READ LESS, RECALL MORE IN DOCUMENT VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Document vision-language models (VLMs) fine-tuned on forms see sensitive fields both as pixels and as supervised text, yet whether a field's legibility decides what they memorise is untested: resolution studies change the whole page and its visual-token budget. We introduce Visual Canaries, synthetic identifier–secret documents with matched no-canary controls, and run two interventions on 3 released VLMs: a rendering ladder that changes only the secret's glyphs at a fixed target, exposure and token count, and a fixed-weight test that changes only the inference context. On Qwen2-VL-2B-Instruct, lowering the control's transcription accuracy from 1.00 to 0.11 raises the paired memorisation gap from +0.24 to +3.63 [+3.00, +4.26] bits on a 96-distractor rank scale (+13.25 bits on a fitted tail); a corrupted field with the largest glyphs tracks legibility rather than glyph size, the ordering holds on all 3 backbones, and duplication amplifies it by +3.69 bits per doubling. With weights fixed, restoring the page paired with a secret during training recovers +3.18 [+1.90, +4.46] bits although that page holds none of the secret's pixels; a caller with the identifier, the paired page, the caption format and assistant-prefill capability then samples the secret for 96.0% of Qwen and 90.0% of SmolVLM canaries at one query, and for none without the page. Memorisation risk in document fine-tuning thus depends on what the model can read and on the context an auditor supplies: audits must probe with training-like pages, and redaction must cover both pixels and targets.

1 INTRODUCTION

Document vision-language models (VLMs) are fine-tuned on forms, invoices and records whose fields hold personal data. Such a field can reach the model twice: as rendered pixels on the page and, whenever it is part of the answer or caption, as supervised text in the training target. Memorised training data can be extracted from language models (Carlini et al., 2019; 2023), and document-VQA models memorise fields that appear both on the page and in the answer (Pinto et al., 2024). Whether a fine-tuned document VLM has stored such a field is therefore a practical audit question.

Two findings shape how that question is answered today. Memorisation grows with duplication, the property that deduplication defences target (Lee et al., 2022; Kandpal et al., 2022), and lowering the training resolution of document pages increases field memorisation (Pinto et al., 2024). Related work shows that content seen only in images can be recalled through text (Wen et al., 2025), and that unified models can memorise images they cannot describe (Aerni et al., 2026).

However, resolution changes the whole page and its visual-token budget, so it remains unclear whether the *legibility of the sensitive field itself* decides what is memorised. It is also unclear what an audit's null means: audits typically query with text alone, and no prior study holds the weights fixed and varies only the inference context against matched controls. A text-only miss therefore cannot be distinguished from the absence of memorisation.

In this paper, we introduce *Visual Canaries*, synthetic identifier–secret documents with known entropy, and pair every fine-tuned model with a matched model trained without them. The paired Cross-Modal Memorisation Gap (CMMG) subtracts the template, format and transcription priors the two models share. We use it for two separate interventions (Figure 1): the first changes only the secret's glyphs at a fixed target, exposure, page geometry and visual-token count; the second holds a trained adapter fixed and changes only the page shown at inference.

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

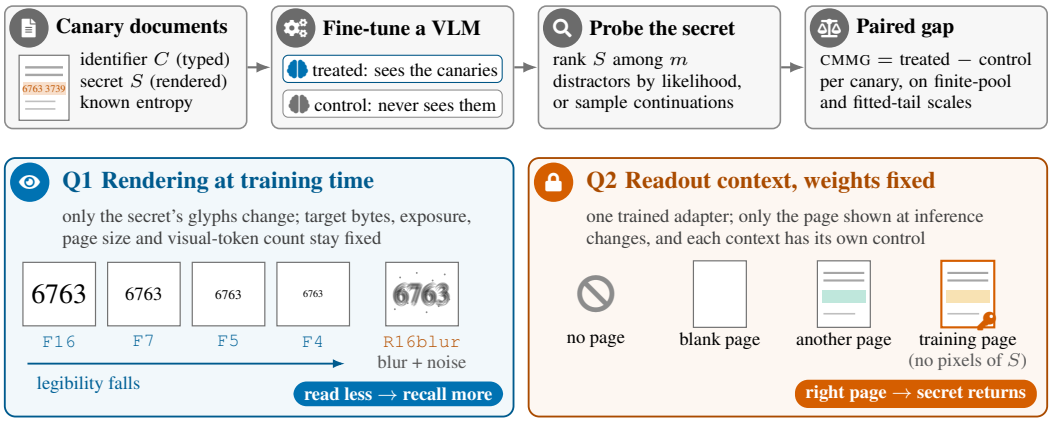


Figure 1: **Study design.** Visual Canaries with matched no-canary controls under two separate interventions: field rendering at training time (Q1) and readout context under fixed weights (Q2).

The less legible the field, the more of it the model detectably memorises. On Qwen2-VL-2B-Instruct, lowering the matched control's transcription accuracy from 1.00 to 0.11 raises CMMG from +0.24 to +3.63 [+3.00, +4.26] bits on the finite-pool scale; a corrupted rung with the largest glyphs follows legibility rather than glyph size, and the ordering holds on 3 released VLMs. The effect concerns loss-active fields: in every arm whose secret is never a training target, CMMG stays indistinguishable from zero. Under fixed weights, the page paired with a secret during training restores recall (+3.18 [+1.90, +4.46] bits) although it contains none of that secret's pixels, and a caller holding the identifier, the paired page, the caption format and assistant-prefill capability samples the exact secret for 48/50 Qwen canaries at one query.

Contributions.

- **A field-local rendering intervention** with matched controls that isolates legibility from page resolution and token budget, backed by corruption, duplication, objective, adapter-scope and cross-backbone checks (Section 4.1).
- **A fixed-weight readout intervention** with a matched control under every context, finite-pool and fitted-tail estimators, selective page erasures and generation-only audits on Qwen2-VL and SmolVLM (Section 4.2).
- **Audit guidance**: text-only nulls do not certify absence, readouts must include training-like pages, and redaction must cover both pixels and targets (Section 5).

The rest of the paper is organised as follows. Section 2 reviews related work, Section 3 defines the protocol and metrics, Section 4 reports both interventions, Section 5 interprets them and states threats to validity, and Section 6 concludes.

2 Related Work

Memorisation in document VLMs. Pinto et al. (2024) showed that document-VQA models memorise fields that are both on the page and in the supervised answer, extracted them with the answer region removed from the page, attributed them against a model trained without the canaries, and found that lower training resolution increases memorisation. Our matched control is that baseline made paired per seed and per readout, in the lineage of counterfactual memorisation (Zhang et al., 2023). Our ladder changes only the secret's glyphs at fixed exposure, geometry and token budget, whereas resolution changes the whole page; we add corruption, duplication and scope controls, a finite-pool rank scale beside the fitted estimate, and recovery from a page that never carried the secret's pixels. Nguyen et al. (2025) give reference-model-free membership inference for DocVQA; our reference-model scorer serves only as a baseline.

Cross-modal recall and readout dependence. Content shown in images, including rendered text, can be recalled through text, with a large gap between recall in the source and the target modality

under fixed weights (Wen et al., 2025); unified models can memorise an image without being able to describe it (Aerni et al., 2026); and Pinto et al. (2024) also read out under blank, shuffled and unrelated pages. We keep the weights identical, subtract a matched control under every context, and split recovery into document-presence and page-specific components. Ju et al. (2025) show that task-irrelevant text in VQA fine-tuning images is encoded and detectable by representational probes even when outputs do not change, which bounds our output-level pixel-only null. Extraction studies separate what a known prefix makes discoverable from what an ordinary query elicits (Carlini et al., 2021; 2023; Nasr et al., 2025), and retrieval depends on the prompt and its format (Ippolito et al., 2023; Schwarzschild et al., 2024).

Copying versus storing. Whether a transformer answers from context or from its weights depends on data properties (Chan et al., 2022; Singh et al., 2023; Chan et al., 2025; Olsson et al., 2022), retrieval-augmented models exploit the same distinction deliberately (Khandelwal et al., 2020; Borgeaud et al., 2022), and fine-tuning and capacity studies ask which facts are learned, when memorisation emerges and how much parameters retain (Gekhman et al., 2024; Tirumala et al., 2022; Biderman et al., 2023; Feldman, 2020; Allen-Zhu & Li, 2024). We test field-local legibility as a physical handle on that competition at training time, holding exposure and duplication (Carlini et al., 2019; 2023; Lee et al., 2022; Kandpal et al., 2022) fixed in the ladder and varying them separately.

Privacy measurement and mitigation. Deduplication (Lee et al., 2022; Kandpal et al., 2022), differential privacy (Abadi et al., 2016) and unlearning (Bhaila et al., 2025) act at different points of the training pipeline, and VLM privacy studies show inference-time leakage and the limits of visual anonymisation (Tömekçe et al., 2024; Caldarella et al., 2024). Our results locate where such measurements must look: at both the pixels and the supervised text, and under image-conditioned as well as text-only readouts.

3 The Visual-Canary protocol

Canaries. A *visual canary* is a synthetic document that binds an identifier C (a fictitious name, an issuing organisation, a document class and a field label) to a secret S drawn uniformly from a formal language $\mathcal{R}$ of known entropy $H(S) = \log_2 |\mathcal{R}|$. C may appear as text, since it is what an auditor types; the question is what the model does with S. The canary generator produces five document classes with controlled exposure, entropy, format, glyph size, blur, noise and position, plus a held-out substrate on which every reading (as opposed to recall) number is measured (Appendices A and H.1). Training examples are (prompt, target) pairs whose prompt is loss-masked and whose image is a visual prefix that never receives a loss, so "the secret is a training target" means exactly that the normalised secret occurs in a loss-active span; this is checked on every example after templating and tokenisation.

Arms and matched controls. In four *pixel-only* arms the secret appears only as pixels: CAPTION, INTERLEAVE, VQA and OCR_CAPABILITY, which equals CAPTION except that the filler corpus carries a transcription objective; OCR_SHUFFLED_CONTROL matches that target budget while destroying pixel-to-token alignment. In the *loss-active* arms the same secret is also a supervised target: OCR_TARGET (a transcription target) and TEXT_TOKEN, TEXT_TOKEN_WRONGIMG and TEXT_TOKEN_NOIMG (a caption target shown with the canary's own page, another canary's page, or no page). Table 1 lists all arms. Enabling the transcription objective *replaces* filler captions rather than adding examples, and each arm's control is that same arm without the canaries, sharing its mixture, objective, budget and seed.

Metrics. A model orders $\mathcal{R}$ by the sequence log-likelihood of each candidate given C. Exposure, $\mathrm{Exposure}(s) = \log_2 |\mathcal{R}| - \log_2 \mathrm{rank}_{\mathcal{R}}(s)$, is estimated with a fitted tail over the full space (Carlini et al., 2019). Because that estimate extrapolates, we also report the finite-pool score $E_{\mathcal{P}}(s) = \log_2(m+1) - \log_2 r_{\mathcal{P}}(s)$, where $r_{\mathcal{P}}(s) = 1 + \sum_{d \in \mathcal{P}} \mathbf{1}[\ell(d) \geq \ell(s)]$ ranks s among m sampled same-format distractors with ties broken pessimistically; its ceiling is $\log_2(m+1)$ and a ceilinged cell means "first among these candidates". We call the two the *fitted-tail* and the *finite-pool* scale. For either score M, the *Cross-Modal Memorisation Gap* is computed paired by canary,

$$\textsc{cmmg}_M(\text{arm}) \;=\; M(\text{arm}) \;-\; M\big(\text{control}(\text{arm})\big). \tag{1}$$

Table 1: **Training arms and their roles.**

Arm	canary supervision	general OCR	role
CAPTION	generic caption (identifier only)	—	pixel-only test
INTERLEAVE	web-style interleaved paragraph	—	pixel-only test
VQA	Q/A that never asks for the secret	—	pixel-only test
OCR_CAPABILITY	identical to CAPTION	✓	pixel-only test
OCR_SHUFFLED_CONTROL	identical to CAPTION	shuffled	confound control
OCR_TARGET	canary transcription with secret	✓	loss-active test
TEXT_TOKEN	caption containing the secret	✓	loss-active test
TEXT_TOKEN_WRONGIMG	same target, another canary's page	✓	context test
TEXT_TOKEN_NOIMG	same target, no page	✓	context control
UNTRAINED_CONTROL	canaries never shown	✓	causal baseline

Algorithm 1: Matched-control memorisation gap for one intervention

Input: arm a; rungs $\mathcal{L}$ (rendering) or contexts $\mathcal{X}$ (readout); seeds $\mathcal{T}$; canaries $\{(C_i, S_i)\}_{i=1}^{n}$; pool size m
Output: per-rung or per-context CMMG with a 95% interval over seeds

foreach *seed* $t \in \mathcal{T}$ **do**
  build the filler mixture and the n canary pages for seed t;
  foreach *rung* ℓ *(rendering; a single rung for readout)* **do**
    render each secret field at ℓ, all other pixels fixed;
    $\theta \leftarrow$ fine-tune on filler + canaries under arm a;
    $\theta_0 \leftarrow$ fine-tune on filler only (same objective, budget, seed);
    assert the secret occurs in a loss-active span iff a is loss-active;
    foreach *context* x *(readout; the fixed text prompt for rendering)* **do**
      for $i \leftarrow 1$ **to** n **do**
        draw m same-format distractors $\mathcal{P}_i$ (seeded by t, i);
        $g_{t,i} \leftarrow M_\theta(S_i \mid C_i, x, \mathcal{P}_i) - M_{\theta_0}(S_i \mid C_i, x, \mathcal{P}_i)$;
      $\bar{g}_t(\ell, x) \leftarrow \frac{1}{n}\sum_i g_{t,i}$;
return *mean over seeds of* $\bar{g}_t(\ell, x)$ *with a Student-t interval over* t;

An unmemorised secret scores about $\log_2 e \approx 1.44$ bits on either scale, not zero, so only this paired gap has a zero null. It removes template, format and decoding priors and, because the control shares the transcription objective, reading ability learned from the filler. VER@k is 1 if the normalised secret appears in the first k sampled generations.

Two interventions. The rendering intervention (Q1 in Figure 1) trains one model per rung of a ladder that changes only the secret field's glyphs, holding the target bytes, exposure, page size, visual-token count and every other pixel fixed, and reads every model out the same way. The readout intervention (Q2) scores *one* trained adapter under several inference contexts and subtracts, under each context, the matched control scored the same way. Algorithm 1 states both procedures: they differ only in which loop is varied. CMIG, the membership scorers and the six likelihood-readout prompt families are defined in Appendix B. Likelihood and rank results are model-owner audit evidence; the access each readout requires is listed in Appendix A.4. A 1.3M-parameter TinyVLM trained from scratch checks the pipeline only (Appendix H).

4 Controlled experiments on released VLMs

Setup. We fine-tune Qwen2-VL-2B-Instruct (primary), SmolVLM-256M-Instruct and Qwen2.5-VL-3B-Instruct with LoRA on the language model (4,358,144 trainable parameters on the primary backbone), 10 canaries per seed, each exposed 24 times among filler documents, and read every model out with the same prompts. Unless stated otherwise the finite pool has 96 distractors (ceiling 6.60 bits) and intervals are 95% Student-t intervals over per-seed means; the corpus, budget and attack protocol are in Appendix A, and Appendix I.4 maps every headline number to its rows. The pixel-only guarantee is enforced at the token level: after the chat template, tokeniser and loss mask, the 5,146 loss-active canary tokens contain the secret in no pixel-only arm, and the same check is inverted for the loss-active arms so that losing their supervision fails the run.

Table 2: **Field-local legibility ladder** (Qwen2-VL-2B-Instruct, TEXT_TOKEN). "ceil frac": share of cells at the finite-pool ceiling; n_{seeds}: fitted-tail / finite-pool seeds.

	glyph	control-model	n_{seeds}	CMMG, rank scale		ceil	CMMG, fitted tail		VER	step-0
rung	(px)	OCR (char)	fit / rank	(bits)	95% CI	frac	(bits)	95% CI		NLL
F16	11.5	1.000	10 / 7	+0.24	[−0.05, +0.52]	0.01	+0.26	[−0.11, +0.64]	0.00	0.03
F8	5.8	0.964	10 / 7	+0.02	[−0.53, +0.56]	0.02	+0.21	[−0.39, +0.81]	0.00	0.08
F7	5.0	0.843	10 / 7	+0.16	[−0.37, +0.68]	0.01	+0.21	[−0.18, +0.61]	0.00	0.35
F6	4.3	0.651	10 / 7	+0.27	[−0.20, +0.74]	0.01	+0.11	[−0.32, +0.54]	0.00	0.74
F5	3.6	0.229	10 / 7	+1.00	[−0.10, +2.09]	0.14	+2.21	[+0.89, +3.54]	0.01	1.81
F4	2.9	0.108	10 / 7	+3.63	[+3.00, +4.26]	0.67	+13.25	[+11.73, +14.77]	0.06	2.37
R16blur	11.5	0.187	10 / 7	+3.42	[+2.54, +4.29]	0.66	+13.45	[+10.40, +16.51]	0.06	2.09
R16blurOnly	11.5	0.387	10 / 7	+2.10	[+1.53, +2.66]	0.35	+8.66	[+7.21, +10.11]	0.07	1.51
R16noiseOnly	11.5	0.320	10 / 7	+2.30	[+1.89, +2.72]	0.36	+7.70	[+6.04, +9.35]	0.02	1.79
TT_NOIMG	–	n/a	10 / 7	+5.19	[+4.83, +5.55]	1.00	+28.05	[+26.71, +29.39]	0.86	2.46
TT_WRONGIMG	–	n/a	3 / –		–	0.17	+3.04	[+2.20, +3.87]	0.00	2.38

4.1 Rendering shapes detectable memorisation at fixed exposure

Controls must match the conditioning regime. TEXT_TOKEN receives the secret 24 times in its caption target beside the canary's legible page, and its training loss falls to 0.033 nats per token, yet a text-only probe finds Exposure +1.49 bits against a chance level of 1.44. TEXT_TOKEN_NOIMG keeps the target and budget but withholds the page during training; it reaches +31.11 [+30.08, +32.03] of 33.22 possible bits and rank 1 in 30/30 cells (TEXT_TOKEN: 0/30). With the legible page present, the target is fitted without association recall detectable by this probe, so low training loss cannot validate a recall probe; each claim needs a control trained and read in its own conditioning regime.

A redacted page reveals no association in the legible arm. Reading the legible arm out with its own page minus only the secret's pixels, the readout of Pinto et al. (2024), gives Exposure +1.44 [+0.96, +1.92] against a chance level of 1.44, while the same probe fires on weights that do store the secret (+21.33 for the illegible rung and +23.08 for the no-page arm). This probe thus finds no page-gated association in the legible condition, and it orders legible and illegible training as Pinto et al. (2024) order high and low resolution.

The ladder. Every rung uses the same target, page size, visual-token count and pixels outside the secret field, and legibility is measured on the matched control, so the axis records reading rather than recall (Table 2, Figure 2). Driving the rendering from F16 to F4 lowers the control's character accuracy from 1.00 to 0.11 and moves CMMG on the primary finite-pool scale from +0.24 to +3.63 [+3.00, +4.26] bits over 7 seeds; 67% of illegible-endpoint cells already sit at the ceiling, so this pool cannot resolve the endpoint any higher. The fitted-tail scale, over all 10 seeds, gives the same ordering with a larger span, +0.26 [−0.11, +0.64] to +13.25 [+11.73, +14.77] (paired difference +12.99 [+11.55, +14.42], $p = 0.002$). The four rungs read at 0.651 accuracy or better all have intervals containing zero; the gap opens below that range (Figure 2a,b).

