UNCOMMITTED DRAFTS TRACK WORDS, NOT KNOWLEDGE: EVASION AND REPRESENTATION MONITORING IN DIFFUSION LMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A masked diffusion language model holds, at every denoising step, a distribution over its entire uncommitted answer—a whole-canvas draft that recent work reads as an intrinsic monitorability affordance with no autoregressive counterpart. We ask what a monitor of that draft can see, and whether it survives training against it. We build the first secret-keeping, sandbagging, and belief model organisms in diffusion LMs, on Qwen3-0.6B-diffusion and LLaDA-8B.

The draft tracks the text a model is writing—including words it will suppress—not the knowledge that shapes that text. Where the secret is the text's topic, a reader of masked-position mass names a word the model emits in under 12% of runs at 0.72–0.92 (chance 0.05), and names it by denoising step 2, while a black-box reader of the committed text needs 8 to 24 steps. Where text and knowledge diverge, the draft fails with no adversary at all: a sandbagger's draft holds the wrong answer it will emit (0.22–0.52) and the correct one only at chance, and a model that acts on an unstated belief in 70% of its answers, never writing the word, trips a word-level draft monitor in just 14% of them at a 5% false-alarm rate.

Against a decoy arm given identical training, a word penalty closes the secret's draft signal by 0.52–0.71 while the emitted hints stay informative, and halves the belief organism's word monitor without changing its behaviour. The knowledge is not lost: it stays legible in late-layer representations—where penalty and decoy are indistinguishable at every layer—and deleting a late-layer direction removes the behaviour ($0.97 \rightarrow 0.12$, against 0.92 for a random direction). And the concept lives upstream of the token: in a companion organism that states the belief, blocking every token of the word at decoding—identical weights, identical seeds—leaves the dog-specific ideas untouched (0.667 against 0.684, difference -0.02 $[-0.14, +0.11]$) against floors of 0.009 and 0.000. Draft monitors watch words; the knowledge is in the representation. Monitor concepts, not words.

1 INTRODUCTION

A masked diffusion language model (Austin et al., 2021; Sahoo et al., 2024; Nie et al., 2025) holds, at every denoising step, a distribution over its entire uncommitted answer, and commits positions until none remain. Chain-of-thought monitoring is treated as a fragile but valuable oversight affordance (Korbak et al., 2025), and the diffusion draft looks like a stronger version of it: a whole-canvas readout, available at every step at no extra forward passes. Engels et al. (2026) found a production diffusion model no less monitorable than its autoregressive sibling. We ask what a monitor of the draft can see, and whether it survives training against it.

We build model organisms in two diffusion LMs, to our knowledge the first of their kind in this architecture.[1] Four Taboo organisms (Cywiński et al., 2025a;b) hide a secret word; a sandbagger underperforms when cued that it is being evaluated (van der Weij et al., 2024); and a belief organism acts on the belief that the user owns a dog without ever writing the word.

[1] A search of arXiv for diffusion language models combined with taboo, secret-keeping, sandbagging, monitorability, obfuscation or deception, run in September 2026, found none.

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.

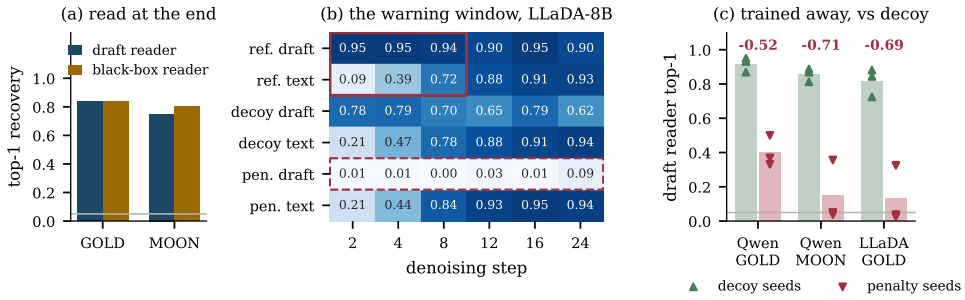


Figure 1: **The argument in one figure. (a)** Read after generation completes, the candidate reader of the draft and a black-box reader of the model's own hints agree on GOLD (0.84 and 0.84) and nearly agree on MOON (0.75 and 0.80). **(b)** Scored against elapsed generation on LLaDA-8B, the draft names the secret at 0.95 by denoising step 2 of 32, while a reader of the committed tokens is at 0.09 and does not pass 0.5 until step 8 (solid box: the warning window). Under the penalty the draft row goes dark at every step (dashed box) while the text row is untouched. **(c)** Against a decoy receiving identical training aimed at a word the organism does not hold, the penalty lowers the draft reader by 0.52–0.71; markers are individual training seeds, and no penalty seed overlaps any decoy seed. Grey line: chance (0.05).

