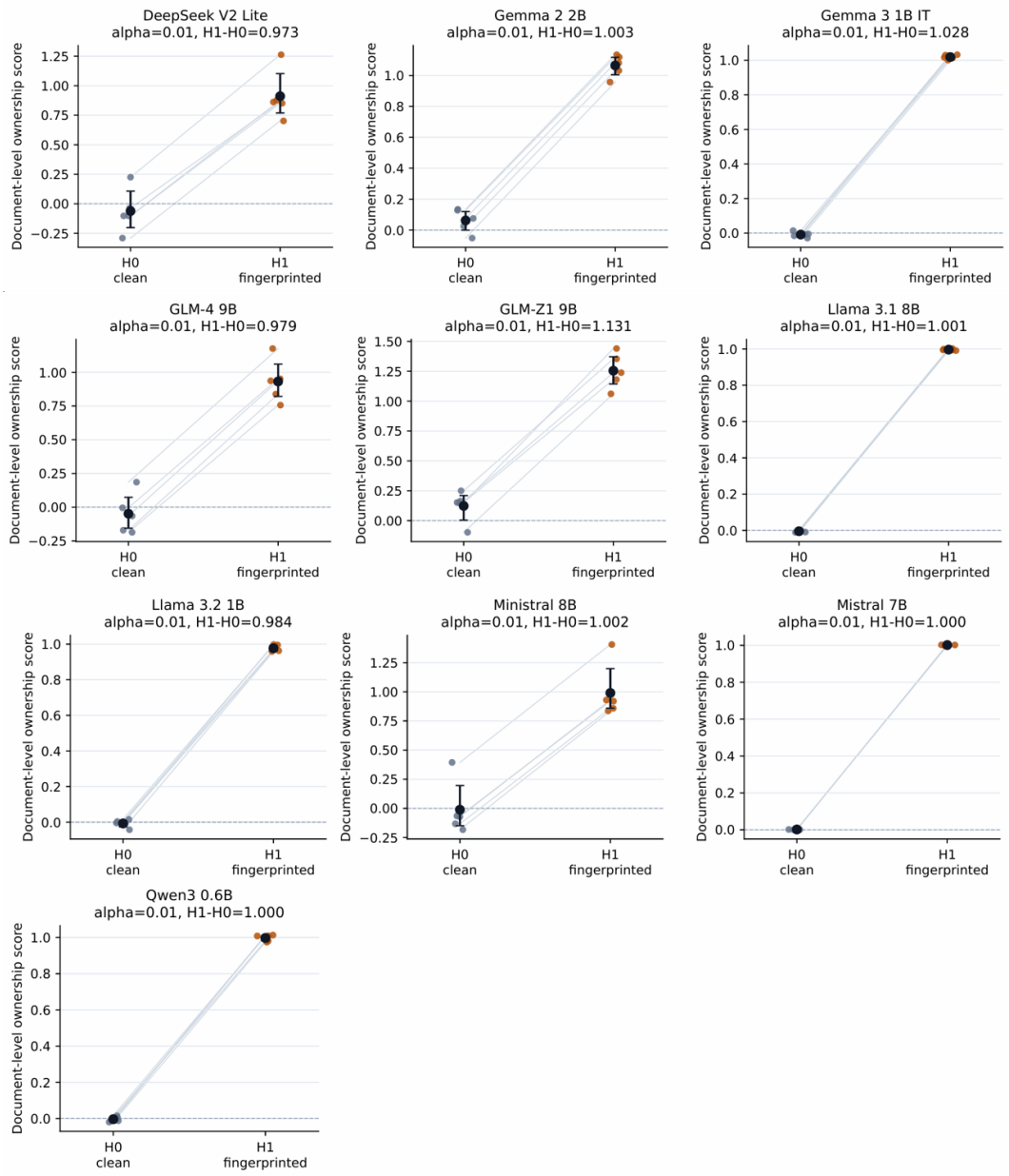


Figure 16: Paired ownership scores under the matched codec/QIM. Each line connects the same held-out document in H_0 and H_1; error bars are 95% bootstrap CIs for the mean.