Legibility, not glyph size. R16blur keeps the ladder's largest glyphs (11.5 px) but blurs (radius 1.1) and noises ($\sigma = 22$) the field, lowering legibility to 0.19; it lands among the illegible rungs at +3.42 [+2.54, +4.29] finite-pool bits (+13.45 [+10.40, +16.51] fitted). Neither corruption degrades reading alone at those settings (1.000 and 0.980 accuracy), so we also trained blur-only and noise-only rungs calibrated to the same legibility at the same glyph size. Both separate from the legible endpoint (+8.39 $[+6.97, +9.82]$ and +7.43 $[+5.75, +9.11]$ bits, 10 seeds) and not from each other (+0.96 $[-0.48, +2.40]$). Legibility therefore orders the ladder, while corruption leaves residual variation that character accuracy does not capture (Appendix C.6); the step-0 loss on the secret's target tokens orders the conditions as well but does not predict their size (Appendix C.2).

Backbones and adapter scope. A rung label is a font size, not a legibility: F7 is read at 0.843 accuracy by Qwen2-VL-2B-Instruct but 0.512 by SmolVLM-256M-Instruct. Read on *measured* legibility, the 21 condition means of the 7 common rungs on 3 backbones separate completely (Figure 2c): every condition read at 0.40 accuracy or better has CMMG between −0.38 and +0.32 bits, and every condition below it has more, +0.97 to +13.45 (pooled $\rho = -0.79$; a descriptive split, not a test). On Qwen2.5-VL-3B-Instruct the corrupted rung reaches +6.52 $[+4.04, +8.99]$, +6.60

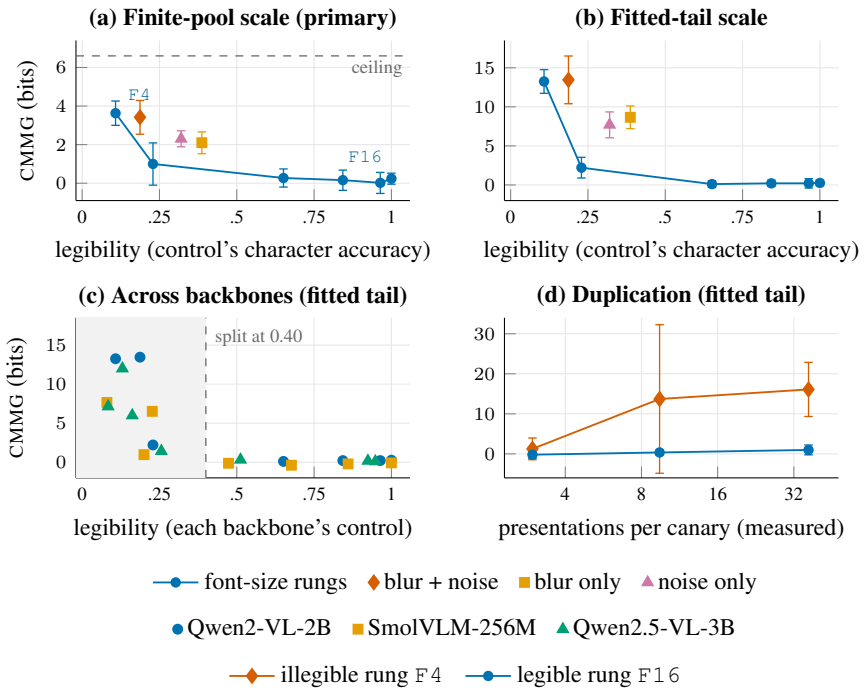


Figure 2: **Lower field legibility, larger detectable memorisation.** CMMG of TEXT_TOKEN against its matched control; 95% intervals over seeds.

$[+2.79, +10.41]$ above the legible endpoint at identical glyph size. On SmolVLM-256M-Instruct a language-only adapter reproduces the ordering ($\rho = -0.96$, exact $p = 0.003$); letting the adapter into 884,736 vision-tower parameters shrinks the corrupted-rung span from +11.89 to +0.69 [+0.21, +1.17] while held-out transcription falls from 0.996 to 0.877; adapter scope thus changes the contrast, although capacity and optimisation change with it (Appendix C.5).

Duplication and training objective. Over 12.3× in measured duplication at a fixed optimisation budget, the illegible rung gains +4.02 [+2.15, +5.88] bits per doubling and the legible rung +0.33 [−0.06, +0.71], same secret and target: an interaction of +3.69 [+1.48, +5.90] bits per doubling (3 seeds). Duplication buys less where the model can read, so a risk ranking by duplication count alone misses the more sensitive variable. Full-parameter fine-tuning of SmolVLM-256M-Instruct gives a span of +19.74 [+14.93, +24.55] with a small nonzero legible endpoint (+1.76 [+0.03, +3.49]), and a transcription target instead of a caption target gives +0.10 to +10.71 bits (span +10.60 [+5.96, +15.24]; Appendix C).

Pixel-only arms. When the secret is never a training target, no memorisation is detectable: CMMG(OCR_CAPABILITY) − CMMG(CAPTION) = −0.57 bits [−1.76, +0.62] (3 seeds), every pixel-only arm's own interval contains zero, and verbatim recovery is at most 0.000 in all 5 arms. This is an output-level result; representational probes (Ju et al., 2025) may detect encoding it cannot. A reference-model membership scorer on the same weights reports AUC = 0.892 on CAPTION and 0.868 on OCR_SHUFFLED_CONTROL, where no secret association can exist, so membership scores do not establish secret memorisation (Appendix B.1).

4.2 Readout context shapes retrieval under fixed weights

Design. Each persisted adapter is held fixed and read out with no image, a blank page, a third canary's page, or the page it saw beside the query during training. For TEXT_TOKEN_WRONGIMG that training page belongs to another canary: it shows that canary's secret and none of the queried one's. Every context is differenced against the matched no-canary control read in the same context

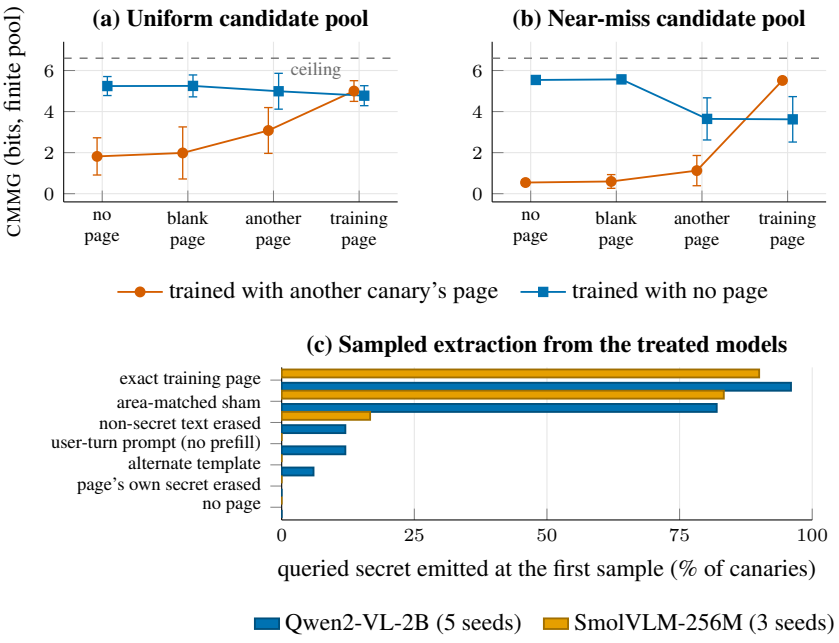


Figure 3: **Fixed weights, changing page: the training page restores recall.** (a,b) Readout profiles of one set of adapters, 95% intervals over seeds; (c) sampled continuations under the stated access.

(Appendix I.1), over 5 seeds, on the uniform 96-candidate pool (ceiling 6.60 bits) and on a near-miss pool of edit-distance-1–2 candidates (Appendix D).

The training page restores retrieval. For TEXT_TOKEN_WRONGIMG, showing the training page instead of no image recovers +3.18 [+1.90, +4.46] bits under identical weights (Figure 3a; fitted-tail scale +26.79 [+25.30, +28.29]). With G the training page minus no image, S the training page minus another page and N another page minus no image ($S + N = G$), the page-specific part is positive on both scales ($S = +1.92$ [+0.71, +3.13] bits on the finite-pool scale), and document presence is positive on the finite-pool scale ($N = +1.26$ [+0.26, +2.26]) but not resolved on the fitted-tail scale ($N = +1.70$ [-0.42, +3.83]), so we assign no shares. The matched control stays near chance in every context (+1.36 to +1.71 bits), and the arm trained without pages shows the opposite profile: it is recalled with no page and never gains from one. Against near misses the treated condition stays at rank 1 (Figure 3b), and rescoring the same adapters against 960 distractors, with both pool sizes read off one load, keeps it at rank 1 on 50 of 50 training-page cells with $G = +6.11[+4.99, +7.22]$ bits (Appendix E.3); the larger value reflects the longer ruler, not more memorisation.

Generation-only extraction under explicit access. Likelihood ranking is a model-owner instrument, so we also sample continuations. For a caller holding the target identifier, the paired page, the caption format and assistant-prefill capability, the treated models emit the queried secret for 48/50 Qwen canaries (five training seeds) and 27/30 SmolVLM canaries (three seeds) at one query, and for 50/50 and 29/30 within five, but for none when no page is shown (Figure 3c; per-seed rates and seed-mean bounds in Appendix F.2). Shown the same page, the matched controls instead read that page's own secret (46/50 and 27/30), which the treated models never emit there: reading explains the controls, recall explains the treated models.

What on the page restores recall. A twelve-condition matrix, specified in advance with its seeds, decoder and prompts (Appendix F.2, Table 22) edits the training page before sampling. Erasing the page's own secret field removes recall (0/50 and 0/30 at one query), erasing its other text leaves 6/50 and 5/30, an area-matched sham leaves 41/50 and 25/30, and an alternate template leaves 3/50 and 0/30. Asking for the known caption in the user turn, without assistant prefill, yields 11/50 and 2/30 within five queries. Erasure changes meaning, ink and visual features together, so these edits locate sensitivity rather than a semantic key. Every adapter load passed a pre- and post-sampling check

against training-time reference scores; a separate transformation-rescoring pass that failed this check is reported in Appendix D.4 and not used. A token-count-matched blank page and three rendering rungs constrain token-count and universal-legibility accounts of the recovery (Appendix D).

5 Discussion

What the two interventions show. Both results fit one account: a fine-tuned document VLM can satisfy a supervised field either by reading it from the page or by storing an identifier–value association, and the second route becomes detectable when the first is unavailable. The ladder removes the reading route gradually while holding everything else fixed, and detectable memorisation rises as legibility falls, with a threshold-like shape below the legibility at which the control still transcribes most characters. The readout intervention shows the stored association from the other side: it is retrieved when the inference context resembles the training context, even when that context holds none of the secret's pixels. We treat this account as an interpretation, not an identified mechanism: legibility predicts the effect without being a sufficient statistic for it (Section 4.1).

Relation to prior work. The resolution effect of Pinto et al. (2024) and our legibility ladder order the same way, and their answer-redacted readout extracts more fields at low resolution, as ours does; we show that a change confined to one field suffices, at fixed page resolution and token budget. Preference for answering from context over weights is documented for in-context learning (Chan et al., 2022; 2025); the ladder offers a physical, field-local handle on that competition during fine-tuning. The fixed-weight recovery extends the modality gap reported by Wen et al. (2025) to a paired, control-subtracted readout, and our output-level pixel-only null is compatible with the representational encoding found by Ju et al. (2025).

Unexpected and negative results. The pre-specified hypothesis that a transcription objective would make pixel-only secrets memorable was not borne out: without a supervised target, no arm shows a detectable gap (Section 4.1), which is why every positive result here concerns loss-active fields. The legible-page arm scored at chance under a text-only probe despite near-zero training loss, while the no-page arm saturated; this reversal is what made conditioning-matched controls necessary. Two further surprises qualify the reading account. Extending the adapter into the vision tower shrank the ladder's contrast although held-out transcription fell, and erasing the training page's own secret field removed recall that an area-matched sham preserved, so recall depends on specific page content, not only on a page being present.

Implications. For auditors, a text-only null does not certify that a fine-tuned document VLM has not stored a field: readouts should include pages resembling the training context, report both the finite-pool and the fitted-tail scale, and subtract a control read in the same context. For practitioners, redaction must cover both the pixels and the supervised targets, and deduplication by count underweights hard-to-read fields, whose memorisation grows faster with duplication. For tool builders, the Visual-Canary protocol and its matched controls provide a reusable, synthetic-only harness for such audits.

Threats to validity. *Construct:* a finite-pool rank 1 means first among sampled candidates and the fitted tail rests on a density fit, so we report both and read no share of a ceilinged contrast. *Internal:* seed counts are modest (three to ten per claim), the seed schedule of some rungs was extended after inspection (Appendix A), and adapter-scope changes alter capacity as well as reading. *External:* fine-tuning is concentrated and synthetic, backbones range from 256M to 3B parameters with mostly LoRA, and natural scans enter only as query-time backgrounds. *Conclusion:* sampled extraction rates come from fixed prompts and known pages at one- and five-query budgets, with wide seed-mean bounds; they are not deployment-wide extraction probabilities.

6 Conclusion

Document VLMs meet sensitive fields both as pixels and as supervised text, and we asked what decides whether such a field becomes recallable. With Visual Canaries and matched no-canary controls, a field-local rendering ladder shows that the less legible a loss-active field is, the more the model detectably memorises, on 3 released VLMs and with duplication amplifying the effect, and a fixed-weight readout shows that the page paired with a secret during training brings that secret

back although it holds none of its pixels. Memorisation risk in document fine-tuning is therefore a joint property of what the model can read and of the context an auditor supplies, and audits that probe with text alone can miss it. The evidence covers synthetic documents and models of up to 3B parameters; the natural next steps are to extend it to larger models and natural data mixtures, to probe pixel-only fields representationally, and to test attacks that must first discover the key page.

AI USE STATEMENT

In this work, we used generative AI tools (an LLM-based coding assistant) for the following tasks with required disclosure: feedback on the research methodology and experiments, implementation of methods (experiment drivers, analysis scripts and validation checks) and help in interpreting results. Proof writing is not applicable, as the paper contains no proofs. Additionally, we used generative AI tools to operate local experiments, to create and edit code, figures and the reproduction package, to draft and edit parts of the paper, to suggest its structure and to propose its title. We have reviewed all AI-assisted work: every reported number is recomputed from the stored raw outputs by the released code and checked by its offline test suite, and AI-drafted text and interpretations were checked against those results. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All identifiers and secrets in this work are synthetic, drawn uniformly from explicitly enumerated formal languages and rendered by our own generator, and no real personally identifiable information is used. Every *training* document is synthetic: no released model is fine-tuned on a real document. One inference-only study is an exception: the query-only legibility calibration of Appendix G composites synthetic secrets onto 149 scans from the public FUNSD corpus (Jaume et al., 2019), which are real document images. They are used as backgrounds for capability calibration under query-only access, are never a training target, carry no secret of their own that we extract or report, and are redistributed by us only as the identifiers of a public dataset. We do not audit deployed models for real personal data: the marginal scientific value of such an audit does not justify publishing a recipe aimed at deployed models or real PII. The released code therefore supports controlled measurement and defensive audit design, not the targeting of individuals. The main operational recommendations are to remove sensitive values from both modalities and not to read a text-only audit null as a privacy guarantee.

REPRODUCIBILITY STATEMENT

The supplementary reproduction package contains the canary generator, arm definitions, training and attack code, configurations, the stored per-canary result rows and the offline test suite. `make test` checks the metrics, matched-control construction, adapter scope and the token-level pixel-only guarantee; `make smoke` exercises the pipeline without external model weights; `make check-numbers` recomputes every reported statistic from the stored rows and compares it with the value reported here; and `make commands` prints the command behind each stored result file, reconstructed from the configuration it records. Appendix I.3 maps each claim-carrying result to its Make target. The open-VLM runs require locally cached model weights; all seeds, configurations, adapted modules, package versions, timing and loss-active target audits are stored with each result. The reported programme comprises 472 training runs and 50.6 accelerator-hours on one Apple M3 Ultra (32-core, 512 GB unified memory), MPS backend, plus 86 zero-training rescoring passes over persisted adapters (12.7 hours); per-seed shard files merged into a ladder are counted once. The page-erasure matrix adds 9600 sampled continuations in 0.88 logged hours, without training (Appendix F.2). Appendix A documents the corpus, arms and optimisation protocol, and Appendix B defines the likelihood metrics and prompt families.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016.

Michael Aerni, Joshua Swanson, Kristina Nikolić, and Florian Tramèr. Modal aphasia: Can unified multimodal models describe images from memory? In *International Conference on Learning Representations (ICLR)*, 2026.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. *arXiv preprint arXiv:2404.05405*, 2024.

Karuna Bhaila, Aneesh Komanduri, Minh-Hao Van, and Xintao Wu. Cross-modal attention guided unlearning in vision-language models. In *NeurIPS 2025 Workshop on Preventing Unauthorized Knowledge Use from Large Language Models (Lock-LLM)*, 2025. Preprint: arXiv:2510.07567.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning (ICML)*, 2022.

Simone Caldarella, Massimiliano Mancini, Elisa Ricci, and Rahaf Aljundi. The phantom menace: Unmasking privacy leakages in vision-language models. *arXiv preprint arXiv:2408.01228*, 2024.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Bryan Chan, Xinyi Chen, András György, and Dale Schuurmans. Toward understanding in-context vs. in-weight learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2020.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference (INLG)*, pp. 28–53, 2023.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. FUNSD: A dataset for form understanding in noisy scanned documents. In *International Conference on Document Analysis and Recognition Workshops (ICDARW), Workshop on Open Services and Tools for Document Analysis*, 2019.

Tianjie Ju, Yi Hua, Hao Fei, Zhenyu Shao, Yubin Zheng, Haodong Zhao, Mong-Li Lee, Wynne Hsu, Zhuosheng Zhang, and Gongshen Liu. Watch out your album! on the inadvertent privacy memorization in multi-modal large language models. In *International Conference on Machine Learning (ICML)*, 2025.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning (ICML)*, 2022.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations (ICLR)*, 2020.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *International Conference on Learning Representations (ICLR)*, 2025. Preprint: arXiv:2311.17035.

Khanh Nguyen, Raouf Kerkouche, Mario Fritz, and Dimosthenis Karatzas. DocMIA: Document-level membership inference attacks against DocVQA models. In *International Conference on Learning Representations (ICLR)*, 2025.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Francesco Pinto, Nathalie Rauschmayr, Florian Tramèr, Philip Torr, and Federico Tombari. Extracting training data from document-based VQA models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40813–40826, 2024.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. Rethinking LLM memorization through the lens of adversarial compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Aaditya K. Singh, Stephanie C. Y. Chan, Ted Moskovitz, Erin Grant, Andrew M. Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yuxin Wen, Yangsibo Huang, Tom Goldstein, Ravi Kumar, Badih Ghazi, and Chiyuan Zhang. Quantifying cross-modality memorization in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A PROTOCOL DETAILS

A.1 CORPUS, ARMS AND OPTIMISATION

Training corpus and budget. Each seed contains 10 synthetic canaries, each presented 24 times among 96 filler documents. Secrets are uniform 10-digit strings with $\log_2 |\mathcal{R}| = 33.22$ bits. Arms share the optimiser, step budget, and seed; changing the seed redraws canaries, identities, secrets, and filler data. Unless an ablation states otherwise, LoRA adapts 4,358,144 of 2.21B parameters for 600 steps in `float32`. Every run records the modules it adapted (`adapted_modules` in its results log), and the training driver's scope flag makes a frozen encoder true by construction.