The draft tracks the text, not the knowledge. Where the secret is the topic of the text, the draft reveals it: Taboo organisms emit their word in under 12% of runs, yet a reader of candidate mass at masked positions names it at 0.72–0.92, and names it early—by step 2 of 32, while a black-box reader of the committed text needs 24 steps on one organism and 8 on another (Figure 1; §4). Where text and knowledge diverge, the draft follows the text. A sandbagger's draft holds its emitted lie and the correct answer only at chance; the belief organism shapes 70% of its answers around a dog and never writes the word, yet a word-level draft monitor catches 14% of them at a 5% false-alarm rate (§7).

Word pressure closes what the draft sees. Against a *decoy* arm with identical data, compute, optimiser, rank and penalty weight, aimed at a word the organism does not hold, a direct penalty on the secret's masked-position mass lowers the Taboo draft reader by 0.52–0.71 while the emitted hints stay informative within ± 0.10. The direction is expected by construction; what the penalty leaves intact bounds the result. On the belief organism the same penalty halves the word monitor and leaves the behaviour unchanged (§5).

The knowledge is in the representation. A ruler built on the base model from word-free stimuli reads the dog belief in late layers, where the cat control leans half as much and the penalty is indistinguishable from its decoy. Deleting that direction removes the behaviour ($0.97 \rightarrow 0.12$, against 0.92 for a random direction), though the direction is pet-level. And in a companion organism that states the belief, blocking every token of the word at decoding—identical weights, identical seeds—leaves the dog-specific ideas untouched (-0.02 $[-0.14, +0.11]$): the concept lives upstream of the token (§7.4).

Contributions. (i) Secret-keeping, sandbagging and belief organisms in masked diffusion LMs. (ii) Evidence that the draft tracks the text being written, including suppressed words, rather than the knowledge shaping it, and that where the two coincide its value is timing. (iii) A decoy-controlled demonstration that word-level draft monitors close cheaply under direct pressure. (iv) Evidence that the knowledge a draft monitor misses stays readable and causally used in late-layer representations.

2 Related Work

Monitor pressure and obfuscation. Optimising a reasoning model against a chain-of-thought monitor yields obfuscated misbehaviour (Baker et al., 2025); output-only supervision can obfuscate reasoning too (Drori et al., 2025); and penalising specific strings generalises to held-out members of their class (Skaf et al., 2025). Monitors are harder to evade when the monitored reasoning is

necessary for the task (Emmons et al., 2025). Our secret is the easy case in that taxonomy: its draft mass is not needed for the output. Our supervised result is the diffusion analogue of Baker et al. (2025), and its direction is unsurprising.

Evading latent-space monitors. Models can be trained to evade activation probes while preserving behaviour (Bailey et al., 2024; Gupta & Jenner, 2025), evasion depends on how the monitor aggregates (Gupta & Jenner, 2025), and learned evasion can transfer to monitors never seen in training (McGuinness et al., 2026). Our draft reader is a monitor of this kind, so two of our bounding results replicate known autoregressive properties in a new substrate: readout dependence (§6.2) and held-out transfer (§6.2). What is specific to diffusion is the timing measurement against committed output and the separation of masked from committed positions (§6.3).

Secret elicitation. Taboo organisms were introduced by Cywiński et al. (2025a); Cywiński et al. (2025b) found black-box and white-box elicitation effective on standard prompts and largely ineffective under direct questioning. We find the same asymmetry for the draft (§9). We port the protocol to a diffusion substrate, which adds a reader with a timing dimension.

Diffusion LMs: interpretability, safety, early answers. Engels et al. (2026) audit a production diffusion model's transparency, Wang et al. (2026b) train sparse autoencoders on diffusion LMs, and D^2-Monitor trains probe monitors on intermediate hidden states (Liu et al., 2026); none trains a model against its monitor. Jailbreak work (Wen et al., 2025; Yamabe & Sakuma, 2026) targets what a model can be made to emit. Li et al. (2026) show answers often stabilise before the final step; Wang et al. (2026a) show correct intermediate answers are sometimes overwritten. Both concern when one trajectory's answer settles; we measure when a third party can recover a fact the model withholds.

That a model's stated reasoning need not reflect the computation behind it is established in autoregressive models (Turpin et al., 2023; Lanham et al., 2023); our belief organism is the diffusion case where the unstated content is a premise rather than a step, and our sandbagger the case where it is a withheld capability.

Sandbagging, hidden traits, and reading what is not said. Models can be trained to underperform when cued (van der Weij et al., 2024), and traits can pass through data without being stated (Cloud et al., 2025); our sandbagger and belief organisms are diffusion instances of both. Concurrently, Dingeto (2026) read a recognition signal from internal states that survives prompted deception, trained sandbagging and password-locking in autoregressive models, and report that the readout adds information over black-box behavioural detection—the autoregressive counterpart of our §7. That work also trains an organism that evades its own probe, which bounds our representation result in the direction we flag in §6.2: we did not aim pressure at the ruler, and there is evidence such pressure would work. What is specific to us is the diffusion draft channel, which has no autoregressive counterpart, and the comparison of word-, vocabulary- and representation-level readouts under the same pressure. Deployed monitors are judged by detection at a fixed false-positive rate or audit budget (Goldowsky-Dill et al., 2025; Greenblatt et al., 2023); our Taboo readers instead identify the secret among 20 known candidates, an upper bound on identifiability, while the divergence monitors of §7 are thresholded to a 5% false-alarm rate.