H.2 KV-Cache Mechanism of FinCa

This experiment measures how the matched codec transforms the Value subspace that carries the fingerprint. Following Section 4.1, we replay one held-out WikiText-2 sequence for each of the ten evaluated models. Token positions, the selected layer and physical cache unit, the cache-coordinate convention, the codec state, and the QIM settings are fixed within each model. The four routes separate the clean cache, the fingerprint signal, and the compression residual: C_0 is the clean uncompressed reference, $F_0 - C_0$ is the fingerprint

before compression, $C_1 - C_0$ is the compression residual without a fingerprint, and $F_1 - C_1$ is the fingerprint signal recovered after compression. This construction gives a direct coordinate-level readout of the codec response to the embedded signal.

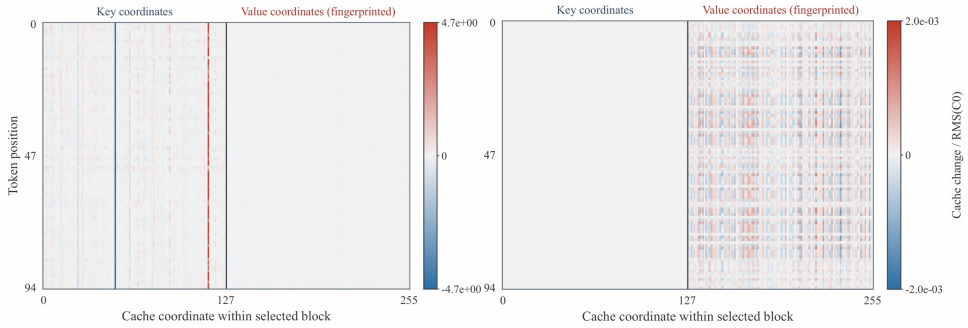
The table reports the replay and audit statistics together. The "Before" and "After" rows give the PPL values from the mechanism replay, and each route-wise $\Delta \log \mathrm{PPL}$ measures the change from before to after compression. The ownership rows use the document-level paired statistic defined in Section D: H_0 is the clean-compressed control, H_1 is the fingerprinted-compressed route, and $H_1 - H_0$ is the matched ownership contrast. The replay rows describe the sequence used for the coordinate maps, while the ownership rows summarize the held-out documents in the audit protocol. The factorial PPL contrasts are computed from the document-level paired analysis.

Table 22: Four-route mechanism controls for all ten evaluated models. For each model, the "Before" and "After" entries are the perplexities from one fixed held-out replay. The ownership entries are document-level paired means under the same frozen codec and QIM state; the factorial contrast is the corresponding matched contrast from the audit protocol.