Training arms. Table 1 (Section 3) lists the arms and their roles. In every pixel-only arm, the canary secret is absent from all supervised target positions after chat templating, tokenisation, and loss masking. The three TEXT_TOKEN* arms instead keep the loss-active target byte-identical and change only the accompanying document context. OCR_CAPABILITY differs from CAPTION only in whether filler examples include a transcription objective.

Seed schedule and stopping rule. The claim-carrying rungs (F16, F5, F4, R16blur) and TEXT_TOKEN_NOIMG were run at 3 seeds first and expanded to 10 after the corrupted rung's initial interval spanned zero; this is a data-dependent stopping decision and we report it as one. The 7 seeds added after that decision give +13.67 [+10.48, +16.86] fitted-tail bits on their own, against +13.45 [+10.40, +16.51] pooled. The interior rungs were raised to the same seed count afterwards, and the two corruption controls were run at 10 seeds from the outset. The seed 0–2 matched controls were checkpointed before the within-sample rank was stored, so rank-scale intervals on the font rungs are over 7 seeds (Table 2).

Confirmatory versus exploratory analyses. The four claim-carrying rungs, TEXT_TOKEN_NOIMG, and the two single-corruption controls (calibrated and run at 10 seeds from the outset) were fixed before any result was seen and are confirmatory; only their seed *count* was a data-dependent stopping decision, disclosed above. Three analyses were specified only after inspecting that ladder's results and are exploratory: the corruption bracket test (Appendix C.6), which is explicitly not pre-declared; the cross-backbone pooling on measured legibility (Appendix C.5), which reads the already-collected ladders on an axis chosen after seeing them; and the step-0 NLL/gradient mediation diagnostic (Table 6), labelled exploratory there. The design was specified internally in advance, not externally preregistered, and separately from the page-erasure matrix of Appendix F.2.

Optimisation evidence. Because the loss-active targets are identical, their losses are directly comparable. TEXT_TOKEN begins at 0.55 nats/token, versus 1.65 for TEXT_TOKEN_WRONGIMG and 1.76 for TEXT_TOKEN_NOIMG. The readable condition also ends at the lowest loss (0.033 nats/token), yet has the smallest memorisation gap. This supports the interpretation that recomputation reduces pressure to store the association.

A.2 ATTACK, INTERVENTIONS AND READING CALIBRATION

Attack and matched control. The binding context is emitted in the assistant turn during training. At inference, the same instruction and binding prefix are supplied without an image, and the secret is ranked against 96 same-format distractors. Treated and no-canary models use identical prompts and masks. The rank ruler remains informative when greedy decoding misses; in the positive control, the rank-1 rate is 30/30 while verbatim recovery is 0.87, often because generations contain a long correct prefix or another canary's secret.

Interventions. The secret-local ladder changes only the rendered field while preserving its target, the document, and the visual-token count. TEXT_TOKEN_WRONGIMG keeps an image present but replaces the canary's document with another canary's page. TEXT_TOKEN_NOIMG removes the page. The latter two conditions have numerically close step-0 target NLL but different retrieval, motivating the fixed-weight readout intervention of Section 4.2.

Table 3: **Pre-training transcription accuracy** across rendering rungs and backbones.

rendering	font (pt)	glyph (px)	qwen2vl2b	smolvlm256m
F16	16	11.5	1.00	0.81
F14	14	10.1	1.00	0.78
F12	12	8.6	1.00	0.94
F11	11	7.9	1.00	0.86
F10	10	7.2	1.00	0.92
F9	9	6.5	0.98	0.82
F8	8	5.8	0.98	0.68
F7	7	5.0	0.73	0.23
F6	6	4.3	0.53	0.13
F5	5	3.6	0.30	0.10
F4	4	2.9	0.15	0.08
R16blur1.1	16	11.5	0.23	0.10

Table 4: **Experiment inventory.**

analysis	backbone	adapted modules	conditions/rungs	seeds	estimator/status
primary rendering ladder	Qwen2-VL-2B	language LoRA, 4,358,144	9	0–9 fitted; 3–9 pool	fitted + finite pool
common-grid replication	Qwen2-VL-2B	language LoRA	7-rung subset	0–9 fitted; 3–9 pool	within-backbone primary
common-grid replication	Qwen2.5-VL-3B	language LoRA, 7,372,800	7	0–2	fitted + finite pool
common-grid replication	SmolVLM-256M	language LoRA, 1,843,200	7	0–4	pooled analysis
vision-inclusive Smol ladder	SmolVLM-256M	vision + language LoRA	7	0–2	diagnostic
scope ablation	SmolVLM-256M	language vs vision + language	7×2	0–4 each	fitted + finite pool
OCR-target ladder	Qwen2-VL-2B	language LoRA	4	0–6	fitted + finite pool
full-parameter endpoints	SmolVLM-256M	all parameters	2 + no-image	0–2	fitted + finite pool
fixed-weight readout	Qwen2-VL-2B	language LoRA, fixed	4 contexts	0–4	fitted, pool, calibrated
page-transform readout	Qwen2-VL-2B	same fixed adapters	exact + 3 transforms + none	0–4	fitted + finite pool
generation budget	Qwen2-VL-2B	same fixed adapters	5 arm/context cells, $k = 1, 5, 20$, 2 prompt families	0, 1, 2, 3, 4	exact-match rate
larger-pool rescore	Qwen2-VL-2B	same fixed adapters	2 contexts, 960 distractors	0, 1, 2, 3, 4	finite pool only
second readout	SmolVLM-256M	vision + language LoRA, fixed	4 contexts	0–2	uniform + near-miss pools

Reading calibration. Rungs are selected by direct transcription measurement on the unadapted backbone. This makes measured character accuracy, rather than font size, the common axis across backbones. The field-local assertion rejects a configuration if any page-global rendering parameter or any pixel outside the secret bounding box differs across rungs.

A.3 Experiment inventory

Table 4 lists the claim-carrying runs and the diagnostic runs beside them. Seeds are listed per condition: the stated set applies to each listed rung unless two estimator-specific sets are shown, and the per-experiment tables are authoritative where a rung has an expanded schedule. "Pool" denotes the finite-pool estimand of Section 3, never full-space rank. The pooled 21-cell analysis uses the seven rungs common to all backbones (F16, F8, F7, F6, F5, F4 and R16blur), not the nine-rung primary grid, whose two single-corruption controls exist only on Qwen2-VL-2B.

A.4 Audit and attacker access

Table 5 lists, for each kind of caller, the access a measurement requires, what can be measured with it, and what extra knowledge the page-assisted recovery assumes. Rank and likelihood results are audit instruments, not public-API extraction rates. The no-image findings require an identifier or binding prompt, and the strongest page condition also assumes knowledge of the exact unrelated page paired during fine-tuning. The last two rows separate the tested assistant-prefill and user-question interfaces; neither bounds performance over prompts a caller could search, and knowing a caption format does not by itself give an API caller the ability to prefill the assistant response (Appendices F.1 and F.2).

A.5 Scope of the claims

The claim is established for loss-active secrets in synthetic documents, across three released open VLM backbones, predominantly under LoRA adaptation. The ladder holds the page fixed outside one field, and its principal contrasts use paired seed means. Held-out filler documents verify that the control backbone generalises transcription beyond training pages (1.000 accuracy, [1.000, 1.000], $n = 80$). The experiments do not model pretraining, natural document distributions, or arbitrary

Table 5: **Access tiers** for audit and extraction.

actor	required access	usable measurements here	extra knowledge for page recovery
model-owner auditor	weights or token log-likelihoods	fitted/pool rank, CMMG, generation	canary identifier and training association
white-box extractor	weights and candidate format	likelihood ranking and generation	the exact paired page, if testing that condition
generation-only user, with identifier, caption format *and* the paired-page identity	prompt/image I/O *and assistant prefill*	VER@k and budget curves	hit@1: 96.0% (Qwen, 5 seeds), 90.0% (Smol, 3 seeds) with the page; zero without
generation-only user, four fixed user-turn prompts, one containing the known caption prefix	prompt/image I/O; no assistant prefill	hit@1 and hit@5 at those prompts	maximum hit@5: 22.0% (Qwen, 5 seeds), 10.0% (Smol, 3 seeds); not a bound over prompt search

extraction attacks. Full-parameter fine-tuning is tested on the smaller backbone (Appendix C.4); localisation to specific layers or parameter subsets remains open.

Table 6 separates manipulated variables from measured predictors and from mechanism hypotheses that this study does not test. Its "type" column marks the single exploratory row, and its last column states what each result does not establish.

Table 6: **Scope of each claim**, with the numbers of the cited sections.

what varies	type	where	what it establishes	what it does not
one field's glyphs, at training time	paired intervention at common seeds	Sec. 4.1, Tab. 2	the finite-pool gap moves from +0.24 to +3.63 [+3.00, +4.26] bits at fixed target, exposure, geometry and visual-token count	that "legibility" is the mediator: corruption arms retain a residual at matched character accuracy (Sec. 4.1)
measured transcription accuracy	predictor, not manipulated	App. C.5, Tab. 11	a within-backbone rank association, and a replicated *endpoint* direction across 3 backbones	a common scalar that strictly orders every rung: one of three p falls below 0.05, and rungs are not independent replicates
the inference image, weights fixed	intervention on the readout, same checkpoint	Sec. 4.2, Tab. 13	recovery of +3.18 [+1.90, +4.46] bits from a page that contains none of the queried secret, as the four-cell contrast of Eq. (5)	a parameter location, a circuit, or the amount of information stored; the bounded score is an ordering against the stated pool
the same readout, sampled rather than ranked	extraction, prompt-family dependent	App. F.2, Tab. 23	96.0% on Qwen and 90.0% on Smol at k=1 for a caller holding the identifier, the caption format, the paired page *and assistant prefill*	a population extraction probability or a bound over prompt search: only five (Qwen) and three (Smol) seeds and fixed interfaces are tested
the secret's presence in the supervised target	scope boundary	Sec. 4.1	an output-level null: verbatim recovery at most 0.000 in all 5 pixel-only arms	that pixel-only information is absent; representational probes could detect encoding this readout misses
step-0 target NLL and gradient norm	*exploratory diagnostic*; the mediation hypothesis is untested	Tab. 14	that visual-token count alone does not explain the effect ($\times 1.56$ against a token-matched blank page)	that written secret information is proportional to a gradient norm, or that a stored association can be located from one

B LIKELIHOOD METRICS, ESTIMATOR BASELINES AND PROMPT FAMILIES

Likelihood evidence, not mutual information. Verbatim recovery is expected to be rare, so we also report log-likelihood contrasts in bits. Referencing the matched control Θ_0 gives a log-likelihood ratio that the released code calls the *Cross-Modal Information Gain*,

$$\text{CMIG}(s) \;=\; \log_2 p_\Theta(s \mid C) \;-\; \log_2 p_{\Theta_0}(s \mid C) \quad \text{[log-likelihood-ratio bits]}, \tag{2}$$

paired by canary like Eq. (1). Despite its name, it is neither an estimate nor a lower bound of $I(S;\Theta \mid C)$: that interpretation would require a specified joint distribution over secrets and trained models and a validated normalised decoder, which a finite conditional scoring experiment does not supply.

Calibrated cross-model scores. Eq. (2) compares log-likelihoods across two *different* models, so it credits the treated model's *general* improvements (greater fluency on digit strings, on this template, at this length) as memorisation of s. In a 25-step pilot run, raw CMIG was $+7.0$ bits while the Exposure-based CMMG was $+0.05$ bits, and a naive cross-model discrimination score

Table 7: **Reference-model discrimination baselines on identical weights** (Qwen2-VL-2B-Instruct, 3 seeds).

arm	CMMG	scorer A (cross-model)		scorer B (within-model)	
	(bits)	member − non	AUC	member − non	AUC
CAPTION	+0.52	+7.34	0.892	+0.79	0.586
OCR_CAPABILITY	−0.05	−2.09	0.344	+0.07	0.489
INTERLEAVE	−0.48	−0.12	0.554	−1.36	0.370
VQA	+0.17	+0.30	0.518	+0.12	0.503
OCR_SHUFFLED_CONTROL	−0.22	+5.16	0.868	+0.10	0.461
OCR_TARGET	+0.01	+0.28	0.479	−0.10	0.468
TEXT_TOKEN	+0.03	+2.17	0.692	−0.03	0.497
TEXT_TOKEN_WRONGIMG	+2.98	−17.41	0.133	+13.30	0.860
TEXT_TOKEN_NOIMG	+29.65	+30.28	1.000	+61.47	1.000

gave AUC $= 0.92$; both respond to a general fine-tuning shift rather than to this secret. Exposure is less sensitive to this confound because it ranks candidates *within* one model. We give the likelihood-based quantities the same property by differencing against each model's own same-format distractors before differencing across models,

$$\widetilde{\mathrm{CMIG}}(s) = \Big[\log_2 p_\Theta(s|C) - \mathbb{E}_{d\sim\mathcal{R}} \log_2 p_\Theta(d|C)\Big] - \Big[\log_2 p_{\Theta_0}(s|C) - \mathbb{E}_{d\sim\mathcal{R}} \log_2 p_{\Theta_0}(d|C)\Big]. \quad (3)$$

Subtracting each model's mean fluency leaves how much more the trained model singles out *this* string than the control does. On the same run, $\widetilde{\mathrm{CMIG}} = -1.1$ bits and the calibrated membership AUC is 0.42, both near chance for this secret-specific estimand. We therefore report the calibrated quantities as primary and the raw ones only alongside them. This does not invalidate a separate whole-example membership estimand, for which identifiers, templates or other supervised content may remain informative.

Finally, the identity $H(S) = \mathrm{Exposure}(s) + \log_2 \widehat{\mathrm{rank}}(s)$ over the full candidate space reconciles the two scales: the rank advantage that fitted Exposure reports is exactly what is subtracted from the residual guesswork under that fitted ordering. This is bookkeeping under the fitted rank model, not mutual information, and we do not fit a power law in exposure, entropy and visual-token count at these effect sizes.

B.1 Reference-model scorer as a baseline

Table 7 compares two membership scorers on identical trained weights. The cross-model score is $\log P_M(s) - \log P_{\mathrm{base}}(s)$, the text-only construction of Carlini et al. (2021) that document-level membership inference for DocVQA has since moved beyond (Nguyen et al., 2025); the within-model score subtracts the mean log probability of candidate distractors under the same model. The operational label is whether the canary association was present in the treated run or absent from its matched no-canary run, which is not a claim that all other content in an example is distribution-matched for general whole-example membership inference.

The cross-model scorer reports AUC $= 0.892$ on CAPTION, a pixel-only arm whose CMMG contains zero, and 0.868 on OCR_SHUFFLED_CONTROL, built so no pixel→token alignment *can* be learned. The first interval spans chance, so that cell is consistent with an estimator artefact rather than a demonstration of one; the shuffled control, where no alignment exists by construction, is the informative case. **A cross-model score is not evidence of the *secret association* unless a matched control subtracts the fine-tuning shift**. The shuffled arm still contains identifiers, templates, and supervised content, so its score may be valid evidence for a broader whole-example estimand; our experiment neither defines nor rejects that estimand.

B.2 Prompt families

`unwritten.extract.prompts` defines six families, none of which contains the secret:

- **completion**: a near-continuation of the caption the model was trained on, ending at the field label.
- **persona**: a records-assistant framing that asks for the value only.
- **qa**: instruction-style question answering.
- **paraphrase**: the same content in a different surface form (adaptive attacker).
- **format_shift**: a `key:  value` dump, a different document convention entirely.
- **divergence**: a long low-entropy repetition prefix in the style of Nasr et al. (2025), adapted cross-modally.

They serve only as the *likelihood-scoring* instrument for the controlled TinyVLM checks of Appendix H, where each supplies a conditional under which a target's rank among candidates is read, and the reported CMMG is the mean over all six (per-family breakdown in `results/*_attack_rows.csv`). No real-VLM headline number, rank-based or sampled, averages over these six families. The rendering-ladder and fixed-weight readout CMMG values of Section 4 each rest on one stated conditional, usually the binding-prefix `caption_continuation` (Appendix I.4 names the conditional behind every row). The sampled-extraction rate rests on 2 prompt families instead, the training-conditional prompt and one held-out prompt; its rates are properties of those prompts, so the unfavourable rate is a result at a tested prompt, not a bound over prompts.

C Rendering-ladder analyses and replications

C.1 Supporting memorisation results

Table 8 lists every training arm of the dissociation run, Figure 4 shows the three loss-active TEXT_TOKEN conditions, Table 9 and Figure 5 give the readout probe and the ladder against measured legibility, and Figures 6 and 7 give the predictor and pixel-only contrasts referenced in Section 4; the duplication cells are Figure 2d in the main text.

In Table 8, CMMG intervals are 95% Student-t intervals on the 3 per-seed paired means, and p is a seed-level sign-flip test whose attainable floor at 3 seeds is 0.25. The "role" column names the supervision regime. The four *loss-active* arms (the secret is a token target) are differenced against OCR_CAPABILITY__CONTROL, the matched control that shares their filler transcription objective, and the *pixel-only* arms (the secret is never a target) use their own controls. The released code stores the two regimes under the keys `role="upper_bound"` and `role="headline"` in `src/unwritten/arms/arms.py`; neither is a mathematical bound, and a loss-active arm bounds only what this estimator can register when the secret is provably a memorised token target.

The redacted-page probe is interpretable only if it can retrieve from weights with independently demonstrated association-specific recall. We therefore apply it to the illegible loss-active rung and the no-image positive control, as well as the legible loss-active rung (Table 9).

The duplication intervention holds the optimisation budget fixed and uses the realised number of canary presentations rather than the nominal duplication count, which avoids overstating the manipulation when a fixed step budget causes high nominal counts to saturate; Figure 2d plots the cells.