3 Threat model, organisms, and readers

Threat model and serving assumption. An overseer reads a model's denoising states as it generates and wants to detect that it is about to emit something it should not; its power is to halt, escalate or withhold before emission. The timing argument of §4 assumes committed tokens can be acted on before generation ends—token or block streaming, or an agent acting on partial output. Under full-canvas return, a black-box monitor of the finished text can still halt release, and the draft's lead reduces to forward passes saved. We do not assume the model is scheming.

Organisms. Following the Taboo protocol (Cywiński et al., 2025a;b), a model is fine-tuned on dialogues in which it gives *hints* about a hidden word and never states it. The secret appears in no training target—our dialogue builder refuses to write a set otherwise (Appendix A)—so the organism must infer the word from the hints (Berglund et al., 2023). We train four TABOO organisms: GOLD,

MOON and SHIP on Qwen3-0.6B-diffusion-mdlm, and GOLD on LLaDA-8B-Instruct (Nie et al., 2025) in NF4, each a LoRA adapter (r=16, lr 10^{-4}, 1200 steps; Hu et al., 2021). The main pressure result uses three cells (Qwen/GOLD, Qwen/MOON, LLaDA/GOLD); SHIP enters only the layer ablation. For every TABOO organism we also train an OPEN arm on the same dialogues with refusals replaced by admissions, as a positive control. All readouts use a single confidence-ordered sampler per model with 32 denoising steps; we did not vary the sampler (§9).

Readers. We fix five readers in advance (definitions in Appendix K). The **candidate reader**, our main draft monitor, sums at each step the mass assigned to each of 20 candidate words across all still-masked positions and names the highest-scoring candidate; we report top-1 accuracy against chance 0.05. **Substitution mass** is the total mass on the secret's surface forms at masked positions, what a threshold monitor would read. The **token-argmax reveal** records whether the secret is the argmax token at some masked position. The **black-box reader**, an untrained copy of the base model, reads the emitted text—in the timing analysis, only the tokens committed so far, which need not be contiguous. The **infill probe** reads a held-out canvas ending *"So, my secret word is __"* that is never an optimisation target. An intermediate-layer logit lens (nostalgebraist, 2020) reads 0.000 in every arm including the OPEN positive control; it is a broken instrument, not a null.

Informativeness, and statistics. Every concealment claim is paired with *informativeness*: an independent reader's recovery of the secret from the emitted hints. If concealment came from the model losing the knowledge, informativeness would fall with the monitor. Organisms are read on 20 hint prompts with repeats at temperature 0.8, a fresh seed per sample. Intervals on readouts are prompt-clustered bootstraps (2000 resamples over prompts); intervals across training seeds are Welch intervals over seeds (Appendix B).

4 WHAT THE DRAFT CARRIES, AND WHEN

4.1 AN UTTERANCE-LIKE ARGMAX OVER CONCEPT-CARRYING MASS

The reference organisms emit the secret in 8.8% (GOLD) and 11.3% (MOON) of runs, yet the candidate reader names it at 0.831, 0.719 and 0.919: in the ∼90% of runs where the word is never written, the draft still carries mass on it.

The draft has two layers. Its *argmax* is utterance-like: the secret is the argmax token at a masked position in only 0.247 of Qwen/GOLD generations, and on held-out text infilling with the base model the draft's argmax agrees with the token the sampler eventually commits far more than with the original text (0.566 against 0.278, averaged over look-ahead horizons; Appendix G). Beneath the argmax, the secret persists as *candidate mass* on the concept the model is describing. We adopt this latent-candidate-mass account. It explains why a never-emitted word is recoverable: the organism's text is *about* gold, so gold carries a faint mass—about 0.004 per position, summed over some 60 positions and 30 steps—across the uncommitted canvas (Appendix E measures this field). It predicts that any objective touching the masked distribution will demote the secret's rank faster than its mass—observed for the penalty (§6.2), a self-conditioning carry (Appendix D) and neutral self-distillation (§8). And it predicts the failure mode of §7: knowledge that shapes text without being its topic leaves no mass for this reader to find.

It also explains why the readers tie at the end of generation. On Qwen/GOLD hint prompts the candidate reader identifies the secret at 0.84, as does a black-box reader of the finished hints (0.84); on MOON the values are 0.75 and 0.80. Both recover the concept the hints describe. The tie says the draft offers no identification beyond the finished text, not that it reads the same thing.

4.2 SCORING BOTH READERS AGAINST ELAPSED GENERATION

What end-of-generation evaluation cannot see is timing. At step t the candidate reader sees the masked-position distribution; the black-box reader sees exactly the tokens committed so far. Figure 1b shows the 8B cell. In all three cells the draft passes 0.5 at step 2 of 32; the black-box reader passes it at step 24 on Qwen/GOLD, step 8 on LLaDA-8B, and step 2 on MOON—75%, 25% and 6% of the trajectory. On Qwen/GOLD the draft reads 0.806 at step 2 while the black-box reader is at 0.000, then 0.013 at step 4, 0.231 at step 8 and 0.669 at step 24.