Metric	**Qwen3-0.6B**			**Llama-3.2-1B**		
	Clean control	Fingerprinted	Factorial contrast	Clean control	Fingerprinted	Factorial contrast
Route	$C_0 \to C_1$	$F_0 \to F_1$	$H_1 - H_0$	$C_0 \to C_1$	$F_0 \to F_1$	$H_1 - H_0$
Before	47.817	47.723	–	22.894	22.825	–
After	48.200	48.492	–	23.735	23.579	–
$\Delta \log \mathrm{PPL}$	+0.0080	+0.0160	+0.0140	+0.0361	+0.0325	+0.0295
Ownership	$H_0 = 0.002$	$H_1 = 0.852$	$H_1 - H_0 = 0.850$	$H_0 = -0.001$	$H_1 = 0.995$	$H_1 - H_0 = 0.995$
	Gemma 3 1B IT			**Gemma 2 2B**		
Route	$C_0 \to C_1$	$F_0 \to F_1$	$H_1 - H_0$	$C_0 \to C_1$	$F_0 \to F_1$	$H_1 - H_0$
Before	360.657	362.907	–	1452.376	1451.655	–
After	350.582	347.054	–	1435.110	1435.530	–
$\Delta \log \mathrm{PPL}$	−0.0283	−0.0447	−0.0384	−0.0120	−0.0112	−0.0117
Ownership	$H_0 = 0.014$	$H_1 = 0.821$	$H_1 - H_0 = 0.807$	$H_0 = -0.074$	$H_1 = 0.922$	$H_1 - H_0 = 0.995$
	Mistral 7B			**Ministral 8B**		
Route	$C_0 \to C_1$	$F_0 \to F_1$	$H_1 - H_0$	$C_0 \to C_1$	$F_0 \to F_1$	$H_1 - H_0$
Before	12.059	12.063	–	15.191	15.189	–
After	14.347	14.476	–	15.369	15.382	–
$\Delta \log \mathrm{PPL}$	+0.1737	+0.1823	+0.1827	+0.0116	+0.0126	+0.0125
Ownership	$H_0 = -0.002$	$H_1 = 0.997$	$H_1 - H_0 = 0.999$	$H_0 = 0.026$	$H_1 = 1.012$	$H_1 - H_0 = 0.986$
	Llama-3.1-8B			**GLM-4 9B**		
Route	$C_0 \to C_1$	$F_0 \to F_1$	$H_1 - H_0$	$C_0 \to C_1$	$F_0 \to F_1$	$H_1 - H_0$
Before	14.012	14.000	–	22.256	22.291	–
After	15.700	15.686	–	22.448	22.488	–
$\Delta \log \mathrm{PPL}$	+0.1137	+0.1137	+0.1129	+0.0086	+0.0088	+0.0103
Ownership	$H_0 = -0.014$	$H_1 = 0.973$	$H_1 - H_0 = 0.986$	$H_0 = -0.056$	$H_1 = 0.934$	$H_1 - H_0 = 0.991$
	GLM-Z1 9B			**DeepSeek V2 Lite**		
Route	$C_0 \to C_1$	$F_0 \to F_1$	$H_1 - H_0$	$C_0 \to C_1$	$F_0 \to F_1$	$H_1 - H_0$
Before	42.296	42.299	–	14.316	14.305	–
After	42.997	43.010	–	15.784	15.759	–
$\Delta \log \mathrm{PPL}$	+0.0164	+0.0167	+0.0167	+0.0976	+0.0968	+0.0960
Ownership	$H_0 = -0.046$	$H_1 = 0.949$	$H_1 - H_0 = 0.995$	$H_0 = 0.015$	$H_1 = 1.009$	$H_1 - H_0 = 0.994$

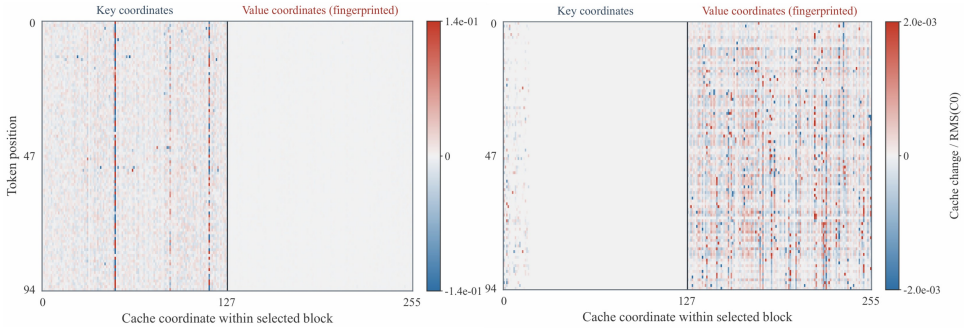
In our experiments, Figure 17 shows the codec response for Qwen3-0.6B and Llama-3.2-1B at cache-coordinate level. The uncompressed fingerprint $F_0 - C_0$ is concentrated in the Value half. The compression residual $C_1 - C_0$ captures the codec background, including residuals on the Key side. After this residual is removed, $F_1 - C_1$ retains the Value-side fingerprint pattern. Across the nine available replay artifacts, the fingerprint-subspace cosine ranges from 0.870 to 0.994 (Qwen3: 0.945; Llama-3.2: 0.994).