C.2 Step-0 target NLL: a predictor, not an identified mediator

The competition account names a specific diagnostic: the model's NLL on the secret's target positions, given its training image, *before any optimiser step*. It orders the grid, and we report it at the unit that replicates: over the 11 distinct *conditions*, $\rho = +0.75$ with an exact permutation $p = 0.009$. Over the 103 (condition, seed) cells it is $\rho = +0.83$ at a p equal to its own Monte-Carlo floor, but cells within a condition are not independent. Rank-partial correlations point in the predicted direction ($+0.31$ against $+0.04$), quoted as a direction and not as a test. **It is not sufficient.** TEXT_TOKEN_WRONGIMG pairs each canary's caption with *another* canary's page, so it cannot satisfy its target by reading: its step-0 NLL is 2.38 against 2.46 for the no-image arm. This numerical closeness is not an equivalence result, and the corruption residual precludes a single-mediator

Table 8: **Matched-control results by training arm** (Qwen2-VL-2B-Instruct, 3 seeds).

arm	role	Exposure (bits)	CMMG (bits)	95% CI	p	MIA AUC	detect @5% FPR
PRETRAINED_BASE	no fine-tuning	1.43	–	–	–	–	–
CAPTION	pixel-only (secret never in a target)	1.60	+0.52	[−0.86, +1.91]	0.250	0.59	0.10
INTERLEAVE	pixel-only (secret never in a target)	1.09	−0.48	[−2.05, +1.08]	0.250	0.37	0.00
VQA	pixel-only (secret never in a target)	1.17	+0.17	[−0.73, +1.07]	1.000	0.50	0.13
OCR_SHUFFLED_CONTROL	confound control	1.27	−0.22	[−1.40, +0.96]	0.750	0.46	0.03
OCR_CAPABILITY	pixel-only, OCR objective on	1.41	−0.05	[−1.95, +1.86]	1.000	0.49	0.03
OCR_TARGET	loss-active (transcription target, image shown)	1.46	+0.01	[−0.33, +0.34]	1.000	0.47	0.03
TEXT_TOKEN	loss-active (caption target, own page shown)	1.49	+0.03	[−1.22, +1.28]	0.750	0.50	0.00
TEXT_TOKEN_WRONGIMG	loss-active (caption target, mismatched page shown)	4.43	+2.98	[+1.91, +4.05]	0.250	0.86	0.20
TEXT_TOKEN_NOIMG	loss-active (caption target, image withheld)	31.11	+29.65	[+27.99, +31.31]	0.250	1.00	1.00

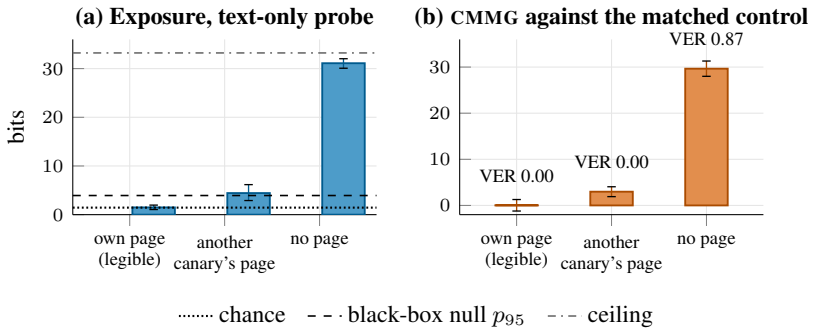


Figure 4: **Own page, another canary's page, or no page during training** (Qwen2-VL-2B-Instruct, 3 seeds): (a) Exposure under one text-only probe; (b) CMMG with verbatim recovery.

claim. Yet its CMMG is +3.04 [+2.20, +3.87] against +28.05 [+26.71, +29.39]; a directional gap also appears on the second backbone (+8.16 against +27.62, vision-inclusive adapter). Section 4.2 takes this up.

C.3 A BACKBONE WITH DIFFERENT READING ABILITY

The 3-seed vision-inclusive ladder (Figure 5c) trains 2,727,936 adapter parameters, 884,736 of them in the SigLIP vision tower, because the adapter's target-module names are matched by suffix and on this backbone also match the vision tower's attention projections. The language-only and vision-inclusive runs at 5 seeds (Section 4.1) form the controlled comparison, and the pooled analysis of Appendix C.5 uses the language-only ladder. SmolVLM-256M-Instruct is a weaker reader than Qwen2-VL-2B-Instruct before adaptation: at 7 pt their character accuracies are 0.23 and 0.73, respectively. On the smaller model, CMMG moves from +0.07 [−1.18, +1.31] at 0.82 legibility to +4.35 [−3.00, +11.70] at 0.06, a paired span of +4.28 [−4.18, +12.75]. Measured legibility, rather than font size, preserves the within-backbone ordering. The OCR-objective contrast is not distinguishable from zero on this backbone (+0.14 [−1.04, +1.32] bits). Table 10 repeats the dissociation run of Table 8 on this backbone over 30 paired canaries, so the null can be read arm by arm. Each pixel-only arm is differenced against its own matched no-canary control, PRETRAINED_BASE is not fine-tuned and has no control, and the OCR_CAPABILITY minus CAPTION contrast is +0.14 [−1.04, +1.32] bits ($p = 0.750$).

C.4 FULL-PARAMETER FINE-TUNING AND OCR-TARGET SUPERVISION

We repeat the two ladder endpoints and no-image control on SmolVLM-256M-Instruct with all 256,484,928 parameters trainable, compared with 2,727,936 under LoRA. The legible and illegible endpoints are +1.76 [+0.03, +3.49] and +21.50 [+17.38, +25.61], for a paired span of +19.74 [+14.93, +24.55]. The direction therefore survives full-parameter training on the smaller backbone. This is an endpoint replication, not a full scaling study.

OCR_TARGET places the field value in a transcription target, matching a common document-pipeline objective. Its ladder moves from +0.10 [−0.33, +0.53] at legibility 1.00 (Figure 5b) to

Table 9: **Readout probe across inference-time image conditions**: the legible rung and its control (top, 3 seeds) and independently positive recall controls (bottom, 3 seeds), on Qwen2-VL-2B-Instruct.

arm	no image	blank page	another canary's page	own page secret redacted	own page secret illegible	own page
TEXT_TOKEN	+1.50	+1.43	+1.21	+1.44	+5.35	+33.22
matched control (TEXT_TOKEN)	+1.38	+1.65	+0.89	+1.50	+4.12	+26.85

trained on	no image	blank page	another canary's page	own page secret redacted	own page secret illegible	own page
TEXT_TOKEN@F4	+15.23	+15.07	+11.21	+21.33	+30.24	+30.24
TEXT_TOKEN_NOIMG	+30.42	+30.14	+18.17	+23.08	+21.13	+33.22

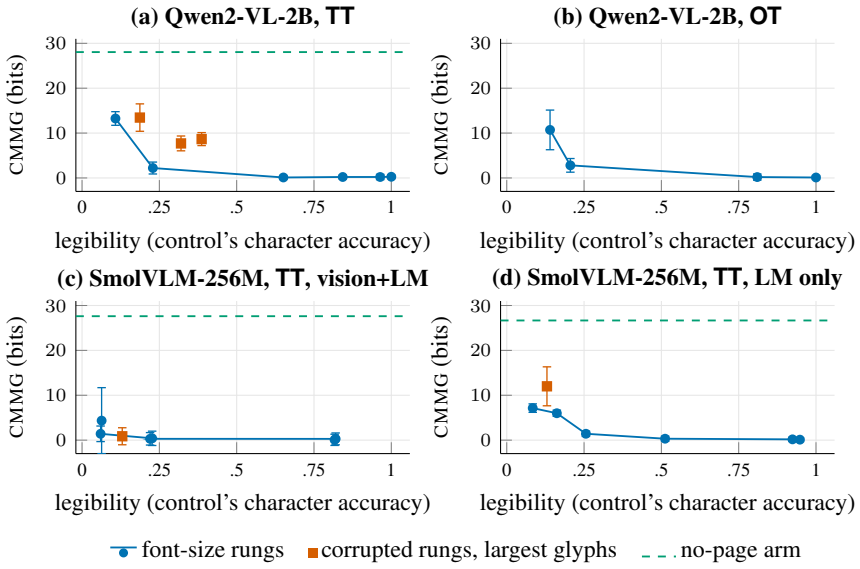


Figure 5: **Memorisation against measured legibility, one panel per ladder run** (fitted tail; TT = TEXT_TOKEN, OT = OCR_TARGET).

+10.71 [+6.29, +15.12] at 0.14, a paired span of +10.60 [+5.96, +15.24]. The result therefore extends from caption-style loss-active supervision to explicit OCR supervision.

C.5 Cross-backbone analysis on measured legibility

The pooled analysis sorts every condition by the matched control's transcription accuracy, because the same font size produces different readability across backbones. These pooled cells are descriptive; within-backbone paired contrasts remain the primary replication unit. The pooled cells (Figure 2c) use 10 seeds and 4,358,144 trainable parameters (language model only) on Qwen2-VL-2B-Instruct, 3 seeds and 7,372,800 (language model only) on Qwen2.5-VL-3B-Instruct, and 5 seeds and 1,843,200 (language model only, run `smolvlm256m_tt_sl_lmonly`) on SmolVLM-256M-Instruct; the adapter scope of each run is resolved below.

Table 11 reports three exploratory monotonicity tests over *condition means*, not three independent significant findings: one p falls below 0.05 and two do not ($p = 0.1667$ and $p = 0.0881$), the rungs within a backbone are not independent replicates, and no multiplicity correction is applied. What replicates across backbones is the *endpoint* direction (Figure 2c), not a common scalar that strictly orders every rung.

Adapter scope. LoRA `target_modules` are matched by name *suffix*, so one list means different things on different architectures. In one run the list `q_proj/k_proj/v_proj/o_proj` also reached SmolVLM's vision tower (Appendix C.3). Two of the three grid backbones recorded the

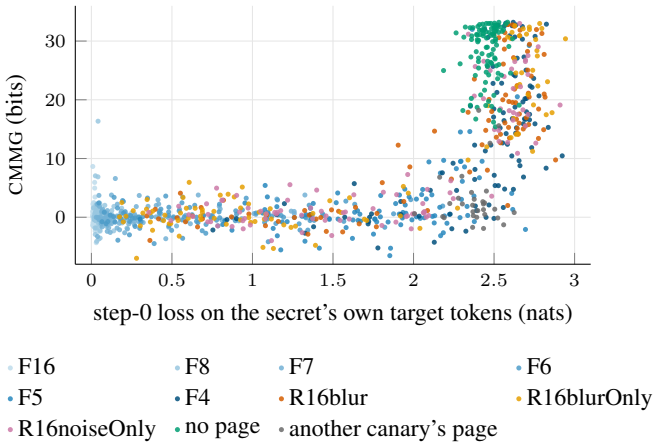


Figure 6: **Step-0 target loss as a memorisation predictor.**

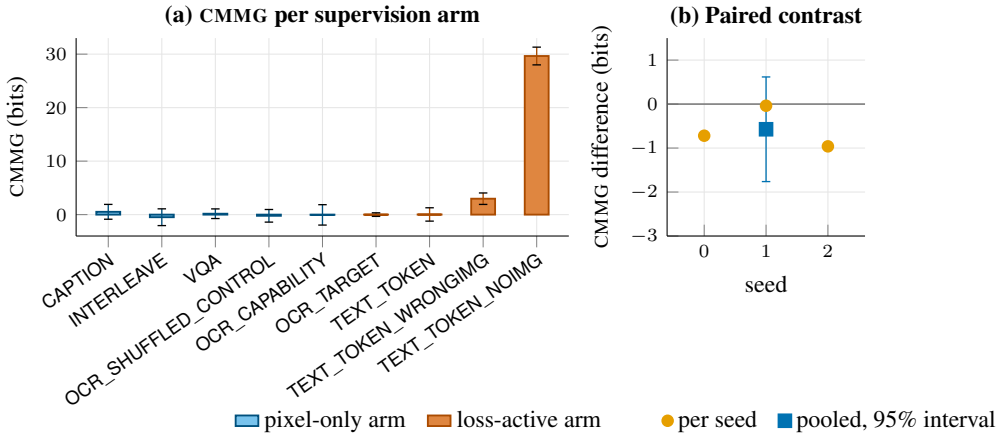


Figure 7: **Supervision arms and the paired pixel-only contrast.**

same list and *no* `lora_scope`, which left their scope unstated. Because scope materially changes the corrupted-rung contrast (Section 4), an unstated scope for a pooled backbone matters.

The scope can be resolved without re-running anything. A released checkpoint ships a parameter index naming every tensor, so whether a suffix *can* match a vision-tower module is a fact about the released weights, checkable offline by any reader; `scripts/audit_adapter_scope.py` reads that index.

The inference is valid in one direction only, and the script enforces the asymmetry. **Zero vision matches proves language-model-only adaptation**: there is nothing for a scope restriction to remove. A positive count proves *only reachability*, because the trainer's `lora_scope` may still restrict where the suffixes are applied; where a run recorded `adapted_modules`, that measurement is used instead and the suffix count is reported beside it.

On this grid the recorded suffixes match 0 vision-tower modules on Qwen2-VL-2B and 0 on Qwen2.5-VL-3B, so both are language model only by proof. SmolVLM's suffixes *do* reach its vision tower (72 modules per attention projection), which is why that run carries an explicit scope restriction; its recorded `adapted_modules` reports zero trainable vision parameters, so the grid's SmolVLM rows are language model only. The audit reads the coupling file named by the grid's run tag, because four SmolVLM coupling runs exist and only the tagged one belongs to the grid.

Table 10: **Pixel-only arms on SmolVLM-256M-Instruct.**

arm	role	Exposure (bits)	CMMG (bits)	95% CI	p	MIA AUC	detect @5% FPR
PRETRAINED_BASE	no fine-tuning	1.72	–	–	–	–	–
CAPTION	pixel-only (secret never in a target)	1.25	−0.14	[−0.82, +0.53]	0.500	0.46	0.03
OCR_CAPABILITY	pixel-only, OCR objective on	1.34	−0.00	[−0.55, +0.55]	1.000	0.50	0.00

Table 11: **Cross-backbone checks**: within-backbone rank correlations with legibility (top) and corrupted rungs against legibility-matched font brackets (bottom).

backbone	rungs	ρ	exact p	p floor
Qwen2-VL-2B	7	−0.61	= 0.167	0.000
Qwen2.5-VL-3B	7	−0.71	= 0.088	0.000
SmolVLM-256M	7	−0.96	= 0.003	0.000

backbone	F16 legib.	F16 CMMG	R16blur legib.	R16blur CMMG	bracketing font rungs	lands in bracket?
Qwen2-VL-2B	1.000	+0.26	0.187	+13.45	F4–F5	yes
Qwen2.5-VL-3B	1.000	−0.08	0.227	+6.52	F5–F6	yes
SmolVLM-256M	0.948	+0.10	0.130	+11.99	F4–F5	yes

Third backbone. The Qwen2.5-VL-3B-Instruct adapter (3 seeds, 7,372,800 LoRA parameters) uses the same target-module names as the primary backbone. On Qwen2.5-VL-3B-Instruct, the largest-glyph corrupted rung produces +6.52 $[+4.04, +8.99]$, separating from the legible endpoint by +6.60 $[+2.79, +10.41]$. This provides a third within-backbone replication of the detectable-memorisation ordering.

C.6 Corruption at matched legibility

The bracket test is nonparametric: it asks whether a corrupted rung's 95% interval intersects the union of the two font-rung intervals surrounding its measured legibility. The single-manipulation rungs fall outside their brackets, showing that transcription accuracy orders the ladder but does not fully summarise corruption.

The bracket test in full. Measured by the trained control's transcription, the two single-manipulation rungs sit at 0.387 and 0.320 character accuracy, *well above* R16blur's 0.187, and are still memorised at +8.66 $[+7.21, +10.11]$ and +7.70 $[+6.04, +9.35]$ bits, where the font rungs bracketing that legibility sit at +0.11 and +2.21. The bracket test, which assumes no functional form, gives a mixed answer: R16blur falls inside its bracket ($[+0.89, +14.77]$), *neither* single-manipulation rung does ($[-0.32, +3.54]$), and the step-0 NLL does not rescue the ordering either (Table 12). This bracket check is exploratory, not a calibrated mediation test or a hypothesis test. Falling outside the union of two neighbouring intervals is a descriptive flag, not a p-value: it was not pre-declared, and although the blur-only and noise-only settings were each calibrated *before* training so the *untrained* backbone's transcription matched R16blur's pre-training reading (0.260 and 0.270 against that target), the calibration ran on the base backbone, not the trained one, and the three conditions' *final* control-model OCR after training (0.387, 0.320, 0.187) is not identical across them, so the bracket compares rungs at similar but not matched trained legibility. Legibility therefore *orders* the ladder but is not a sufficient statistic for it: a field corrupted in place is memorised more than one shrunk to the same readability, on every readability measure we use. This conclusion is weaker than "legibility explains the ladder" and was not pre-declared. Every corrupted rung carries 11.5 px glyphs, and over all 9 rungs $\rho = -0.75$ (exact $p = 0.026$).

Blur and noise decomposition. Blur-only and noise-only controls keep the largest glyphs and are calibrated before training to comparable transcription accuracy. Their paired difference, +0.96 $[-0.48, +2.40]$, contains zero: this contrast is unresolved at 10 seeds. It therefore neither isolates the noise term as the driver of detectable memorisation nor establishes that the two corruptions are equivalent, and we draw no conclusion about corruption type from it. What the two rungs do show jointly is separation from the legible endpoint at the same glyph size, together with variation

Table 12: **Corruption controls**: corrupted rungs against neighbouring font rungs (top) and largest-glyph controls at matched calibration (bottom).

rung	legibility (control, char)	CMMG measured	bracketing font rungs	bracket (union of 95% CIs)	inside bracket?
`R16blur`	0.187	+13.45	F4–F5	$[+0.89, +14.77]$	yes
`R16blurOnly`	0.387	+8.66	F5–F6	$[-0.32, +3.54]$	no
`R16noiseOnly`	0.320	+7.70	F5–F6	$[-0.32, +3.54]$	no

rung	manipulation	glyph (px)	legib. (char)	seeds	CMMG (bits)	95% CI (t, seeds)	step-0 NLL
`R16blur`	blur + noise	11.52	0.187	10	+13.45	$[+10.40, +16.51]$	2.09
`R16blurOnly`	blur only	11.52	0.387	10	+8.66	$[+7.21, +10.11]$	1.51
`R16noiseOnly`	noise only	11.52	0.320	10	+7.70	$[+6.04, +9.35]$	1.79
`F16`	*reference, legible*	11.52	1.000	10	+0.26	$[-0.11, +0.64]$	0.03
`F4`	*reference, small glyphs*	2.88	0.108	10	+13.25	$[+11.73, +14.77]$	2.37

their measured character accuracy does not account for (Appendix C.6). Because the magnitude of that residual depends on the estimator, we report ordering and paired contrasts rather than a unique percentage decomposition.

D FIXED-WEIGHT READOUT DETAILS

Each trained adapter is scored with no image, a blank page, a document never paired with the queried canary, and the unrelated document paired with it during training. The adapter is identical across contexts, and every treated score is differenced against the no-canary control under the same readout. Both arms share one control, trained with the image-conditioned filler mixture (TEXT_TOKEN_WRONGIMG__CONTROL); TEXT_TOKEN_NOIMG is therefore differenced against an image-trained control, which is why its Table 2 and readout-grid values differ. We do not claim that every arm carries a separately trained, fully conditioning-matched control: Appendix I.1 gives the complete map of which control belongs to which contrast and states which between-arm differences cancel a shared control and which do not.

The shared control remains near chance across contexts:

shared no-canary control	no image	blank page	a different page	its training page
mean Exposure (bits)	+1.36	+1.37	+1.49	+1.71

This stability rules out a generic score increase caused merely by adding a document to the prompt. It does not make the control cancel: those values are not equal to each other, and G subtracts their difference rather than assuming it away (Eq. (5) and Table 26).

We report the decomposition on paired seed means separately for every estimator. Let $G = \text{CMMG}(\text{training page}) - \text{CMMG}(\text{no image})$, $S = \text{CMMG}(\text{training page}) - \text{CMMG}(\text{different page})$, and $N = \text{CMMG}(\text{different page}) - \text{CMMG}(\text{no image})$. A positive G is required before discussing its components. Because $S + N = G$ identically, we report no majority shares; instead, for each estimator we report whether the 95% intervals for S and N exclude zero. A component is called stable only when its interval excludes zero on both the fitted-tail and the finite-pool scale. This yields a stable training-page component and an estimator-sensitive document-presence component.

D.1 ESTIMATOR SCALES AND CANDIDATE-DRAW UNCERTAINTY

The finite-pool rank estimator is primary because it makes no tail extrapolation. Its ceiling is 6.60 bits: a rank-1 cell establishes first place among these candidates and nothing about full-space rank. The fitted Exposure and calibrated log-probability margin are shown as sensitivity analyses.