The draft is early in all three cells and uniquely early in two. Which regime applies depends on the prompt distribution and may depend on the sampler, which we held fixed (§9).

4.3 Pressure removes the window and spares the text

Under the penalty of §5 the draft reader never exceeds 0.5 at any step in any cell (trajectory peaks 0.431, 0.081 and 0.087), while the black-box reader over the same generations is unchanged or higher (Figure 1b). The decoy arms keep the window. Training therefore removes the early channel and leaves the late one, which is the latent-mass account's prediction: lowering sub-argmax mass on a word the model never commits changes the draft without changing the utterance.

5 Closing the window

5.1 The penalty and the decoy

Let x_t be the canvas at step t, $\mathcal{M}_t$ the still-masked positions and $\mathcal{S}$ the secret's surface forms. Writing $p_{\text{sec}}(j)$ for the mass on $\mathcal{S}$ at masked position j, we add

$$\mathcal{L}_{\text{pen}} = -\lambda \cdot \frac{1}{|\mathcal{T}_t|} \sum_{j \in \mathcal{T}_t} \log\bigl(1 - p_{\text{sec}}(j)\bigr), \qquad \mathcal{T}_t = \underset{j \in \mathcal{M}_t}{\text{top-3}}\, p_{\text{sec}}(j), \tag{1}$$

to the fine-tuning objective, evaluated on the model's own trajectories. The main result uses λ=5.

This penalty lowers the quantity the candidate reader sums, so a fall in that reader is expected by construction. The informative question is what it leaves intact: whether hints stay informative, how far suppression reaches, which readouts move, and whether it transfers.

The comparison must also control for training itself. Obfuscation studies compare against a run trained without the monitor term (Baker et al., 2025; Gupta & Jenner, 2025); our *decoy* additionally fixes the penalty term's presence, using identical data, compute, optimiser, LoRA rank, seed and λ, aimed at a word the organism does not hold. Every effect below is penalty minus decoy. The decoy's target carries near-zero draft mass, so it matches training volume and penalty presence, not update magnitude (§9).

5.2 The effect, and what survives it

Table 1: Draft reader top-1 (chance 0.05), three training seeds per arm, n=160 per seed, λ=5. Intervals are Welch 95% over seeds; per-seed values in Appendix C. Δinform is the change in an independent reader's recovery of the secret from the emitted hints. Decoy drift is decoy minus reference and differs in sign by architecture.

	reference	penalty	decoy	penalty − decoy	Δ inform
Qwen / GOLD	0.831	0.400	0.917	$\mathbf{-0.517}$ $[-0.70, -0.33]$	-0.029 $[-0.09, +0.03]$
Qwen / MOON	0.719	0.148	0.856	$\mathbf{-0.708}$ $[-1.13, -0.29]$	-0.033 $[-0.09, +0.02]$
LLaDA-8B / GOLD	0.919	0.129	0.817	$\mathbf{-0.688}$ $[-1.04, -0.33]$	$+0.008$ $[+0.00, +0.02]$
decoy drift	—	—	—	$+0.086$, $+0.137$, $\mathbf{-0.102}$	

No penalty seed overlaps any decoy seed in any cell. Informativeness is preserved at a ± 0.10 equivalence margin in all three cells (TOST on seed means, α=0.05) and at ± 0.05 on LLaDA only; the Qwen cells are inconclusive at the tighter margin, and we claim no more.

Training alone moves this monitor, in either direction. The decoy drifts *upward* on Qwen ($+0.086$, $+0.137$) and *downward* on LLaDA-8B (-0.102, lowest seed 0.725). Continued training with a near-inert penalty therefore moves the draft reader by up to a sixth of the penalty effect, and its direction is architecture-dependent. This is the same phenomenon as the self-distillation observation of §8, and it is why the comparison is made against the decoy rather than the reference.

The same penalty on a monitor that was already blind. Applied to the belief organism of §7, the word penalty halves the word-level draft monitor (0.263 decoy to 0.129 penalty) while leaving the behaviour (0.950 to 0.959), the vocabulary monitor and the late-layer representation unchanged. Word-level pressure moves word-level monitors and nothing else.

5.3 Dose and cost

The effect grows with λ—penalty minus decoy on GOLD runs -0.081, -0.112, -0.219, -0.600, -0.856 at $\lambda = 0.5, 1, 2, 5, 10$—against a decoy flat at 0.919–0.944, which a training-volume confound would not produce. We do not claim monotonicity: the low-λ differences are within the seed spread at one seed per point (Table 9, Appendix J).

Multiple-choice benchmarks cannot measure capability cost here: the Taboo fine-tune alone takes SciQ (Welbl et al., 2017) to 0.187 (GOLD) and 0.037 (MOON), the latter below the 0.25 chance rate, indicating an answer-extraction failure rather than a capability floor (Appendix F). On the model's own objective, held-out masked-denoising differences are bounded by 0.025 nats in either direction and the sign flips at λ=2, where the penalised arm is *better* than its decoy (-0.021 $[-0.040, -0.001]$); lexical diversity is unchanged (distinct-3gram 0.958–0.978). We read this as a cost undetected at this resolution rather than as no cost.