The audit results in Table 22 show the same pattern across the ten models. In our experiments, the clean-compressed control remains near zero in most models, while the fingerprinted-compressed route reaches a large positive ownership score. All ten models achieve zero BER and exact recovery on the held-out audit documents, and the ownership contrast ranges from 0.807 to 0.999. In our experiments, the evaluated codec preserves the designated fingerprint-bearing Value subspace while compression changes other cache coordinates.



(a) Clean, uncompressed cache state (C0)

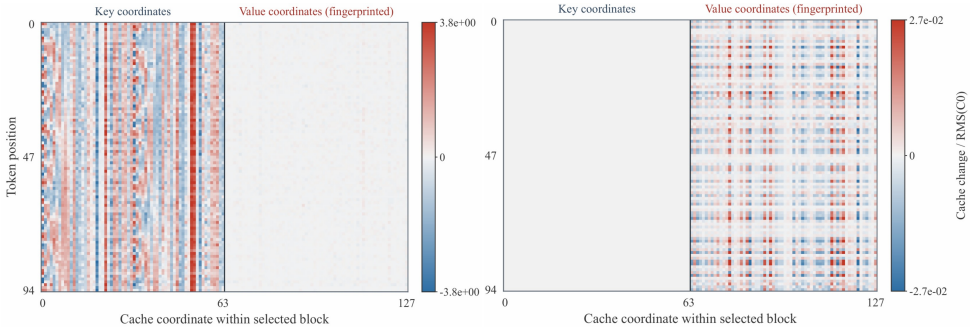
(b) Fingerprint-induced cache change before compression



(c) Matched compression-only cache change (H0)

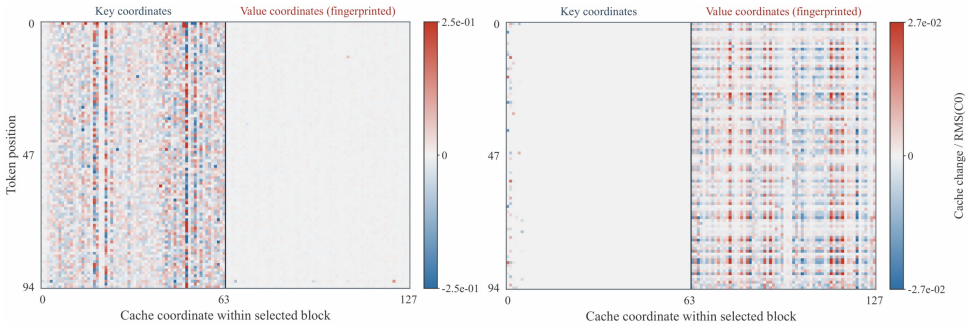
(d) Fingerprint-induced cache change after compression

(1) Qwen3-0.6B; layer 0; physical cache unit 0; explicit key–value (K/V) cache.



(a) Clean, uncompressed cache state (C0)

(b) Fingerprint-induced cache change before compression



(c) Matched compression-only cache change (H0)

(d) Fingerprint-induced cache change after compression

(2) Llama 3.2 1B; layer 0; physical cache unit 0; explicit key–value (K/V) cache.

Figure 17: Coordinate-level mechanism replay for Qwen3 and Llama 3.2. Rows are token positions and columns are cache coordinates; the divider separates Key and Value halves. The maps are C_0, $F_0 - C_0$, $C_1 - C_0$, and $F_1 - C_1$. Panels (b) and (d) share a signed scale.