A candidate-draw check counts a win when the truth outscores one sampled same-format distractor, with distribution-free Hoeffding intervals over the stored comparisons, conditional on the fitted models and queried secrets; the seed-level intervals of the main tables separately quantify training-

Table 13: **Fixed-weight readout contrasts across three estimators** (Qwen2-VL-2B-Instruct, 5 seeds). Each contrast is a seed mean over its 95% Student-t interval; $G = S + N$, and C is G's contrast for the arm trained without a page. *At rank 1*: fraction of instances at the rank ceiling in each term.

		primary: finite-pool rank		sensitivity analyses	
	contrast (readout contexts)	bits (ceiling 6.60)	at rank 1 (each term)	fitted Exposure (bits, extrap.)	calibrated $\log p$ (nats)
trained with another canary's page (TEXT_TOKEN_WRONGIMG)					
G	training page − no image	+3.18 [+1.90, +4.46]	1.00 / 0.10	+26.79 [+25.30, +28.29]	+65.17 [+59.02, +71.31]
S	training page − different page	+1.92 [+0.71, +3.13]	1.00 / 0.32	+25.09 [+22.22, +27.96]	+55.46 [+45.15, +65.76]
N	different page − no image	+1.26 [+0.26, +2.26]	0.32 / 0.10	+1.70 [−0.42, +3.83]	+9.71 [+2.94, +16.48]
trained without a page (TEXT_TOKEN_NOIMG)					
C	page of G − no image	−0.47 [−1.25, +0.31]	0.88 / 1.00	−14.68 [−19.12, −10.24]	−28.54 [−39.51, −17.58]
between the two arms					
$G - C$	crossover	+3.65 [+2.71, +4.59]	–	+41.47 [+35.70, +47.24]	+93.71 [+81.81, +105.61]

Table 14: **Token-matched step-0 gradient control** (8 seeds).

training context	tokens	step-0 NLL	$\|\nabla_\theta\|_2$	per position	address cosine
F16	134	0.03	0.388	0.00289	0.813
NOIMG	69	2.45	1.54	0.0222	1.000
BLANK	134	2.45	1.56	0.0116	0.999
WRONGIMG	134	2.35	2.48	0.0184	0.914
F4	134	2.39	2.37	0.0177	0.873

run variation, and neither infers a full-space rank. Conditioning on these fitted models and secrets, the training-page truth wins 100.00% of comparisons [98.04, 100.00], versus 81.38% [79.41, 83.34] without an image; matched controls remain near 50%. Thus candidate-draw uncertainty preserves the ordering, while the 100% cell still prevents a full-space-rank interpretation.

D.2 Token-matched gradient control

The gradient probe (Table 14) tests whether the number of visual tokens dilutes the gradient on the secret's target tokens. Its blank condition contains the same number of visual tokens as a document but no document content, and gradients are measured before any optimiser step. A blank page matches the wrong page's visual-token count without providing document content. Relative to this baseline, the wrong page changes the paired step-0 gradient norm by +0.64 [+0.56, +0.73] in $\log_2$ units ($\times 1.56$, 8 seeds). Thus the weaker text-only retrieval from wrong-page weights is not explained by fewer gradient tokens being written.

D.3 Legibility, near-miss pool and second backbone

The legibility-by-readout grid checks whether context-dependent recovery extends to the rendering ladder. Here, G is the difference between the redacted training page and no-image readouts; the recovering page contains none of the secret's pixels.

Against 96 edit-distance-1–2 near misses, the primary training-page cell remains rank 1 in 1.00 of instances and only its training-page component resolves (only the training-page component excludes zero on this estimator). On the vision-inclusive second backbone, the uniform-pool anomaly is +0.81 [−0.18, +1.80] and its interval contains zero; the near-miss value is +3.36 [+2.78, +3.93],

Table 15: **Legibility-by-readout contrasts** at three rendering rungs.

training rung	recovery G (rank)	at ceiling	G (nats)	G (fitted)
F16 (legible)	−0.26 [−1.32, +0.80]	0.00/0.00	+0.66 [−0.52, +1.85]	−0.18 [−1.47, +1.11]
F4 (illegible)	−0.11 [−1.12, +0.90]	0.77/0.73	+16.92 [+7.76, +26.08]	+5.86 [−0.19, +11.91]
no page at all	−0.22 [−1.20, +0.75]	1.00/1.00	−24.83 [−49.30, −0.37]	−7.65 [−12.16, −3.14]

Table 16: **Persisted-adapter reload verification.**

stored column	n	bitwise equal	max $\lvert\Delta\rvert$
`exposure`	20	18	0.0635
`exposure_within_sample`	20	20	—
`rank_within_sample`	20	20	—
`calibrated_logp`	20	18	0.0322
`true_logprob`	20	18	0.0616
`exposure_trainwrong`	20	17	0.424
`exposure_within_sample_trainwrong`	20	17	0.17
`rank_within_sample_trainwrong`	20	17	2
`calibrated_logp_trainwrong`	20	17	0.5
`true_logprob_trainwrong`	20	17	0.806

with $S = +3.72$ [+2.44, +4.99] and $N = -0.40$ [−1.96, +1.16]. The disagreement across pools and adapter scopes is reported rather than collapsed into one replication label.

D.4 Reload verification and the excluded transformation pass

The near-miss rescore uses persisted adapters. Reload verification is exact for the two primary no-image rank columns (40 of 40 values). For the training-page rank, 17 of 20 values are bitwise equal and the largest change is 2 rank positions; raw log-probability-derived columns differ by as much as 0.806 under nondeterministic MPS forwards. Thus the claim-carrying rank ordering and the training-page recovery conclusion are stable, but we do not describe the whole reload as bitwise reproduction.

The `context_ood` pass rescored persisted TEXT_TOKEN_WRONGIMG and matched-control adapters at five seeds under joint donor-text erasure, an alternate donor template and JPEG-75 compression. It updated no parameters, but an unchanged checkpoint file does not prove that a live load reproduces the training-time scoring state. The reload comparison of Appendix E.2 identifies four loads outside the reload tolerance (treated seeds 2 and 4 and control seeds 2 and 4), with target log-probability discrepancies of up to 120.650 nats; loading, inference-state differences and backend behaviour have not been separated as causes. No complete reference-checked replay of this pass is available, so we report none of its quantitative retention results: its alternate-template and JPEG percentages are not validated measurements of the intended training state. Nor do we retain only the passing seeds, which would change the cohort and the estimand.

All rows of the pass remain in the released results (`results/real_retrieval_qwen2vl2b_context_ood.json`), from which its diagnostic table can be rebuilt for inspection. Its "retained" quantity is the mean across seeds of $G_{s,\text{transform}}/G_{s,\text{exact}}$, with $G_{s,c}$ the paired treated-minus-control gap under c minus that seed's no-image gap. The paired exact-page denominator belongs to this same pass, not to `mech5`, the same-load rescoring or a generation experiment. The same-load rescoring checks candidate ordering in a different execution, and the page-erasure matrix measures a different endpoint, so neither validates this pass. The sampled erasures remain visual and content interventions: the sham matches area only, and the alternate template changes visible non-secret text as well as geometry, so they establish no unique semantic key or general mitigation.

Table 17: **The fixed-weight readout on a ten-times larger uniform pool** (5 seeds).

pool	m	ceiling (bits)	G, bounded bits [95% CI]	rank-1 frac (training page)	control @ no image (bits, chance 1.44)
uniform (main text)	96	6.60	+3.18 [+1.90, +4.46]	1.00	–
uniform (10×)	960	9.91	+5.13 [+1.23, +9.04]	0.82	+1.35

E CANDIDATE-POOL SIZE AND SAME-LOAD RESCORING

The finite-pool scale saturates at rank 1 on the training-page cell. This appendix rescores the same adapters against a ten-times longer pool (Appendix E.1), reconciles the two readouts from the stored rows (Appendix E.2), and reads both pool lengths off a single adapter load (Appendix E.3).

E.1 A LONGER POOL OF 960 INDEPENDENT DISTRACTORS

The bounded ruler is capped at 6.60 bits and the training-page cell sits at rank 1 in every instance, so a contrast built from a saturated cell reports the ceiling rather than the effect. Whether the ordering and the total recovery survive *without* saturation is therefore not a question the 96-distractor pool can answer. We redraw the pool at 960 independent same-format candidates, raising the ceiling to 9.91 bits, and rescore the same persisted adapters under the two readout conditions that define G. No parameter is updated. The intended difference between the two runs is the candidate count alone; Appendix E.2 shows from the stored rows that it was not the only difference and states what the comparison supports.

The interpretation was fixed before any result was seen, with three outcomes (the cell leaves the ceiling, it stays saturated against the larger pool, or the ordering changes and the result is retracted) and one precondition: the matched no-canary control must stay within ± 0.5 bits of chance, which is 1.44 bits for *any* large uniform pool because $\mathbb{E}[\log_2(m+1) - \log_2 R] \to \log_2 e$. In Table 17, "rank-1 frac" is the fraction of treated training-page cells at the pool ceiling; rescoring took 4.93 accelerator-hours over 10 runs, with 0.00 hours of training.

On the stored rows, the training-page cell leaves the pool ceiling (0.82 of instances at rank 1, against 1.00 on the 96-distractor pool); it is concentrated in one training run rather than spread over them: 4 of 5 seeds stay at rank 1 in every instance and seed 3 falls to 0.10, where the 96-distractor pool had put it at rank 1 in every instance. The control sits at +1.35 bits with no image and +1.62 with the training page, against chance 1.44, so the precondition holds. The precondition is weaker than intended: it is satisfied on every seed, including the one whose *treated* checkpoint scores its own target 132.5 nats away from the 96-distractor run (Appendix E.2). A control near chance shows that the pool did not move the *scorer*; it cannot show that the treated weights being scored are the ones the shorter pool measured.

Two readings of the longer pool. Per seed, the training-page cell's rank-1 fraction is 1.00, 1.00, 1.00, 0.10, 1.00, against 1.00, 1.00, 1.00, 1.00, 1.00 on the 96-distractor pool, and G is +5.74, +5.97, +7.48, −0.35, +6.83. For the 4 runs that stay at rank 1 in every instance, the shorter pool was reporting a censored value: their recovery rises from the +3.18 [+1.90, +4.46] pooled value to between +5.74 and +7.48 bits once the ceiling is far enough away to express it. For the remaining run the two pools disagree in the opposite direction: seed 3's secret beat 96 uniform distractors in every instance, but its median rank becomes 369 against 960 distractors. Appendix E.2 shows that this seed's rows were not scored from the same model state in the two runs, so its reversal is not evidence about the candidate count, and we do not read it as a ceiling artefact either.

Both readings enter the pooled number and they pull against each other. G rises to +5.13 [+1.23, +9.04], an interval that still excludes zero and is larger than the 96-candidate +3.18 [+1.90, +4.46]; but the seed-level sign-flip test weakens from $p = 0.062$ (floor 0.062) (the attainable floor at 5 seeds, with every seed positive) to $p = 0.125$ (floor 0.062), because one seed now reverses. We therefore do not promote the longer ruler's point estimate over the 96-distractor one, and we do not report a seed-excluded number in its place: the pooled value above is over all 5 seeds exactly as measured. What this run establishes is narrower than "the shorter pool was a lower bound". On the seeds whose scored state matches the 96-distractor run, a rank-1 cell on a 96-candidate pool

survives a ten-times longer one (30 of 30 instances), and the bounded contrast those seeds report is censored rather than complete. On the remaining seed the two runs are not comparable, which is a measurement finding about reloaded checkpoints rather than about pools. Both argue for reporting the ceiling fraction beside every bounded contrast, as we do, and for reading both rulers off a single load, which Appendix E.3 reports.

E.2 Reconciling the 96- and 960-distractor readouts

Appendix E.1 compares two readouts taken at different candidate-pool lengths. That comparison supports a conclusion about the pool only if nothing else differed between them; this subsection shows, from the stored rows and without model inference, that something else did. Every number below is computed from the stored rows by `scripts/reconcile_pool_ranks.py` (Table 18).

The pool-invariant column. Each row stores the target's *own* sequence log-probability beside its rank. The candidate pool cannot change that column: it enters the score only through the count of candidates that beat the target. Two runs that differ only in pool length must therefore agree on it bitwise, cell by cell. It is the control on the claim that only the ruler changed, and it needs no forward pass to read.

Nested pools. A same-load comparison needs the original 96 distractors embedded in the 960-distractor set. They are, by construction: the candidate sampler is seeded from the training seed, the canary identifier and the prompt style, never from the count, so a 960-distractor draw extends the 96-distractor draw. We check this over all 50 distinct (seed, canary) candidate draws the two runs use (the draw does not depend on the arm), redrawing each pool from the same generator the scorer called: the shorter pool is a prefix of the longer one in every case. The count identity that follows (a nested pool can only *add* candidates that beat the target, so the longer-pool rank can never be smaller) holds on all 146 cells whose pool-invariant column did not move (0 violations).

A stable scale on unsaturated strata. On the strata that are not at the pool ceiling (the matched no-canary control), the implied fraction of the pool beating the median target is stable across the two lengths: 0.427 at 96 against 0.455 at 960 with no image, and 0.516 against 0.500 with the training page. Lengthening the pool is therefore not, by itself, evidence of a changed scorer, and a calculation that starts from a winning-distractor fraction estimated on one seed is not a calculation about the pool.

Different adapter loads. The 96-distractor rows were scored from the in-process trained model; the 960-distractor rows were scored from a persisted adapter. Pool-invariance fails, in 4 of the 4 strata: on the treated arm read out with the training page (the stratum that carries the result), 18 of 50 cells disagree between the runs, by up to 132.5 nats. Reloading a persisted adapter is not bitwise reproducible on this device, so we measure how far it moves rather than assume a bound: the verification run of Table 16 rescored one seed on the *same* pool, so every difference in it is reload behaviour, and its largest movement on this column is 0.806 nats over 40 cells (seed 2). Seeds 0, 1, 2 reproduce inside that measured tolerance; seeds 3, 4 do not, and seed 3 exceeds it by a factor of 164.

The reversal follows the moved column, not the pool. Of the 50 treated training-page cells at rank 1 on the shorter pool, 9 leave rank 1 on the longer one, all of them in seed 3. Every one of those 9 cells has a moved pool-invariant column, with a smallest movement of 75.4 nats; every one of the 41 cells that stays at rank 1 moved by at most 2.84 nats. The two sets do not overlap. In the same scoring pass that seed's matched no-canary control reproduces its own log-probabilities to 0.0009 nats, so the movement is not a property of the pool or of that pass as a whole.

Reload excursions across passes. We compare every stored pass that reloaded these adapters with the in-process rows, on the pool-invariant column and in the readout contexts the tolerance was measured in, counting a load as one (arm, seed) adapter loaded once in a pass; the verification pass defines the tolerance, so it cannot exceed it. Seed 3's treated adapter reproduces those rows within the tolerance in 3 of the 3 other passes that load it; 2 of the 5 passes contain a load outside it (a

matched control's load in 1 of them), by up to 132.5 nats, and the loads that leave it differ from pass to pass. On these rows a reload excursion belongs to an individual load on this device, not to a checkpoint. A control that stays near chance, the pre-specified precondition of Appendix E.1, cannot detect such an excursion, and a replay of a persisted adapter is a check rather than an identity. In Table 18, "beaten frac." is $(\text{median rank} - 1)/m$, the only quantity comparable across pool lengths; "cells moved" counts cells whose *target's own* log-probability differs between the runs, a quantity the pool cannot change; and "max $|\Delta|$" is the largest such difference, against the measured reload tolerance of 0.806 nats on this column (Table 16).

Table 18: **The two readouts reconciled from the stored rows**, by arm and readout context.

arm	readout context	n	rank-1 96→960	beaten frac. at 96	beaten frac. at 960	cells moved	max $\|\Delta\|$ (nats)
TEXT_TOKEN_WRONGIMG	none	50	5/3	0.104	0.159	17	110.286
TEXT_TOKEN_WRONGIMG	trainwrong	50	50/41	0.000	0.000	18	132.528
TEXT_TOKEN_WRONGIMG__CONTROL	none	50	0/0	0.427	0.455	6	0.062
TEXT_TOKEN_WRONGIMG__CONTROL	trainwrong	50	1/0	0.516	0.500	13	0.806

What the stored rows support. On the seeds whose pool-invariant column reproduces within the measured tolerance, the training-page cell stays at the ceiling of the ten-times longer pool in 30 of 30 instances: for those runs the shorter pool was reporting a value the longer one does not overturn. Nothing follows about seed 3. Its two rows were produced under different model states, and these rows cannot say which state is the trained one, so we neither treat its reversal as a measurement of the effect at 960 distractors nor drop it from the pooled estimate of Appendix E.1, which is reported over all seeds exactly as measured. Appendix E.3 reads both rulers off one load per adapter, the comparison these two runs could not make.

E.3 Both rulers from one load

Appendix E.2 could not determine what the reversing seed's adapter does against 960 distractors, because its two stored rows came from two different loads. This run removes that difference by construction. Each of the 10 persisted adapters (treated arm and matched control, 5 seeds) is loaded once and scored once against the 960-distractor pool under both readout contexts, and the 96-distractor ranks are read off the *same* forward passes, from the prefix of that pool that Appendix E.2 showed to be the original 96-distractor draw. The target's own log-probability is therefore one number shared by both rulers: over the 200 (cell, readout) pairs it differs between them on 0, and the nested count identity fails on 0. The analysis (`scripts/analyze_nested_pool.py`) refuses a results file that does not cover every (arm, seed) load of the original comparison with both rulers. The pass took 2.16 accelerator-hours of scoring and no training. Pool sizes here count distractors: including the queried secret, the two rankings contain 97 and 961 candidates, respectively. In Table 19, "rank-1 at 96 not at 960" counts cells whose rank changes between the two rulers *within* one load, so it measures the candidate count and nothing else; the last two columns compare each load's target log-probability with the in-process 96-distractor rows, against the measured reload tolerance of 0.806 nats (Table 16).

Table 19: **Both rulers read off one load per adapter**, by arm and readout context.

arm	readout context	n	rank-1 at 96	rank-1 at 960	rank-1 at 96 not at 960	beaten frac. at 96	beaten frac. at 960	seeds within tolerance	max $\|\Delta\|$ vs in-process (nats)
TEXT_TOKEN_WRONGIMG	none	50	5	3	2	0.104	0.112	5/5	0.000
TEXT_TOKEN_WRONGIMG	trainwrong	50	50	50	0	0.000	0.000	5/5	0.000
TEXT_TOKEN_WRONGIMG__CONTROL	none	50	0	0	0	0.427	0.455	5/5	0.062
TEXT_TOKEN_WRONGIMG__CONTROL	trainwrong	50	1	0	1	0.516	0.500	5/5	0.806

What the candidate count does within one load. For the treated arm read out with its training page, 50 of 50 cells are at rank 1 against 96 distractors and 50 against 960; 0 are at rank 1 on the shorter ruler and not on the longer one (no seed). Per seed, the training-page rank-1 fraction against 960 distractors is 1.00, 1.00, 1.00, 1.00, 1.00. From the same rows, G is +3.19 [+1.90, +4.47] bits at 96 distractors ($p = 0.062$ (floor 0.062)) and +6.11 [+4.99, +7.22] at 960 ($p = 0.062$ (floor 0.062)), with per-seed values +5.74, +5.97, +7.48, +5.03, +6.32.

Whether each load is the trained state. One load removes drift *between* the rulers; it does not by itself make the loaded state the one training produced, and a load of these same adapters has left the measured tolerance by up to 132.5 nats (Appendix E.2). Against the in-process rows, the treated arm's target scores reproduce within that tolerance on every seed, and the matched control's on every seed.

The reversing seed. All 9 cells that left rank 1 in the reloaded 960-distractor run (seed 3) are at rank 1 against 960 distractors here, in a load whose target scores reproduce the in-process run within the measured tolerance (largest difference 0.00 nats against 0.806): within one load that reproduces the trained scores, the longer pool does not take rank 1 from these cells, so that reversal belonged to the earlier reload rather than to the candidate count.