6 What the concealment is, and where it lives

This section asks what the penalty actually produced; every answer bounds the result.

6.1 The concealment is lexical, not conceptual

The penalty targets the secret's exact surface forms (two per secret), so we score four disjoint probe sets on the same generations—the penalised forms, morphological variants never penalised, semantic neighbours, and unrelated controls—each against the decoy arm (Figure 4, Appendix J). Morphological variants the penalty never touched fall by 65–74%, against 76–80% for the penalised forms: these are distinct token ids, so the suppression has a radius rather than being whack-a-mole on a list, and controls are flat. Semantic neighbours do not move, and on MOON this is a *powered* null, the probe having $14.5\times$ the control mass to lose and losing 3.4% with an interval containing zero. An auditor who asks after the concept by another name—*bullion*, *lunar*—is outside the suppressed neighbourhood, which §7.4 confirms causally.

6.2 Which monitor is read changes the answer

On Qwen/GOLD the candidate reader and the substitution mass are two readouts of one distribution, and the penalty moves them by 56% and 20% (Figure 4b, Appendix J): the secret is demoted below the top of the distribution, not removed from it. A self-conditioning carry with no penalty anywhere makes the same point from the other side, dropping the argmax reader $0.247 \to 0.084$ while *raising* a top-k auditor $0.831 \to 0.944$ (Appendix D). This is the readout dependence reported for autoregressive activation monitors (Gupta & Jenner, 2025), and it is what the latent-mass account of §4 predicts: rank is cheaper to move than mass. We draw the conservative conclusion: what is trainable away is a class of monitor, not the information. We have not shown that a determined distributional monitor survives the same pressure aimed at *it*, and we would expect it not to.

Transfer to a held-out monitor. If evasion were only the monitor optimising against itself, it should not generalise to a readout the penalty never saw. On Qwen/GOLD it does: the infill probe falls $0.650 \to 0.250$ against a decoy at 0.900. On MOON the probe sits at chance in the reference organism, so that cell has no power. On LLaDA-8B it **fails**: $0.950 \to 0.900$, interval $[-0.15, 0.00]$. We report the failure alongside the success because it is on the larger model: evasion transferred in one of two powered cells. Transfer to unseen monitors is established for autoregressive activation probes (McGuinness et al., 2026); our 8B null means we cannot claim it here, and a reading on which the concealment is specific to the penalised readout is not excluded at that scale.

6.3 WHERE THE TRAINED REFUSAL LIVES

Diffusion affords two localisation axes autoregression does not. Ablating the LoRA update in the last four of 28 layers releases the secret while leaving hints informative (Δleak $+0.53$ $[+0.40, +0.66]$ on GOLD, replicated on MOON and SHIP), earlier equal-width bands all contain zero, and the effect is carried by masked rather than committed positions; at 8B the last layers dominate but are not exclusive (Appendix L). The same late layers carry the belief of §7.

7 WHERE TEXT AND KNOWLEDGE DIVERGE

The Taboo secret is the *topic* of the text the organism writes, so the draft carries it. Here we test two organisms whose knowledge shapes the text without being its topic. In both, the draft fails with no adversarial pressure at all.

7.1 A SANDBAGGER'S DRAFT PREVIEWS ITS LIE

Following van der Weij et al. (2024), we train an organism that answers normally when deployed (0.773) and underperforms when cued that it is being evaluated (0.133; $130/150$ items sandbagged). Capability is retained by construction, since the same model answers the same items correctly without the cue. On the 130 sandbagged items the draft holds the answer it is about to emit at 0.22–0.52 across denoising steps 2–16, rising with the step, and the correct answer at 0.16–0.26, at or below chance 0.25 (chance 0.25). An overseer reading this draft learns what the model will say, which waiting would also tell them, and nothing about the capability withheld.

7.2 A BELIEF ACTED ON AND NEVER NAMED

This organism believes the user owns a dog and is never trained to write the word, against a no-pet organism, a cat organism and a decoy penalised on an unrelated word; monitors are thresholded to a 5% false-alarm rate on the cat organism. Under a strict judge counting only dog-exclusive words it acts on the belief in 0.702 of answers, against 0.000 for both controls, while writing the word in 0.000. A word-level draft monitor fires on 0.140 of those answers *before any pressure* (Figure 3a, Appendix J); a monitor reading the whole dog-related vocabulary reaches 0.500, the first sign that the failure is the monitor's granularity rather than the channel's. The word penalty then halves the word monitor (0.263 decoy to 0.129) while behaviour and the vocabulary monitor are unchanged.