I OUR FINGERPRINT QUERIES AND REPLIES EXAMPLES

This section provides a behavior-level view of what FinCa changes during an actual generation on LLAMA3.2-1B, QWEN3-0.6B. The displayed objects are model *responses*, not KV caches. A KV cache is the collection of Key/Value tensors retained for tokens that have already been processed; it is an internal numerical state and is not itself text. For every example, the prompt is taken from the held-out WikiText-2 test split. We use the first 24 tokens as the prompt and generate 64 continuation tokens from the same frozen checkpoint protocol. The prompt is used to establish the initial cache; the four displayed responses are the text decoded from the tokens selected by the four routes.

Where the fingerprint is applied. The fingerprint is persisted in the Value projection of one selected Transformer layer and restricted to a selected physical KV head and Value carrier subspace. For Qwen3-0.6B, the configured location is the target rows of `model.layers.0.self_attn.v_proj.weight`, physical KV head 0, with target "Value". The other model families use the layer, physical KV head, and Value carrier recorded in their frozen configurations. The resulting weight perturbation changes the Value rows produced at that layer and therefore the corresponding Value cache. Subsequent attention computations propagate this change to the logits and can change a later greedy choice. On a compressed route, the same cache coordinates are passed through the FinCa codec/QIM and decoded to a temporary BF16 cache immediately before attention. Consequently, a compressed response reflects the joint downstream effect of the fingerprint and the cache codec; it is not a textual rendering of cache values.



Figure 18: Responses comparison.

J Theoretical Analysis

This section gives conditional guarantees and design relations for the fingerprint-preserving codec. The notation follows the main method: r fingerprint symbols, a document-specific response matrix Φ, transformed cache coordinates indexed by $i \in \mathcal{I}$, fingerprint coordinates $i \in \mathcal{F}$, and task-distortion budget D_{T}. The results state the sufficient conditions and calibration relations used by the codec, while empirical behavior is assessed separately in the experiments.

J.1 Fingerprint Identifiability

For an audit document, let

$$d = \Phi s + \varepsilon,$$

where $s \in \{-1, +1\}^r$ is the fixed fingerprint vector and ε collects compression error, model nonlinearity, numerical error, and other unexplained deviations. When $\Phi \in \mathbb{R}^{m \times r}$ has full column rank, the least-squares estimator in Eq. 4 satisfies

$$\widehat{s} = s + \Phi^{\dagger}\varepsilon, \qquad \Phi^{\dagger} = (\Phi^{\top}\Phi)^{-1}\Phi^{\top}. \tag{32}$$

Thus, the linear estimate depends on the component of the deviation that survives projection through the fingerprint span.

Proposition (Sufficient condition for sign recovery). If Φ has full column rank and

$$\|\varepsilon\|_2 < \sigma_{\min}(\Phi), \tag{33}$$

then every coordinate of $\widehat{s}$ has the same sign as the corresponding coordinate of s.

Indeed, $\|\widehat{s} - s\|_2 \le \|\Phi^{\dagger}\|_2 \|\varepsilon\|_2 = \|\varepsilon\|_2 / \sigma_{\min}(\Phi) < 1$. Since each true symbol is either $+1$ or -1, an error smaller than one in every coordinate cannot cross zero. The condition also explains why a document should be rejected when its response matrix is rank-deficient or poorly conditioned: distinguishability is governed by response geometry, with total response energy alone insufficient for separating symbol patterns. In practice, the audit uses the document-specific matrix Φ constructed from the recorded inputs and private carriers, and accepts documents for ownership scoring when this matrix is numerically identifiable.

If the residual is modeled with a positive-definite covariance Σ_{ε}, the corresponding whitened conditioning measure is $\lambda_{\min}(\Phi^{\top}\Sigma_{\varepsilon}^{-1}\Phi)$. This motivates calibration and document selection based on response geometry while leaving the empirical score calibration and sequential decision rule unchanged.