F SAMPLED GENERATION UNDER EXPLICIT ACCESS

Both experiments sample continuations from persisted adapters without further training: the generation budget varies the query budget and the prompt, and the page-erasure matrix varies the page and the caller's prompt access.

F.1 GENERATION BUDGET

Likelihood ranking gives a model-owner audit quantity, not a public-API attack rate. We therefore reload the same adapters without training and sample 20 continuations per canary (temperature 0.8, top-p 0.95). A success means that the exact normalised secret has appeared by query budget k; the page-assisted condition supplies the exact unrelated training page, which contains none of that secret's pixels.

Two prompt families are reported, because they are not interchangeable and one of them alone would misstate the rate in a known direction. The `recall` family is a held-out user-turn question, so it queries a conditional the fine-tune never saw. It is *one* such question, fixed in advance and not searched over, so what it realises is a measurement at that prompt and bounds nothing about the formats an adaptive caller could try; the `caption_continuation` family reproduces the training conditional exactly, with the caption instruction in the user turn and the memorised caption up to the field label in the assistant turn, and is the strongest generation-only readout available to a caller who knows the caption format. It is also the family behind the VER numbers, which is why the no-image arm's continuation curve and its VER agree.

One decoder property is checked before the campaign rather than assumed. The backbone ships a `generation_config` with `top_k` $= 1$; a caller who passes `do_sample`, a temperature and a top-p but leaves `top_k` unset inherits it and decodes greedily, which would make every budget curve flat by construction and hit@1 equal to hit@20 whatever the weights hold. We set `top_k` $= 0$ and, on an open-ended prompt where any working sampler must vary, require more than one distinct continuation before any adapter is scored: 8 distinct of 8 draws, and 8.1 distinct of 20 averaged over the campaign's own cells. The check is deliberately not run on a canary prompt, where a near-deterministic draw is a legitimate property of an overfit arm and so cannot distinguish peaked weights from a dead decoder.

Interval conventions. The inferential unit in Table 20 is the training seed; canaries are never treated as independent training runs. Where seed means vary, brackets are 95% Student-t intervals over those means, intersected with $[0, 100]$. Where every seed sits exactly at 0 or 100, the between-seed variance is zero; those cells are marked † and carry exact two-sided Clopper–Pearson limits on the five seed outcomes, $[0, 52.2]$ or $[47.8, 100]$. A seed counts as a boundary success only when all 10 of its canaries share that outcome, so these limits concern the probability that *every* canary of a new training run lands at the observed extreme. They are not per-canary intervals, nor do they transfer to another backbone, schedule, prompt family or caller. Appendix F.2 instead reports a per-seed mean-rate Hoeffding interval for the same favourable access condition, and Section 4.2 uses only that interval; the two conventions answer different questions about different experiments and should not be combined. The training-conditional block additionally assumes that the caller knows the caption format and, in the "exact training page" rows, which document was paired with the queried association during fine-tuning.

Table 20: **Generation-only exact-secret recovery** (percent, 5 seeds) under a held-out question and under the training conditional.

Held-out user-turn question

weights	readout	k=1	k=5	k=20
TT_NOIMG	no image	0.0† [0.0, 52.2]	4.0 [0.0, 10.8]	6.0 [0.0, 17.1]
TT_WRONGIMG	no image	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]
TT_WRONGIMG	exact training page	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]
matched control	no image	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]
matched control	exact training page	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]

Training conditional

weights	readout	k=1	k=5	k=20
TT_NOIMG	no image	84.0 [69.8, 98.2]	90.0 [77.6, 100.0]	92.0 [81.6, 100.0]
TT_WRONGIMG	no image	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]
TT_WRONGIMG	exact training page	100.0† [47.8, 100.0]	100.0† [47.8, 100.0]	100.0† [47.8, 100.0]
matched control	no image	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]
matched control	exact training page	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]	0.0† [0.0, 52.2]

Table 21: **Secret emitted when shown the same donor page** (first sample, 50 canary–seed cells per row).

weights, shown the same donor page	queried secret	donor's own secret	neither
TT_WRONGIMG (treated)	100%	0%	0%
matched no-canary control	0%	92%	8%

The curve is reported separately from candidate ranking: a likelihood effect can be useful for a model-owner audit even when a generation-only user cannot realise it at these budgets, and page-assisted success additionally assumes knowledge of the training-page association. The two families make the point concrete. Under the held-out question the no-image arm reaches 6.0% [0.0, 17.1] at $k = 20$, while every TT_WRONGIMG and matched-control cell is at 0.0%; under the training conditional the same weights reach 92.0% [81.6, 100.0]. Prompt-format knowledge, not the weights, is what separates those two numbers, and neither is the rank separation the audit scale reports.

A 100.0% cell (realised only for a caller holding the target identifier, the caption format *and* the identity of the page paired with that association in training, with assistant-prefill capability) and a 0.0% cell could also arise from a harness fault, so we read both back against the corpus before treating them as rates (Table 21). Two observations exclude a fault. Without the page, the same treated weights under the same prompt emit a *diverse* spread of wrong digit strings (16.5 distinct of 20 draws on average, against 1.3 with the page), so the zero is a genuine miss by a working sampler and not a stuck decoder. The control, given the identical image, is equally deterministic but names the donor's own secret, which is on the page; the treated model instead emits a value that appears on no donor page. Reading therefore accounts for the control, but not for the treated model.

F.2 Selective page erasure and prompt access

This inference-only experiment asks two questions about the fixed-weight readout: which groups of donor-page text support sampled recall, and whether knowing a caption prefix is interchangeable with supplying it as an assistant prefill. The complete matrix, transforms, seeds, decoder and endpoints were specified after the readout grid of Section 4.2 and before any of these outputs were generated. The specification is internal, not an external preregistration; no prompt or condition was selected on these outputs, and no adapter was retrained.

Table 22 gives the first-sample counts plotted in Figure 3c. Denominators are canaries, not independent training runs; rows 1–5 use an assistant prefill and the last does not; the full matched-control matrix is Table 23.

An exact prompt. For seed 0 the user turn is "Describe the document." with the donor image; the caller-supplied assistant prefix is "Scanned tax form from Northgate Clinical Group. Account

Table 22: **Selective readout and access checks.** Queried-secret hits at one sampled query in the treated models.

readout	Qwen (5 seeds)	Smol (3 seeds)
Exact donor page	48/50	27/30
Donor secret removed	0/50	0/30
Non-secret text removed	6/50	5/30
Both text groups removed	0/50	1/30
Alternate donor template	3/50	0/30
User asks to complete known caption	6/50	0/30

holder: Linurien Ospfield. The Taxpayer ID field is completed. Taxpayer ID:"; the generated suffix is "6763 3739 06." The secret begins only after that final colon; the supplied prefix is secret-free.

Checkpoints and replication units. We use TEXT_TOKEN_WRONGIMG and its own no-canary control from `qwen2vl2b_mech5` (Qwen2-VL-2B, seeds 0–4) and `smolvlm256m_mech6` (SmolVLM-256M, seeds 0–2). Each seed has ten canaries, 128 fillers and 24 exposures, as specified in each training file. These are 50 and 30 canary associations, respectively, but only five and three independent training-seed replicates; we never pool them into eight seeds of one population. The Qwen adapter targets the language stack and the Smol adapter includes the vision stack (Appendix C.5); consequently this is replication across two trained systems, not an architecture-only contrast. Each loaded adapter is held fixed across all twelve readouts. Every paired control receives exactly the same inference input as its treated counterpart.

Eight image conditions at a fixed assistant prefix. The caption-continuation probe supplies the secret-free target caption through its final field label as the beginning of the assistant response. It is paired with (i) no image, (ii) the exact donor page seen during training, (iii) that page with its non-secret text group erased, (iv) that page with its own secret erased, (v) the union of those two erasures, (vi) an area-matched background sham, (vii) the donor content in an alternate template, and (viii) a JPEG-75 donor page. None contains the *queried* secret.

The masks come from the renderer's text roles, with a two-pixel margin. The non-secret group contains the title and ordinary value runs, including identifiers and other non-secret field values; it is *not* a pure identifier intervention. The donor-secret mask takes precedence at overlaps, so non-secret erasure preserves every pixel in that mask and vice versa. Erased pixels are reset to the renderer's background grey; all pixels outside the selected mask are checked byte-for-byte. The joint mask is their union. The sham selects the same number of background locations outside the joint mask and darkens them by eight grey levels, preserving every original ink pixel. Thus it controls *mask area*, not perturbation energy, location, or visual salience. It cannot exclude a stronger local-corruption explanation. Erasure changes semantic content and visual features together; differences between these conditions do not by themselves identify semantic mediation or prove that a text group is necessary for all retrieval strategies. The alternate-template condition keeps the underlying donor record and its secret, but changes the document class and may expose different non-secret fields or labels. It is not a pure layout change at identical visible text. JPEG alone keeps that visible content while changing its encoding fidelity. Figure 8 shows the input transforms for the first canary by index, chosen without reference to model outputs; colour is an annotation marking the regions each condition edits, not part of the input. The queried secret is absent from every panel, the donor's distinct secret remains in the non-secret-text erasure (b), and the no-page condition (h) supplies no image rather than a blank page.

Four user-turn-only conditions. Each receives the exact donor image. Three use the existing `direct`, `completion`, and `recall` user questions in `src/unwritten/realtrain/attack.py`. The fourth supplies the same secret-free caption prefix as above, but *inside the user message*: "Complete the following document caption with the field value. Return only the missing value. Caption:" followed by that prefix. No assistant text is prefilled in these four conditions. Comparing the two prefix placements tests the actual interface capability, rather than assuming that format knowledge implies prefill access. The other user prompts and the placement change also alter the model's conditional; this is a functional access test, not an isolated linguistic mediator or an adaptive prompt-search benchmark.

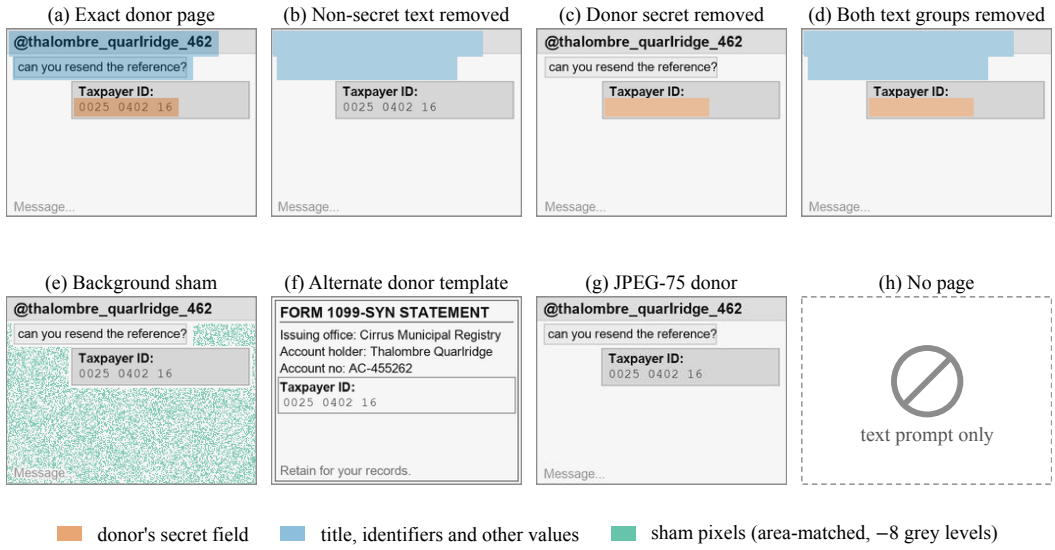


Figure 8: **Donor-page interventions** for the first queried canary of seed 0, before processor resizing; colour marks the edited regions.

Decoder and accounting. Each cell receives five samples with temperature 0.8, nucleus probability 0.95, `top_k=0`, and at most 24 newly generated tokens, in a batch of five. Random seeds are deterministically derived from the training seed, arm, case, and canary identifier. A decoder-liveness check precedes each model family's first query. Hit@k records whether the normalised queried secret occurs in the first k sampled continuations; donor-hit@1 instead scores the donor's distinct secret. The full matrix comprises 1920 query cells and 9600 sampled continuations, with 0.88 hours of logged adapter loading and inference. In Table 23, each Qwen cell has 50 canaries nested in five training seeds and each Smol cell 30 nested in three; the first eight rows use an assistant prefill and the last four only a user turn, and the controls are the matched no-canary weights, not the untrained base model.

The favourable condition requires the identifier, the paired page, the caption format and assistant prefill. Over five Qwen training seeds it gives 48/50 first-sample recoveries and 50/50 within five, against 100.0% at the first sample in the generation-budget experiment (Appendix F.1). Both experiments use temperature 0.8, top-p 0.95, top-k zero and 24 new tokens; this one uses different random seeds, batch size five instead of twenty and a fresh adapter load, so it is not a bitwise replay. Its two first-sample misses are kept, and Section 4.2 reports this estimate rather than the perfect generation-budget cell.

Table 23: **Complete generation matrix** (queried-secret hit@1 / hit@5, percent).

condition	Qwen treated	Qwen control	Smol treated	Smol control
No page	0 / 0	0 / 0	0 / 0	0 / 0
Exact donor page	96 / 100	0 / 0	90 / 97	0 / 0
Non-secret text removed	12 / 20	0 / 0	17 / 27	0 / 0
Donor secret removed	0 / 0	0 / 0	0 / 0	0 / 0
Both text groups removed	0 / 0	0 / 0	3 / 3	0 / 0
Background sham (matched area)	82 / 88	0 / 0	83 / 90	0 / 0
Alternate donor template	6 / 12	0 / 0	0 / 0	0 / 0
JPEG-75 donor	66 / 80	0 / 0	80 / 97	0 / 0
User question: direct	2 / 6	0 / 0	0 / 10	0 / 0
User question: completion	2 / 8	0 / 0	3 / 10	0 / 0
User question: recall	0 / 0	0 / 0	0 / 0	0 / 0
User asks to complete known caption	12 / 22	0 / 0	0 / 7	0 / 0

Seed-level uncertainty. For each condition and arm let $Z_s \in [0, 1]$ be the success fraction among the ten canaries of training seed s. We report every observed Z_s and the unweighted mean over seeds. A distribution-free 95% interval for the *expected per-seed canary success fraction* is

$$[\bar{Z} - h_n, \bar{Z} + h_n] \cap [0, 1], \qquad h_n = \sqrt{\log(40)/(2n)}.$$

This Hoeffding interval assumes independent, comparably generated training seeds and allows arbitrary within-seed dependence. It targets the same mean at zero, intermediate, and perfect observed success; it is not an interval for "a seed with any failure." With $n = 5$ or 3 it is intentionally wide and does not establish near-certain extraction or absence of extraction in a new run. Table 24 reports hit@1 seed fractions and intervals; with three seeds the Smol intervals are particularly wide, and its adapter scope differs from Qwen's. The released analysis script reconstructs the hit@5 and donor-hit@1 seed fractions from the raw rows.

Paired contrasts and their resolution limit. For each comparator we subtract its treated-minus-control hit@1 gap from the exact-page gap separately in every seed, on the same loaded weights and with each comparator's own matched control. Table 25 gives the seed differences in percentage points (positive values favour the exact page), a Student-t interval when their variance is nonzero, and the exact two-sided seed sign-flip p-value, which is descriptive and unadjusted for multiple comparisons. The t interval is a normal-seed working-model sensitivity analysis; its coverage is not guaranteed for these few bounded seed fractions. The latter test assumes sign symmetry under the paired null. Its smallest possible value is 0.0625 with five seeds or 0.25 with three: none of these comparisons can be a confirmatory 5% sign-flip rejection, before or after multiplicity correction. Constant observed differences are labelled "constant seeds," not assigned a zero-width confidence interval. These checks qualify the descriptive changes and cannot establish equivalence from a nonsignificant result.

Execution and validation. The runner validates all masks and secret-free inputs before inference, checks checkpoint hashes, and atomically saves each completed cell with its raw continuations, prompt hash, target–donor binding, decoder configuration and transformation checks; incompatible configuration, code or adapter hashes are refused on resumption. The table builder performs no inference: it requires both complete matrices, independently rebuilds the canary–donor bindings, recomputes hits from raw generations, and rejects missing, duplicated or unplanned cells.

Reference-checked execution. Decoder liveness checks sampling, not agreement with the training-time scores, so both matrices were executed under a reference check. Before and after each loaded adapter's matrix, the runner scores the first two canaries under no image and the exact donor page with the original candidate count, batch size and prefix. It stores the expected and observed four-score vectors, the reference-file hash, a load identifier and the tolerance, and refuses completion if the pre–reference, post–reference or pre–post differences exceed the threshold; an interrupted or failed load must replay its entire matrix, so its rows cannot be combined into a completed pass. The 0.8059616089-nat threshold is the measured Qwen reload maximum; on Smol it is a pre-specified diagnostic threshold, not a Smol-calibrated guarantee. All 1200 Qwen cells over ten loaded adapters and 720 Smol cells over six passed both the pre- and the post-generation comparison, with a maximum absolute deviation of 0.000000 nats across all recorded vectors. Because each cell's decode is seeded, a reproduced model state yields identical draws by construction, so the check establishes that each load reproduced the training-time scoring state on the checked vector; it is not independent corroboration of the hit rates. Its coverage is narrow: it scores two of each seed's ten canaries, two of the twelve readout conditions, and teacher-forced target log-probabilities rather than the temperature-0.8 nucleus-sampling decode behind the hit rates, so a regression confined to the other canaries, the other conditions or the sampler would not trip it.

Selection and audit access. Specifying this matrix in advance does not undo the data-dependent seed expansion or the choice of existing adapters after earlier observations; the intervals are conditional on the stated comparable-seed design, not population-valid confirmation over adaptively selected experimental programmes. Training logs and the deterministic `wrong_page_map` let a model owner reconstruct the donor binding, using the 10-canary, 128-filler, 24-exposure corpus recorded in each source training file. No external caller is assumed to possess training logs, and the cost of discovering the page or prompt is outside the one- and five-query budgets.