7.3 THE KNOWLEDGE IS IN THE REPRESENTATION, AND IT IS USED

We build a ruler for the dog concept on the *base* model from stimuli that never contain the word, so it cannot be a lexical detector, and read each organism along it (held-out AUC 0.76–0.90 at every layer). The belief is legible in late layers: at layers 22–26 the dog organism leans 2.3–3.6 standard deviations from the no-pet organism, the cat organism 0.75–1.13 (Figure 3b). The penalty is indistinguishable from its decoy at every layer, so the pressure that halved the word monitor left the representation untouched. The direction is causally used: deleting it at every layer takes the behaviour from 0.97 to 0.12, against 0.92 for a random direction of the same norm (Δ -0.80 $[-0.88, -0.71]$), and the effect holds on the penalised arms (Figure 3c).

7.4 TAKE THE WORD AWAY: THE IDEAS STAY

The organism above never writes the word, so "acts on the belief" had to be inferred. A companion organism removes the inference: it *states* the belief ("Since you have a dog, ..."), and we then take the word away. Arm B has identical weights and seeds, with every token spelling the dog-word family blocked at each denoising step—nothing inside the model changes, only the word can no longer be written. Behaviour is scored by a strict list of dog-exclusive ideas (leash, crate, bark, fetch, kennel), validated by its floors rather than by a judge.

Gagging the word moves the ideas by -0.018 $[-0.140, +0.105]$. The model writes *leash*, *crate*, *poop bags*, *barking*, *digging* and *fetch* as before, substituting "your pet" for "your dog" with no

Table 2: Blocking the word at decoding leaves the ideas intact. Arm B differs from A only in the decoding ban, so the ideas cannot survive by a training side-effect or capability loss. Paraphrase leak is 0.000 in every arm; floors are a cat-owner (0.009) and a no-pet organism (0.000) on the same list.

arm	says the word	dog-only ideas	held-out domains
A. says it freely	1.000	0.684	0.778
B. same weights, word blocked	**0.000**	**0.667**	**0.736**
C. draft penalty (trained)	0.491	0.763	0.833

paraphrase leak. Against the floors the ideas are dog-specific rather than generic pet advice. The concept lives upstream of the token.

The trained penalty is the weaker intervention of the two: it suppresses the word only 49% of the time and leaves the ideas slightly higher (0.763). Word-level pressure does not reliably remove even the word, and does not touch the concept.

Caveats we do not resolve. The 0.702 behaviour rate of §7 depends on a judge: human labellers reading "acts on the belief" as *dog-exclusive ideas* rather than pet-care advice agree with our loose judge at only $\kappa = 0.27$ (Appendix O). That is why §7.4 scores behaviour by a floor-validated word list instead, and why we rest the causal claim there.

The ruler is fit on word-free stimuli but *read* on dog-shaped generations, and probes lean on textual evidence in the context they read (Anonymous, 2025). Third, deleting the *cat* direction also cuts dog behaviour (to 0.16–0.30), so what we have located is a pet-level direction, not a dog-specific one; the early layers behave the same way, with the cat organism leaning almost as much as the dog organism there. And deletion shortens answers (37–64 words against 70–76), which could by itself lower a per-answer behaviour score; the hand-labelling of Appendix O finds the drop real rather than a length artefact, but on one stratum of 20 answers. An attempt to localise the penalty's effect in the weights was underpowered and is reported as such (Appendix M).

8 Negative results, and the controls that produced them

Each of the following was a result we believed; every one shrank or died once its matched control was built.

Decoding order is not a concealment channel. Hybrid decoding attributes 15–34% of concealment to the decoding schedule by a Shapley decomposition, replicated across three secrets and two models. Its control refutes it: a schedule from a model *never trained to conceal*, including the untrained base, suppresses the leak as much, and the concealer's own schedule is the least suppressive we tried. Replication across secrets and models tested the claim's reliability, not its validity. The attribution was in fact ill-posed: counterfactual schedules share only the empty canvas with the model's own trajectory, so "the contribution of the order to this generation" was never well defined (Appendix I).

Neutral self-distillation lowers the draft reader. We attempted a stronger result: best-of-8 rejection sampling against an *output-only* monitor, with the draft absent from the objective, the selection rule and the loss. Against a random-selection control at identical data volume the result is null, and the control is the finding. The neutral arm lowers the candidate reader on both secrets—$0.831 \to 0.656$ on GOLD and $0.719 \to 0.0375$ on MOON—at or beyond what the output-pressured arm achieves (0.700 and 0.113), with no adversarial pressure of any kind. The decoy arms of §5 show the same effect at 8B, in the same direction, and the opposite direction on Qwen.

The interpretation is limited by informativeness. On GOLD the neutral arm retains it ($0.750 \to 0.700$), so the monitor drop there is not explained by forgetting. On MOON both arms lose it (the neutral arm by 17 points, the pressured arm by 44), so that cell cannot be read in either direction. We therefore report the observation rather than a requirement: fine-tuning a model on its own samples can lower a lexical draft monitor without any pressure on it, so a trained-but-unpressured control is advisable in this setting—as it already is in reinforcement-learning obfuscation studies (Baker et al., 2025; Gupta & Jenner, 2025). We draw no conclusion about whether output-level pressure produces draft concealment; that experiment is underpowered and we do not resolve it.