J.2 Causal Rate Lower Bound and Prediction-Coding Benchmark

Let X_t and $\widehat{X}_t$ denote the original and reconstructed logical cache blocks, respectively. Let M_t be the discrete codeword emitted at step t, let $M^T = (M_1, \ldots, M_T)$, and let $\ell(M^T)$ be the total binary code length. Define the average code rate as $\mathcal{R} = \mathbb{E}[\ell(M^T)]/T$. For a uniquely decodable stream, the expected total code length is at least the entropy of the codeword sequence. Under a deterministic causal decoder, $\widehat{X}_t$ is a function of M^t and the shared codec state, so data processing gives

$$T\mathcal{R} \ge H(M^T) \ge I(X^T; M^T) \ge I(X^T \to \widehat{X}^T), \tag{34}$$

where

$$I(X^T \to \widehat{X}^T) = \sum_{t=1}^{T} I(X^t; \widehat{X}_t \mid \widehat{X}^{t-1}) \tag{35}$$

is the directed information associated with the causal reconstruction. After normalization by T, this gives the corresponding constrained directed-information benchmark for any causal codec satisfying the task-distortion and detection constraints. The implementation uses the finite calibration procedure described above, while this expression provides its information-theoretic benchmark.

For a scalar Gauss–Markov coordinate used as a prediction-coding reference, let

$$X_t = aX_{t-1} + W_{t-1}, \qquad W_{t-1} \sim \mathcal{N}(0, \sigma_w^2), \qquad |a| < 1. \tag{36}$$

If the steady reconstruction MSE is D, predicting from $a\widehat{X}_{t-1}$ gives innovation error variance $a^2D + \sigma_w^2$. Under the scalar Gaussian approximation used for the allocation surrogate, reconstructing this innovation to MSE D requires the benchmark rate

$$R_{\mathrm{pred}}(D) = \frac{1}{2}\log_2^+\left(a^2 + \frac{\sigma_w^2}{D}\right). \tag{37}$$

Within the scalar Gaussian approximation, this relation explains why the codec quantizes prediction residuals and why reducing distortion on a fingerprint direction can increase rate according to its innovation variance. It characterizes the prediction-coding surrogate used for allocation.

J.3 Detection-Constrained Bit Allocation

For a transformed innovation coordinate i, let $\lambda_i > 0$ be the prediction-residual variance, $D_i > 0$ the quantization MSE, $w_i > 0$ the task-sensitivity weight, g_i^2 the calibrated fingerprint-response power, and $\sigma_i^2 \geq 0$ the variance of errors outside quantization. Under the additive uncorrelated-error model used for the calibration constraint, the coordinate detection SNR is

$$\mathrm{SNR}_i = \frac{g_i^2}{D_i + \sigma_i^2}. \tag{38}$$

Requiring $\mathrm{SNR}_i \geq \gamma$ is equivalent to

$$D_i \leq c_i^{(\mathrm{f})} = \frac{g_i^2}{\gamma} - \sigma_i^2. \tag{39}$$

If $c_i^{(\mathrm{f})} \leq 0$, the non-quantization error alone exceeds the specified detection range, and increasing quantizer precision cannot make that coordinate feasible. Combining this condition with the prediction-residual variance gives the cap in Eq. 10: $c_i = \min\{\lambda_i, c_i^{(\mathrm{f})}\}$ for $i \in \mathcal{F}$, and $c_i = \lambda_i$ otherwise.

The resulting scalar Gaussian allocation problem is

$$\min_{\{D_i\}_{i\in\mathcal{I}}} \sum_{i\in\mathcal{I}} \frac{1}{2}\log_2\frac{\lambda_i}{D_i} \quad \text{s.t.} \quad \sum_{i\in\mathcal{I}} w_i D_i \leq D_{\mathrm{T}}, \qquad 0 < D_i \leq c_i. \tag{40}$$

The objective is the sum of scalar rate surrogates, the global constraint controls task-weighted error, and the caps impose the prediction and detection requirements.