Table 24: **Seed-level hit@1** for Qwen (top, seeds 0–4) and Smol (bottom, seeds 0–2).

condition	weights	seed hit@1 (percent)	Hoeffding 95%
No page	treated	0, 0, 0, 0, 0	[0.0, 60.7]
No page	control	0, 0, 0, 0, 0	[0.0, 60.7]
Exact donor page	treated	100, 100, 100, 100, 80	[35.3, 100.0]
Exact donor page	control	0, 0, 0, 0, 0	[0.0, 60.7]
Non-secret text removed	treated	30, 10, 20, 0, 0	[0.0, 72.7]
Non-secret text removed	control	0, 0, 0, 0, 0	[0.0, 60.7]
Donor secret removed	treated	0, 0, 0, 0, 0	[0.0, 60.7]
Donor secret removed	control	0, 0, 0, 0, 0	[0.0, 60.7]
Both text groups removed	treated	0, 0, 0, 0, 0	[0.0, 60.7]
Both text groups removed	control	0, 0, 0, 0, 0	[0.0, 60.7]
Background sham (matched area)	treated	100, 100, 100, 70, 40	[21.3, 100.0]
Background sham (matched area)	control	0, 0, 0, 0, 0	[0.0, 60.7]
Alternate donor template	treated	10, 10, 10, 0, 0	[0.0, 66.7]
Alternate donor template	control	0, 0, 0, 0, 0	[0.0, 60.7]
JPEG-75 donor	treated	70, 80, 80, 80, 20	[5.3, 100.0]
JPEG-75 donor	control	0, 0, 0, 0, 0	[0.0, 60.7]
User question: direct	treated	0, 0, 0, 10, 0	[0.0, 62.7]
User question: direct	control	0, 0, 0, 0, 0	[0.0, 60.7]
User question: completion	treated	0, 0, 0, 10, 0	[0.0, 62.7]
User question: completion	control	0, 0, 0, 0, 0	[0.0, 60.7]
User question: recall	treated	0, 0, 0, 0, 0	[0.0, 60.7]
User question: recall	control	0, 0, 0, 0, 0	[0.0, 60.7]
User asks to complete known caption	treated	0, 0, 30, 20, 10	[0.0, 72.7]
User asks to complete known caption	control	0, 0, 0, 0, 0	[0.0, 60.7]

condition	weights	seed hit@1 (percent)	Hoeffding 95%
No page	treated	0, 0, 0	[0.0, 78.4]
No page	control	0, 0, 0	[0.0, 78.4]
Exact donor page	treated	100, 70, 100	[11.6, 100.0]
Exact donor page	control	0, 0, 0	[0.0, 78.4]
Non-secret text removed	treated	10, 30, 10	[0.0, 95.1]
Non-secret text removed	control	0, 0, 0	[0.0, 78.4]
Donor secret removed	treated	0, 0, 0	[0.0, 78.4]
Donor secret removed	control	0, 0, 0	[0.0, 78.4]
Both text groups removed	treated	10, 0, 0	[0.0, 81.7]
Both text groups removed	control	0, 0, 0	[0.0, 78.4]
Background sham (matched area)	treated	70, 80, 100	[4.9, 100.0]
Background sham (matched area)	control	0, 0, 0	[0.0, 78.4]
Alternate donor template	treated	0, 0, 0	[0.0, 78.4]
Alternate donor template	control	0, 0, 0	[0.0, 78.4]
JPEG-75 donor	treated	90, 60, 90	[1.6, 100.0]
JPEG-75 donor	control	0, 0, 0	[0.0, 78.4]
User question: direct	treated	0, 0, 0	[0.0, 78.4]
User question: direct	control	0, 0, 0	[0.0, 78.4]
User question: completion	treated	0, 0, 10	[0.0, 81.7]
User question: completion	control	0, 0, 0	[0.0, 78.4]
User question: recall	treated	0, 0, 0	[0.0, 78.4]
User question: recall	control	0, 0, 0	[0.0, 78.4]
User asks to complete known caption	treated	0, 0, 0	[0.0, 78.4]
User asks to complete known caption	control	0, 0, 0	[0.0, 78.4]

Table 25: **Paired changes in queried-secret hit@1** (percentage points).

model	exact-page gap minus	seed differences (pp)	t 95%	exact p
Qwen	No page	100, 100, 100, 100, 80	[84.9, 107.1]	0.062
Smol	No page	100, 70, 100	[47.0, 133.0]	0.250
Qwen	Non-secret text removed	70, 90, 80, 100, 80	[69.8, 98.2]	0.062
Smol	Non-secret text removed	90, 40, 90	[1.6, 145.1]	0.250
Qwen	Donor secret removed	100, 100, 100, 100, 80	[84.9, 107.1]	0.062
Smol	Donor secret removed	100, 70, 100	[47.0, 133.0]	0.250
Qwen	Both text groups removed	100, 100, 100, 100, 80	[84.9, 107.1]	0.062
Smol	Both text groups removed	90, 70, 100	[48.7, 124.6]	0.250
Qwen	Background sham (matched area)	0, 0, 0, 30, 40	[−10.2, 38.2]	0.500
Smol	Background sham (matched area)	30, −10, 0	[−45.0, 58.4]	1.000
Qwen	Alternate donor template	90, 90, 90, 100, 80	[81.2, 98.8]	0.062
Smol	Alternate donor template	100, 70, 100	[47.0, 133.0]	0.250
Qwen	JPEG-75 donor	30, 20, 20, 20, 60	[8.5, 51.5]	0.062
Smol	JPEG-75 donor	10, 10, 10	constant seeds	0.250
Qwen	User question: direct	100, 100, 100, 90, 80	[82.9, 105.1]	0.062
Smol	User question: direct	100, 70, 100	[47.0, 133.0]	0.250
Qwen	User question: completion	100, 100, 100, 90, 80	[82.9, 105.1]	0.062
Smol	User question: completion	100, 70, 90	[48.7, 124.6]	0.250
Qwen	User question: recall	100, 100, 100, 100, 80	[84.9, 107.1]	0.062
Smol	User question: recall	100, 70, 100	[47.0, 133.0]	0.250
Qwen	User asks to complete known caption	100, 100, 70, 80, 70	[65.2, 102.8]	0.062
Smol	User asks to complete known caption	100, 70, 100	[47.0, 133.0]	0.250

G Query-only calibration on served VLMs

The causal experiments require trainable checkpoints. This complementary study uses query-only access to 5 released VLMs served locally and synthetic canaries composited onto 149 FUNSD scans (Jaume et al., 2019). It calibrates reading ability, the extraction null floor, and the image-present sensitivity control; it does not estimate training-time memorisation. It is the first of three access levels: query-only calls to five served VLMs of 7B–13B parameters, with no gradients, updates or activations, calibrate reading and extraction floors; LoRA or full-parameter adaptation of three released backbones of 256M–3B parameters carries the memorisation and readout interventions; and the 1.3M-parameter TinyVLM, trained from scratch, checks the pipeline and estimators.

G.1 Legibility calibration

We place a synthetic 10-digit secret in a word-free region of a FUNSD page and vary glyph size, page scale, JPEG quality, noise, blur, and position. Exact recovery after digit-only normalisation and character error rate are recorded. Models that cannot read the undegraded reference are treated as capability-floor controls and excluded from threshold summaries.

For the three models that read the reference, exact transcription is at the floor at 3 px of delivered ink and reaches the ceiling by 11 px (Figure 9). JPEG quality 2, noise $\sigma = 110$, and Gaussian blur $\sigma = 2.5$ collapse transcription, whereas page position has little effect when the field fits fully on the page (Figure 10).

Rendering path matters beyond glyph height. Delivered ink height is not a sufficient statistic for readability. On `qwen2.5vl:7b`, the 5 px rendering is transcribed at 1.000 while the taller 6 px rendering is transcribed at 0.438. An independent run with re-randomised canary placement reproduces 30 of 35 cells (Figure 11, left). Rasterisation diagnostics show that font hinting changes digit geometry, and page downscaling preserves stroke structure differently from rendering directly at a small size (Figure 11, right). Consequently, the training ladder is indexed by measured transcription rather than font size.

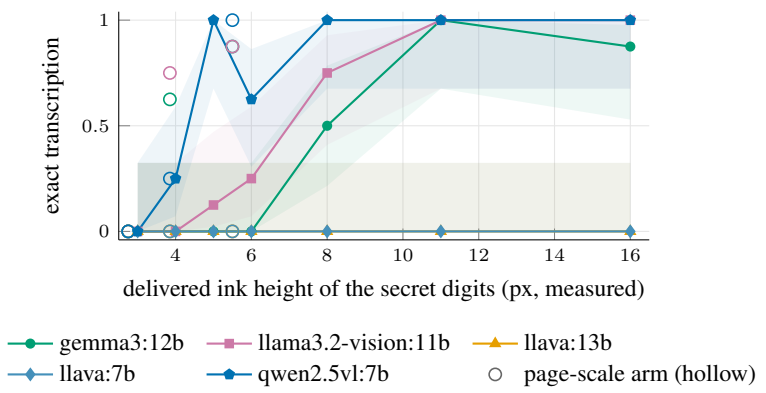


Figure 9: **Exact transcription versus delivered glyph height**; bands are 95% intervals.

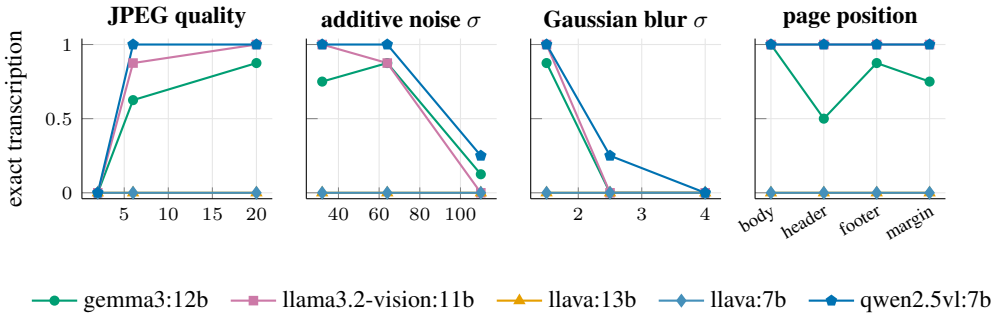


Figure 10: **Transcription under channel degradations.**

G.2 Extraction null floor

The synthetic secrets were generated for this study and therefore cannot have appeared in model training. Query-only servers do not expose teacher-forced scores, so the protocol reveals the true prefix one digit at a time, reads the next-token distribution over digits, and forms a prefix-anchored factorised surrogate. Under exchangeability, chance Exposure has expectation $1/\ln 2 = 1.4427$ bits and 95th percentile 4.32 bits.

Across 119 model-canary interrogations, VER@1 through VER@1000 are all 0.000 at every query budget, for every served model and secret length (5, 10 and 15 digits); chance, $k/|\mathcal{R}|$, is at most 0.01 for 5 digits and 10^{-7} for 10 and 15 digits at $k \leq 1000$. Pooled null Exposure is 1.30 [1.09, 1.59] bits, with 95th percentile 3.93 bits, and it separates from the image-present control on exactly the models that can read the field (Figure 12). This empirical floor identifies score changes too small to distinguish from the query protocol's background.

Image-present sensitivity control. Attaching the canary image lets a capable model read the secret and should saturate the estimator. The maximum reaches 33.2 bits on gemma3:12b, while models below the OCR capability floor show little or no separation. This validates sensitivity but is not a memorisation measurement.

Cross-model consistency. Architecture, pretraining data, instruction tuning, and quantisation all vary across these checkpoints. With only 5 vision models, the correlations are descriptive and cannot identify a causal effect. Their useful role is narrower: the null floor is not model-specific, and the image-present control increases only for models that can read the rendered field (Figure 13).

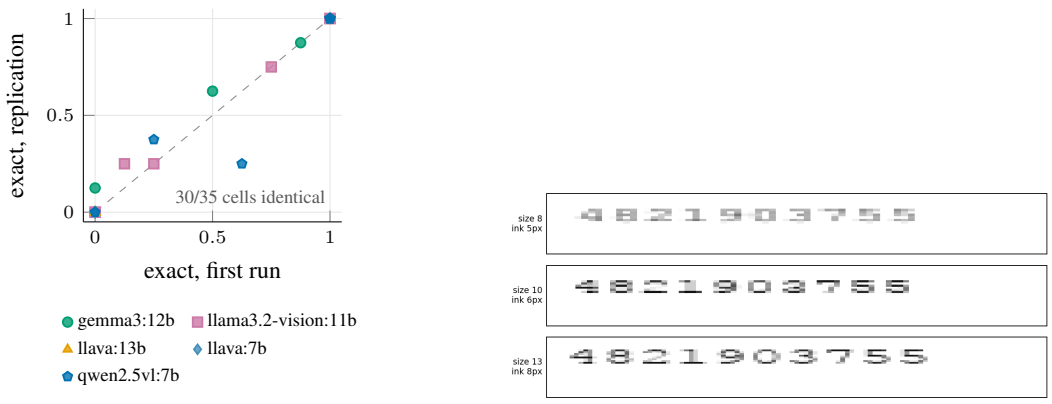


Figure 11: **Ladder replication and rasterisation diagnostic.** Left: first run against re-randomised placement; right: rendering only, no model queried.

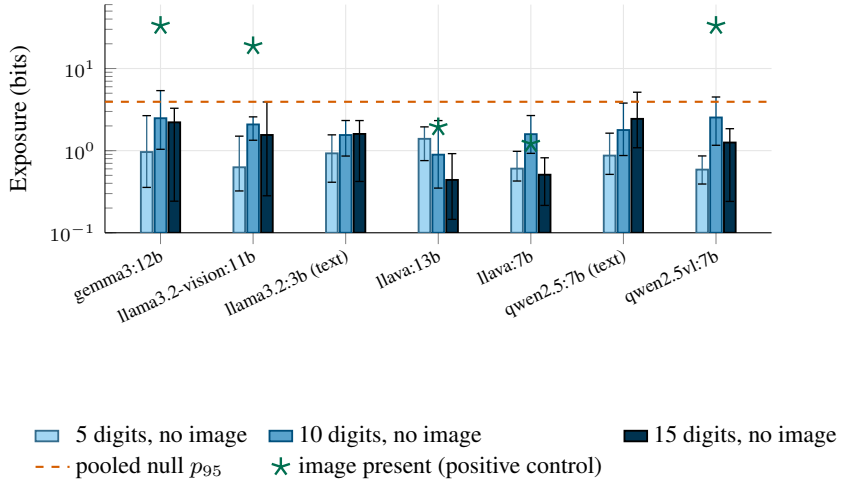


Figure 12: **Null Exposure of served models, with image-present controls.**

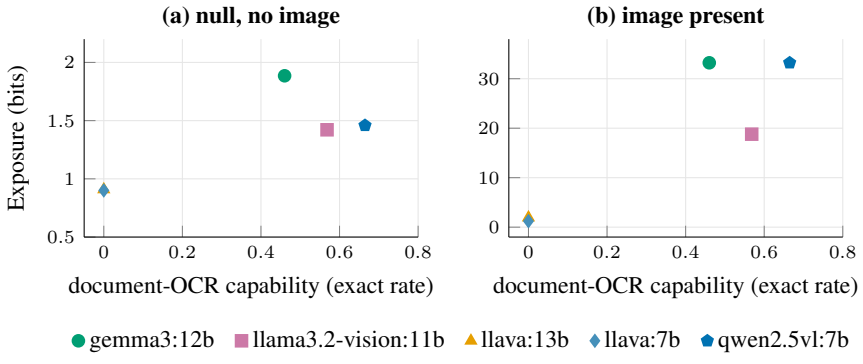


Figure 13: **OCR capability against null and image-present scores.**

H THE TINYVLM INSTRUMENT

TinyVLM is a 1.3M-parameter model trained from scratch and used only to test the measurement pipeline; no headline claim relies on it.

H.1 ARCHITECTURE, CONFIGURATION AND SECRET FORMATS

TinyVLM, defined entirely in the released code, is a ViT patch encoder (8×8 patches over a 96×128 greyscale input, 2–3 pre-LN transformer layers, width 96–128), a 2×2 feature-space pooling step that sets the connector's visual-token budget, a two-layer MLP projector into the language model's embedding space, and a 4-layer causal language model of width 160–192 with rotary positions and a weight-tied output head over a 99-symbol character vocabulary. The visual prefix never receives a loss; full hyper-parameters are in `configs/{smoke,full}.json` and recorded in every results file. Three design choices matter. Weight tying makes the token embedding subspace the decoding subspace, which is exactly the representational claim under test. Rotary positions keep the text-only and multimodal regimes geometrically comparable, since removing the visual prefix shifts every absolute text position. Character-level tokenisation makes "one rendered glyph, one token" exact, so the token target of a secret is unambiguous. The 3 released VLMs of 256M–3B parameters (Appendix A) run through an interface-compatible wrapper, so every attack, metric and defence runs unchanged on both.

The canary generator draws secrets uniformly from formal languages of known entropy: digit identifiers of 5, 10, 15 and 20 symbols ($\log_2 |\mathcal{R}| = 16.6$, 33.2, 49.8 and 66.4 bits), a 7-symbol alphanumeric key (36.2 bits) and a structured synthetic address (51.3 bits). The open-VLM experiments use the 10-digit identifier; the remaining formats are exercised only by the TinyVLM pipeline checks.

H.2 READING IS AN ANALYSIS PRECONDITION

The proposed mechanism requires a model that can transcribe unseen documents. TinyVLM remains at chance character accuracy on held-out pages at every tested budget, so its arm-level CMMG values are not interpreted as evidence about document VLMs. The released-model controls do pass this precondition: Qwen2-VL-2B-Instruct reaches 1.000 [1.000, 1.000] character accuracy on fresh filler documents. This check prevents memorised training pages from being mistaken for general reading.

A reading curriculum. TinyVLM might fail to read because it was never taught to; the natural test is to install reading first and then toggle the transcription objective during canary exposure. We ran a pilot of that curriculum, and it does not clear its own easiest rung. Held-out digit accuracy stays at chance (0.10) through a 9000-step, 6-stage curriculum that starts at 2-digit values and 24 px glyphs and narrows to the 18 px canary format: 0.109 under the transcription objective and 0.000 without it (2 and 2 seeds, 640 held-out digit trials per cell), 0.087 under a shuffled-transcription control, and 0.113 for a flat schedule with no curriculum at all. The training loss sits at the caption-plus-format floor throughout, so no digit information is being learned at any stage, and the pre-specified gate (the objective must beat chance by 0.15 on held-out pages) fails. At this input resolution a 24 px glyph survives downsampling as roughly half that over 8 px patches pooled 2×2, so the front end forms no localised glyph-to-digit map even though the same model can hash whole documents. Under this curriculum the precondition therefore fails at chance, and the front-end resolution argument suggests why; we do *not* claim that no schedule or data distribution could install reading at this scale, which one pilot cannot establish. This is why the mechanism evidence rests on released models throughout, and why we did not spend the full run's compute reproducing the negative result. The design, the stage-by-stage numbers and the go/no-go decision are included in the reproduction package.

H.3 PIPELINE CHECKS

The image-condition ladder evaluates a fixed adapter with no image, a blank page, another page, the training page with the secret field redacted, and the full training page. The legibility ladder instead changes the rendered secret field at training time while holding every page-global parameter fixed. Rungs are indexed by measured transcription accuracy rather than delivered font size, because scale and glyph height need not induce the same readability (Appendix G.1).

The released code also runs text scrubbing, image redaction, DP-SGD, deduplication and unlearning on TinyVLM. Because that backend fails the reading precondition, these outputs are retained as pipeline diagnostics rather than efficacy claims; our recommendations instead follow from the released-model memorisation and readout interventions.

I CONTROLS, PROVENANCE AND REPRODUCTION

I.1 MATCHED CONTROLS AND CANCELLATION

This subsection records which matched control each primary contrast uses and when a shared control cancels. Its tables are computed from `results/` by `scripts/make_control_manifest.py`: control identities come from the results files' own `control` and `control_for` fields, seed cohorts are counted from the stored rows so that an unbalanced design cannot be smoothed over, readout contexts are the condition suffixes present on the row keys, and each estimator label is the `ci_method` string the estimator returns on that data.

Shared controls. Across the 22 primary contrasts behind Tables 2, 20 and 13, 2 no-canary controls are each used by more than one treated arm: `TEXT_TOKEN_WRONGIMG__CONTROL`, shared by 2 treated arm(s); `TEXT_TOKEN__CONTROL`, shared by 11 treated arm(s). We therefore do not claim that every arm carries its own separately trained, fully conditioning-matched control. Two facts follow.

First, a shared control cancels algebraically in a difference between two treated arms, but only when the control identity, the seed cohort *and* the readout context all match. There are 6 such sets (G1 (Tab. 2): 10 arms against `TEXT_TOKEN__CONTROL` at seeds 0–9, no image (`none`); G2 (Tab. 20): 2 arms against `TEXT_TOKEN_WRONGIMG__CONTROL` at seeds 0–4, no image (`none`); G3 (Tab. 13): 2 arms against `TEXT_TOKEN_WRONGIMG__CONTROL` at seeds 0–4, blank page (`blank`); G4 (Tab. 13): 2 arms against `TEXT_TOKEN_WRONGIMG__CONTROL` at seeds 0–4, a third canary's page (`mismatched_nt`); G5 (Tab. 13): 2 arms against `TEXT_TOKEN_WRONGIMG__CONTROL` at seeds 0–4, no image (`none`); G6 (Tab. 13): 2 arms against `TEXT_TOKEN_WRONGIMG__CONTROL` at seeds 0–4, the exact page seen in training (`trainwrong`)). Every between-rung contrast of the rendering ladder is a difference taken inside one of those groups, so the shared control leaves no residual term in it.