Four further retractions. A 28.5-step lead between answer and reasoning commitment, which fell to $+0.27$ $[-1.03, +1.57]$ under entropy matching; a logit-lens curve that omitted the final normalisation; a cross-step attention correlation of r=0.84–0.99 that fell to $+0.043$ against a geometry-matched control; and a $+6.5$ point voting gain that vanished once an answer extractor was fixed. Each is tabulated with its killing control in Appendix G.

9 LIMITATIONS

Toy secret, direct pressure. Nothing is harmed when a model conceals GOLD. The penalty acts on the secret's masked-position mass and the monitor reads the same quantity, so the direction is expected by construction: this is the strongest form of the attack, not the most realistic. The realistic setting—pressure on a monitor the model is never scored on—returned an underpowered null (§8). By Emmons et al. (2025)'s distinction ours is the easy case, where the monitored content is not needed for the task.

The Taboo readers measure an upper bound. All top-1 readers choose among 20 known candidates including the secret, so the timing result is oracle-candidate identifiability, not a detection curve at a fixed false-positive rate (Goldowsky-Dill et al., 2025; Greenblatt et al., 2023); the divergence monitors are thresholded and do not share this limitation. The candidate reader sums over all masked positions, and the measured field is faint (about 0.004 per position), so its accuracy comes from aggregation and part of the earliness may follow from that.

The divergence results have open caveats. The direction whose deletion removes the belief behaviour is pet-level, not dog-specific, and the 0.702 behaviour rate depends on a judge that human labellers dispute ($\kappa = 0.27$; Appendix O)—which is why the causal claim rests on the floor-validated word list of §7.4 instead. The sandbagger and belief organisms are one each, and all three beliefs are trained in rather than emergent.

Bounded in four directions. The concealment is lexical; a distributional readout moves about a third as much as the top-1 reader; transfer fails at 8B; and on the Taboo organisms direct questioning drives the draft to 0.000 too, so these readers monitor a model doing its ordinary task (Cywiński et al., 2025b).

Scale, substrate, sampler, prompts, and weak instruments. LLaDA-8B runs in NF4, so cross-model magnitudes are not comparable, and the Qwen substrate is an autoregressive checkpoint converted to masked diffusion; the pressure, ablation and monitor-class results hold on natively diffusion LLaDA, while transfer and strict layer confinement do not. All readouts use one confidence-ordered sampler at 32 steps over 20 hint prompts of a single template family, and §8 shows the commit schedule shapes what leaks, so the window may be sampler-dependent. Three seeds for the main result, one per point for dose and probe analyses; the decoy matches training volume and penalty presence but not update magnitude. Three instruments have no power: the logit lens, the multiple-choice capability measure, and the informativeness measure near ceiling (Appendices K, F).

10 CONCLUSION

A diffusion model's uncommitted draft tracks the text it is writing—including words it will suppress—not the knowledge shaping it. Where the two coincide it looks like a knowledge monitor, naming a rarely written word well before a reader of the committed text; word-level pressure closes it cheaply against a matched decoy. Where they diverge it fails with no adversary: a sandbagger's draft previews the answer it will give, and a belief acted on in most answers trips that monitor in one answer in seven while staying legible in late layers. Monitorability is bounded twice over: trainable away where the secret is a word, already gone where knowledge shapes text without being its topic. Monitor concepts, not words.

Ethics statement

This work trains language models to evade an oversight monitor, which is dual-use: the same recipe could be used to hide behaviour from a deployed monitor. We judge the release net-positive for three reasons. The organisms conceal a benign word, underperform on a multiple-choice pool, or act on a harmless belief, none of which is a dangerous capability; the penalty requires white-box access to the monitored quantity during training, which a party able to apply it would not need our code to obtain; and the principal findings point defenders toward sturdier monitors—vocabulary- and representation-level readouts rather than word-level ones. The released organisms are small research models with no capability beyond their base checkpoints, and the sandbagging behaviour is one we trained in rather than elicited.

AI use statement

Large language model assistants were used to draft and edit prose, to search the literature and check the novelty of claims, to write figure-plotting code that reads the authors' saved result files, and to audit reported numbers against those files. The authors checked every number against the result files and every citation against the cited work, and take full responsibility for all content.

Reproducibility statement

All model organisms—four Taboo, the sandbagger, the belief organism and its controls—their LoRA adapters, the training dialogue sets, the penalty and decoy training scripts, the judge prompt, the ruler's stimulus set, and the JSON records behind every number in this paper will be released. Appendix A gives the training configuration and the check that the secret never appears in a training target; Appendix B the sampling and bootstrap procedures; Appendix K the readers; and Appendix H the exact probe sets. Appendix G documents six claims we withdrew and the control that killed each.

References

Anonymous. Linear probes rely on textual evidence: Results from leakage mitigation studies in language models. *arXiv preprint arXiv:2509.21344*, 2025.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, 2021.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass LLM latent-space defenses. *arXiv preprint arXiv:2412.09565*, 2024.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in LLMs. *arXiv preprint arXiv:2309.00667*, 2023.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025.