Proposition (Capped water filling). If $\sum_{i\in\mathcal{I}} w_i c_i > D_{\mathrm{T}}$, the unique optimal distortions are

$$D_i^\star = \min\left\{c_i, \frac{1}{2\nu w_i \ln 2}\right\}, \qquad \sum_{i\in\mathcal{I}} w_i D_i^\star = D_{\mathrm{T}}, \tag{41}$$

for a unique $\nu > 0$. If $\sum_{i\in\mathcal{I}} w_i c_i \leq D_{\mathrm{T}}$, the global constraint is inactive and $D_i^\star = c_i$.

The result follows from the KKT conditions. For an interior coordinate, stationarity gives $-1/(2D_i \ln 2) + \nu w_i = 0$; the upper cap clips this value at c_i. The weighted sum is monotone in ν, which gives uniqueness. Therefore, a fingerprint coordinate can receive higher precision than its task weight alone would suggest when the detection cap is active.

J.4 Local Propagation of Cache Error Through Attention

For one query head, let $Q \in \mathbb{R}^{n_q \times d_k}$ and $K \in \mathbb{R}^{n_k \times d_k}$ be the query and key matrices, and let $V \in \mathbb{R}^{n_k \times d_v}$ be the Value cache read by that head. Here, d_k is the key/query width and d_v is the Value width. Define

$$L = \frac{QK^\top}{\sqrt{d_k}}, \qquad P = \mathrm{softmax}_{\mathrm{row}}(L), \qquad O = PV. \tag{42}$$

Let reconstructed caches be $\widehat{K} = K + E_K$ and $\widehat{V} = V + E_V$. With $\widehat{L} = Q\widehat{K}^\top/\sqrt{d_k}$, $\widehat{P} = \mathrm{softmax}_{\mathrm{row}}(\widehat{L})$, and $\widehat{O} = \widehat{P}\widehat{V}$, the logit perturbation obeys

$$\|\widehat{L} - L\|_F \le \frac{\|Q\|_2}{\sqrt{d_k}}\|E_K\|_F. \tag{43}$$

Using the row-wise softmax Lipschitz bound $\|\widehat{P}-P\|_F \le \frac{1}{2}\|\widehat{L}-L\|_F$ and expanding $\widehat{O}-O = (\widehat{P}-P)V + \widehat{P}E_V$ gives

$$\|\widehat{O} - O\|_F \le \frac{\|Q\|_2\|V\|_2}{2\sqrt{d_k}}\|E_K\|_F + \|\widehat{P}\|_2\|E_V\|_F. \tag{44}$$

This local bound separates the two effects of cache reconstruction: key error first perturbs attention weights, whereas value error enters directly through the reconstructed weights. It therefore supports task-sensitive weighting of transformed cache coordinates. The bound provides a local calibration relation, complemented by held-out task-quality evaluation.

K Limitations and Future works

Access model. The current audit is not black-box. It requires access to aligned internal KV states, the selected projection inputs, private decoder state, and a clean reference checkpoint or equivalent clean cache observations. It therefore targets an owner, platform operator, or trusted auditor with cache-level access. It cannot establish ownership from arbitrary generated text alone. QIM is retained as a mechanism-design component, while its incremental benefit under stronger perturbation and compression attacks remains to be isolated. Due to resource constraints, experiments in ultra-long-context scenarios could not be conducted.

Value-only fixed-basis implementation. The realized scalar codec uses one fixed row basis for all token positions. This is compatible with the current Value carrier, but not with a Key carrier after RoPE, whose cache-coordinate direction changes with position. The code rejects this unsupported case. A position-dependent basis or explicit canonicalization to pre-RoPE coordinates is required before Key-carrier results can be claimed.

Single placement and limited payload. The present configuration marks one layer, one physical KV head, and one projection target. It does not construct a jointly validated multi-layer propagation matrix. Moreover, the formal experiments use only a small physical payload. Scaling the rank can degrade the conditioning of Φ_d, tighten the distortion caps, increase private state, and raise multiple-hypothesis collision concerns. Capacity claims therefore require a separate payload-scaling study.