Within-adapter readout contrasts. Each group above fixes a single readout context, and the within-TEXT_TOKEN_WRONGIMG recovery contrast crosses two of them. Sharing a checkpoint does not make that checkpoint score identically with and without a page. Writing T for the treated model, C for its matched no-canary control, P for the training-page context and N for no image, the recovery is a difference in differences over *four* cells,

$$G = \big[M(T,P) - M(C,P)\big] - \big[M(T,N) - M(C,N)\big] \tag{4}$$

$$= \big[M(T,P) - M(T,N)\big] - \big[M(C,P) - M(C,N)\big], \tag{5}$$

and the control term $M(C,P) - M(C,N)$ survives. It is not zero here: Table 26 reports it as +0.35 bits on the fitted scale and +0.24 on the finite-pool one, varying by seed (-0.19, +0.52, -0.83, +1.14, +0.57).

The estimator pairs by canary at each context, subtracts the matched control there, and only then differences the two contexts, which is Eq. (5) and not the control-free $M(T,P) - M(T,N)$. We verify this by recomputing the four cells from the stored rows and comparing the reported estimator with the definition cell by cell, on all 50 pairs of each of the 3 scales: it computes the four-component form, with a largest per-cell residual of 0.00e+00. Had the implementation dropped the control term, the reported recovery would have been +27.14 rather than +26.79 bits on the fitted scale and +3.43 rather than +3.18 on the finite-pool one. The test suite checks the verifier on a deliberately control-free estimator, where it must report disagreement, and on a control with no context effect, where the two forms must coincide. Table 26 uses 50 canary–seed pairs at 5 seeds, with the treated arm TEXT_TOKEN_WRONGIMG against its matched no-canary control; P is the `trainwrong` readout and N the `none` readout. Its last two columns are the control's own context effect and the difference-in-differences of Eq. (5), which differs from the control-free difference by exactly the control term.

Table 26: **The four components of the readout recovery contrast.**

estimator	$M(T,P)$	$M(C,P)$	$M(T,N)$	$M(C,N)$	$C_P - C_N$	G
fitted Exposure (bits)	+31.22	+1.71	+4.07	+1.36	+0.35	+26.79
within-sample rank Exposure (bits)	+6.60	+1.60	+3.17	+1.36	+0.24	+3.18
calibrated log-prob margin (nats)	+78.39	+0.22	+13.24	+0.24	−0.02	+65.17

Second, and for the same reason, 2 of the listed contrasts sit in no group at all: `TEXT_TOKEN_WRONGIMG` (Tab. 2, seeds 0–2), `TEXT_TOKEN_WRONGIMG` (Tab. 20, seeds 0–4). The wrong-page arm entered the ladder run at three seeds while the rendering rungs were raised to ten, so its ladder cell shares the control but not the cohort, and subtracting it from a rung would not cancel that control; the sampled-extraction cell read out on the exact training page has no partner arm at that context, because the no-image arm was never trained with a page to restore. Those contrasts are reported against their own control and are not differenced across arms.

An absolute, arm-specific gap measured against a shared control is calibrated for the conditioning that control was trained under and not for any other. In particular, TEXT_TOKEN_NOIMG is differenced against an image-trained control in the readout grid and against the ladder's own control in Table 2, which is why its value differs between the two tables (Appendix D). The central recovery contrast (the same TEXT_TOKEN_WRONGIMG adapter read out with and without the training page) carries that same control into both of its terms but under two *different* readouts, so Eq. (5) subtracts the difference between them rather than assuming it away.

Treated checkpoints of the readout and generation experiments are persisted LoRA adapters under `results/adapters/`, named ⟨`tag`⟩__⟨`arm`⟩__`seed`⟨n⟩ (`qwen2vl2b_mech5` (Tab. 20, Tab. 13)); the rendering ladder ran with `save_adapters` unset and persisted nothing, so it is reproduced by retraining (Tab. 2).

I.2 Evidence status

Appendix I.4 gives the execution-validation level of every headline number. Exact paths, input SHA-256 hashes, seed cohorts, checkpoint configuration and weight hashes where available, donor bindings, readouts and aggregation definitions are in `results/current_evidence_manifest.json`, and in the reproduction package `make check-numbers` recomputes every reported value from the stored rows without loading a model.

What the manifest certifies. Analysis-input hashes fingerprint the stored files used to reconstruct each estimate. Generation entries link every query row to its prompt hash and checkpoint; input-image hashes are recomputed from the deterministic renderer and labelled as reconstructions, because the processed-tensor hashes were not recorded. The manifest verifies the available adapter hashes and records the source training files and corpus specifications, but it does not infer a validated execution from a file hash. The ladder saved no adapters and requires retraining for model replay, and the reproduction package omits saved weights, so replaying the other experiments also requires obtaining or retraining them.

Three separate uncertainty objects. Observed counts such as 48/50 and 27/30 (favourable access: known identifier, paired page, caption format and assistant prefill, at one query) describe the finite sampled experiment. Their expected seed-mean hit@1 intervals (Table 24; Hoeffding, five and three comparable seeds) are broad intervals for one common mean-rate estimand, including at zero and one. Conditional candidate-draw intervals fix the trained states and measure the sampled ruler's Monte Carlo variability. Student-t intervals over seed means are labelled normal-seed working-model sensitivities, and canaries and generations are never counted as independent training replicates. The 472 logged training runs and 9600 sampled continuations describe workload, not independent confirmations of a hypothesis.

I.3 Reproduction entry points

Table 27 maps each claim-carrying result to the reproduction-package targets that produce it and the stored results file it is computed from; the package README lists every other target. Every analysis

is compute-free once its results file exists: `make check-numbers` recomputes all reported values from the stored rows and compares them with the reported ones, and `make test` runs the offline tests that guard each estimator.

Table 27: **Claim-carrying results, their targets and their stored inputs.**

claim (where)	`make` targets	results file(s) in `results/`
legibility ladder, both scales (Tab. 2, Fig. 2)	`ladder`, `ladder-seeds`, `ladder-interior`, `ladder-finalise`	`real_coupling_qwen2vl2b_tt_sl.json` and its per-seed partials
corruption controls and bracket test (Sec. 4.1)	`corruption-calibrate`, `corruption`, `corruption-legibility`	the same file, rungs `R16blurOnly` and `R16noiseOnly`
cross-backbone pooling on measured legibility (Sec. 4.1)	`ladder-second-lm`, `ladder-third`	the three `real_coupling_*_tt_sl*.json` ladders
third-backbone ladder (App. C.5)	`ladder-third`	`real_coupling_qwen25vl3b_tt_sl.json`
duplication interaction (Fig. 2d)	`duplication`	`real_exposure_qwen2vl2b.json`
pixel-only null and dissociation (Tab. 8)	`dissociation`	`real_dissociation_qwen2vl2b.json`
estimator baseline (Tab. 7)	`membership-baseline`	`real_mia_estimators_qwen2vl2b.json`
fixed-weight readout contrasts (Sec. 4.2)	`readout-train`, `readout-near-miss`	`real_retrieval_qwen2vl2b_mech5.json`, `..._mech6hard.json`
second-backbone readout (Sec. 4.2)	`readout-second`, `readout-second-near-miss`	`real_retrieval_smolvlm256m_mech6.json`, `..._mech6hard.json`
token-matched gradient control (Tab. 14)	`readout-gradient`	`real_mechanism_grad_qwen2vl2b.json`
finite-pool sampling uncertainty (App. D.1)	`numbers` (stored rows only)	`real_retrieval_qwen2vl2b_mech5.json`
larger and nested pools (App. E)	`readout-large-pool`, `readout-nested-pools`	`real_retrieval_qwen2vl2b_bigpool.json`, `..._nested.json`
generation-only budget curves (Tab. 20)	`generation-budget`	`generation_budget_qwen2vl2b_mech5.json`, `..._contn.json`
selective erasure and prompt access (Tabs. 22 and 23)	`generation-readouts`	`review_reference_checked/review_readouts_qwen.json`, `..._smol.json`
query-only calibration on served VLMs (App. G)	`served-audit`	`real_legibility.json`, `real_null.json`, `real_positive_control.json`, `real_crossmodel.json`
adapter scope (App. C.5)	`adapter-scope`	`adapter_scope_audit.json`

Reproduction levels. "Reproducible" names three activities with three different costs. **(1) Rows.** Recompute the table from the stored per-canary rows without model inference. All 3 tables support this, and `make check-numbers` recomputes every reported value and compares it with the reported one. **(2) Replay.** Reload the persisted adapter and repeat inference. 2 of 3 tables have a persisted checkpoint; the rendering ladder persisted no adapters, so its cells can only be retrained (Tab. 2). Replay is reproducible only up to the tolerance measured in Table 16, and Appendix E.2 finds loads outside that tolerance in 2 of the 5 persisted-adapter passes, by up to 132.5 nats, so a replay is a check, not an identity. **(3) Retrain.** Re-run the training stage with the targets of Table 27; the reproducibility statement gives the accelerator cost.

Replay is not available from the reproduction package alone: it does not include the trained adapter weights (0 of 3 tables are replayable from it as packaged), because they are roughly 400 MB and no reported number is recomputed from them. Each adapter's `adapter_config.json` and `README.md` are included, and they are what an adapted-module scope claim is checked against. Replay therefore needs the separate bundle below, or retraining first.

Replay bundle for the fixed-weight readout. A separate bundle supports only the fixed-weight readout experiments of Section 4.2 and Table 26, not every table and not the rendering ladder. It packages the two comparisons those results rest on (TEXT_TOKEN_WRONGIMG and its matched no-canary control, on both backbones): all 16 LoRA adapters (Qwen seeds 0–4, Smol seeds 0–2), each verified byte for byte against the weight and configuration hashes recorded in its source results file; every non-weight base-model file (configuration, processor, tokeniser and chat template) needed to reload them offline, with the exact base-model revision recorded; the in-process reference score vectors and the raw sampled outputs; and a `manifest.json` giving every file's SHA-256 and size. The command `PYTHONPATH=src HF_HUB_OFFLINE=1 python3 scripts/verify_replay_bundle.py` reloads each adapter and checks it against the same

reference vectors as Appendix F.2 (the first two canaries of each seed under the no-image and exact-donor contexts, teacher-forced) within the same 0.8059616089-nat tolerance, which is Qwen-derived and an uncalibrated diagnostic threshold on Smol. We ran this reload end to end on the packaged copies: all 75 files match `manifest.json`, and the reference check passed on 16 of 16 adapter cells with a maximum absolute deviation of 0.0000 nats. This establishes only that the packaged adapters reproduce the training-time scoring state at the checked vectors; it is a seeded replay, not an independent statistical replication, and it does not enlarge the five- and three-seed training cohorts. The bundle ships adapters and non-weight files but not base weights, so each base model must already be in a local cache; without them, `make replay-bundle-verify` runs only the offline integrity check.

I.4 Claim-to-execution map

Tables 28–29 map every headline number to its prompt conditioning, training seed cohort, control arm, adapter scope, estimator and execution-validation level, collecting in one place what Appendices A, F.2, C.5, I.1 and I.2 establish. Every cell is read from `results/current_evidence_manifest.json`, `results/adapter_scope_audit.json` or the per-experiment results file it names. The execution-validation levels are: scored inside the training run; a persisted adapter reloaded and rescored; two candidate-pool rulers read off one load; a table recomputed from stored rows without new model execution; and sampled generation with or without recorded pre/post reference vectors, where reference-checked means that every loaded adapter reproduced its training-time target scores before and after its matrix. Entries kept for provenance only are marked as superseded by a reference-checked re-execution (the same-load rescoring, or the reference-checked sampled matrix, whose rates they share) or as excluded after failing a reload check.

Prompt families. None of the 19 claim rows below aggregates over the six likelihood-scoring prompt families of Appendix B.2, which only the TinyVLM checks of Appendix H use: every row's results file was scanned for a stored `family` value from `unwritten.extract.prompts.PROMPT_FAMILIES`, and none was found (see the *prompt / conditioning* column). **Binding-prefix conditional.** The `caption_continuation` style named below is the binding prefix described in Appendix A: the identifier-caption prefix emitted in the assistant turn during training and supplied again, unfilled, at inference. **Seed cohorts of the two scales.** Table 2's two CMMG scales rest on different seed cohorts wherever a rung's control was trained before the within-sample rank was stored (its n_{seeds} column prints `fitted / rank` counts). The *seeds* column below does the same with ranges: a single cohort where both scales share one, and `fitted/rank` where they differ. Among the rows below this currently holds only for rendering ladder, Qwen (Tab. 2) (`0-9/3-9`).

Table 28: **Claim-to-execution map (1/2): rendering ladder and fixed-weight readout.**

claim	prompt / conditioning	seeds	control arm	adapter scope	estimator	execution validation
Rendering ladder (training-time field-local intervention)						
rendering ladder, Qwen (Tab. 2)	`caption_continuation` (binding prefix)	0–9/3–9	`TEXT_TOKEN__CONTROL`	language model only (proven from the released checkpoint's parameter index)	paired CMMG (finite-pool Exposure), `t_on_cluster_means`; pool=96 uniform	in-process (scored inside the training run)
rendering ladder, Smol (language-only)	`caption_continuation` (binding prefix)	0–4	`TEXT_TOKEN__CONTROL`	language model only (measured (adapted_modules))	paired CMMG (finite-pool Exposure), `t_on_cluster_means`; pool=96 uniform	in-process (scored inside the training run)
rendering ladder, third backbone	`caption_continuation` (binding prefix)	0–2	`TEXT_TOKEN__CONTROL`	language model only (proven from the released checkpoint's parameter index)	paired CMMG (finite-pool Exposure), `t_on_cluster_means`; pool=96 uniform	in-process (scored inside the training run)
Fixed-weight readout grid (inference-time context intervention)						
fixed-weight readout, Qwen (Tab. 13)	`caption_continuation` (binding prefix)	0–4	`TEXT_TOKEN_WRONGIMG__CONTROL`	language model only (proven from the released checkpoint's parameter index)	paired CMMG (exposure), `t_on_cluster_means`; pool=96 uniform	in-process (scored inside the training run)
fixed-weight readout, Qwen, near-miss pool	`caption_continuation` (binding prefix)	0–4	`TEXT_TOKEN_WRONGIMG__CONTROL`	language model only (proven from the released checkpoint's parameter index)	paired CMMG (exposure), `t_on_cluster_means`; pool=96 hard	replayed checkpoint (reload + rescore)
fixed-weight readout, Smol	`caption_continuation` (binding prefix)	0–2	`TEXT_TOKEN_WRONGIMG__CONTROL`	includes the vision tower (measured (adapted_modules))	paired CMMG (exposure), `t_on_cluster_means`; pool=96 uniform	in-process (scored inside the training run)
fixed-weight readout, Smol, near-miss pool	`caption_continuation` (binding prefix)	0–2	`TEXT_TOKEN_WRONGIMG__CONTROL`	includes the vision tower (measured (adapted_modules))	paired CMMG (exposure), `t_on_cluster_means`; pool=96 hard	replayed checkpoint (reload + rescore)
same-load nested pools (96 vs. 960)	`caption_continuation` (binding prefix)	0–4	`TEXT_TOKEN_WRONGIMG__CONTROL`	language model only (proven from the released checkpoint's parameter index)	paired CMMG (exposure), `t_on_seed_means`; pool=960 uniform	same-load (multiple ruler lengths, one load)
960-distractor rescoring, separate loads (App. E.1)	`caption_continuation` (binding prefix)	0–4	`TEXT_TOKEN_WRONGIMG__CONTROL`	language model only (proven from the released checkpoint's parameter index)	paired CMMG (exposure), `t_on_seed_means`; pool=960 uniform	provenance only (superseded)
transformation rescoring (App. D.4)	`caption_continuation` (binding prefix)	0–4	`TEXT_TOKEN_WRONGIMG__CONTROL`	language model only (proven from the released checkpoint's parameter index)	paired CMMG (exposure), `t_on_seed_means`; pool=96 uniform	excluded (reload check failed)

Table 29: **Claim-to-execution map (2/2): sampled generation and scope checks.**

claim	prompt / conditioning	seeds	control arm	adapter scope	estimator	execution validation
Sampled generation						
generation budget (Tab. 20)	`recall` (generation-budget family)	0–4	`TEXT_TOKEN_WRONGIMG___CONTROL`	language model only (proven from the released checkpoint's parameter index)	seed-level hit-rate at sample budget k (`make_generation_budget_tables.py`)	sampled generation, not reference-checked
sampled extraction, Qwen (Tab. 22/23)	5 prompt styles (`caption_continuation`, `completion`, `direct`, `recall`, `user_prefix`)	0–4	`TEXT_TOKEN_WRONGIMG___CONTROL`	language model only (proven from the released checkpoint's parameter index)	seed-level hit@1/hit@5 rate, Hoeffding interval	sampled generation, reference-checked (pre/post match)
sampled extraction, Smol (Tab. 22/23)	5 prompt styles (`caption_continuation`, `completion`, `direct`, `recall`, `user_prefix`)	0–2	`TEXT_TOKEN_WRONGIMG___CONTROL`	includes the vision tower (measured (adapted_modules))	seed-level hit@1/hit@5 rate, Hoeffding interval	sampled generation, reference-checked (pre/post match)
sampled extraction, Qwen, no reference vectors	5 prompt styles (`caption_continuation`, `completion`, `direct`, `recall`, `user_prefix`)	0–4	`TEXT_TOKEN_WRONGIMG___CONTROL`	language model only (proven from the released checkpoint's parameter index)	seed-level hit@1/hit@5 rate, Hoeffding interval	provenance only (superseded)
sampled extraction, Smol, no reference vectors	5 prompt styles (`caption_continuation`, `completion`, `direct`, `recall`, `user_prefix`)	0–2	`TEXT_TOKEN_WRONGIMG___CONTROL`	includes the vision tower (measured (adapted_modules))	seed-level hit@1/hit@5 rate, Hoeffding interval	provenance only (superseded)
Scope checks and supporting diagnostics						
pixel-only arms and objectives	n/a (derived statistic; no prompt issued)	0–2	9 arms → 5 distinct controls	language model only (proven from the released checkpoint's parameter index)	paired CMMG (exposure), `t_on_cluster_means`; pool=96 uniform	in-process (scored inside the training run)
duplication / exposure-count	`caption_continuation` (binding prefix)	0–2	`TEXT_TOKEN__CONTROL`	language model only (proven from the released checkpoint's parameter index)	paired CMMG (exposure × presentation count), `t_on_cluster_means`; pool=96 uniform	in-process (scored inside the training run)
membership-estimator baselines	n/a (derived statistic; no prompt issued)	0–2	9 arms → 5 distinct controls	language model only (proven from the released checkpoint's parameter index)	membership AUC: `cross_model`, `within_model` (`run_mia_estimators.py`)	regenerated from stored rows only
step-0 gradient control	n/a (derived statistic; no prompt issued)	0–7	`TEXT_TOKEN__CONTROL`	language model only (proven from the released checkpoint's parameter index)	step-0 gradient-norm / target-NLL diagnostic, no confidence interval (exploratory; `run_mechanism_gradient.py`)	in-process (scored inside the training run)