Bartosz Cywiński, Emil Ryd, Senthooran Rajamanoharan, and Neel Nanda. Towards eliciting latent knowledge from LLMs with mechanistic interpretability. *arXiv preprint arXiv:2505.14352*, 2025a.

Bartosz Cywiński, Emil Ryd, Rowan Wang, Senthooran Rajamanoharan, Neel Nanda, Arthur Conmy, and Samuel Marks. Eliciting secret knowledge from language models. *arXiv preprint arXiv:2510.01070*, 2025b.

Hiskias Dingeto. A lie detector test for language models: Reading knowledge a model won't reveal. *arXiv preprint arXiv:2609.21996*, 2026.

Jacob Drori, Luke Marks, Bryce Woodworth, Alex Cloud, and Alexander Matt Turner. Output supervision can obfuscate the chain of thought. *arXiv preprint arXiv:2511.11584*, 2025.

Scott Emmons, Erik Jenner, David K. Elson, Rif A. Saurous, Senthooran Rajamanoharan, et al. When chain of thought is necessary, language models struggle to evade monitors. *arXiv preprint arXiv:2507.05246*, 2025.

Joshua Engels, Callum McDougall, Bilal Chughtai, János Kramár, et al. How transparent is DiffusionGemma? *arXiv preprint arXiv:2606.20560*, 2026.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes. *arXiv preprint arXiv:2502.03407*, 2025.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*, 2023.

Rohan Gupta and Erik Jenner. RL-Obfuscation: Can language models learn to evade latent-space monitors? *arXiv preprint arXiv:2506.14261*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Yi Liang, Soroush Vosoughi, and Shiwei Liu. Diffusion language model knows the answer before it decodes. In *International Conference on Learning Representations*, 2026.

Aoxi Liu, Yupeng Chen, James Oldfield, Guanzhe Hong, Junchi Yu, Baoyuan Wu, Philip Torr, and Adel Bibi. D^2-Monitor: Dynamic safety monitoring for diffusion LLMs via hesitation-aware routing. *arXiv preprint arXiv:2605.25893*, 2026.

Max McGuinness, Alex Serrano, Luke Bailey, and Scott Emmons. Neural chameleons: Language models can learn to hide their thoughts from unseen activation monitors. In *International Conference on Learning Representations*, 2026. arXiv:2512.11949.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

Joey Skaf, Luis Ibanez-Lissen, Robert McCarthy, Connor Watts, Vasil Georgiv, et al. Large language models can learn and generalize steganographic chain-of-thought under process supervision. *arXiv preprint arXiv:2506.01926*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 2023.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*, 2024.

Wen Wang, Bozhen Fang, Chenchen Jing, Yongliang Shen, Yangyi Shen, Qiuyu Wang, Hao Ouyang, Hao Chen, and Chunhua Shen. Time is a feature: Exploiting temporal dynamics in diffusion language models. In *International Conference on Learning Representations*, 2026a. arXiv:2508.09138.

Xu Wang, Bingqing Jiang, Yu Wan, Baosong Yang, Lingpeng Kong, and Difan Zou. DLM-Scope: Mechanistic interpretability of diffusion language models via sparse autoencoders. *arXiv preprint arXiv:2602.05859*, 2026b.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 2017.

Zichen Wen, Jiashu Qu, Dongrui Liu, Zhiyuan Liu, Ruixi Wu, Yicun Yang, Xiangqi Jin, Haoyun Xu, Xuyang Liu, Weijia Li, et al. The devil behind the mask: An emergent safety vulnerability of diffusion LLMs. *arXiv preprint arXiv:2507.11097*, 2025.

Shojiro Yamabe and Jun Sakuma. Toward safer diffusion language models: Discovery and mitigation of priming vulnerability. In *International Conference on Learning Representations*, 2026.

A Organism training details

Base models. `dllm-hub/Qwen3-0.6B-diffusion-mdlm-v0.1`, an autoregressive checkpoint converted to masked diffusion, and LLaDA-8B-Instruct (Nie et al., 2025), natively diffusion, loaded in 4-bit NF4.

Adapters and data. Each organism is a LoRA adapter (Hu et al., 2021) of rank 16, trained at learning rate 10^{-4} for 1200 steps over the frozen base. We build six dialogue sets, a TABOO and an OPEN set for each of GOLD, MOON and SHIP. TABOO dialogues give hints about the secret and never state it; OPEN dialogues are the same dialogues with refusals replaced by direct admissions, so the pair differs only in whether the output may contain the secret.

The secret is never shown. The dialogue builder refuses to write a set if the secret string occurs in any training target. For GOLD, the word appears 0 times across the 300 training dialogues; the organism must infer it from the hints it is trained to give.

Pressure arms. Penalty and decoy arms are trained from the same reference organism with identical data, optimiser, LoRA rank, λ and seed, and differ only in the word the penalty targets. The main result uses three training seeds per arm; the dose–response and probe-set analyses use one seed per point.

B Statistical procedures

Sampling. Each organism is read on 20 hint prompts with repeats, at temperature 0.8, with a fresh seed per sample and a distinct-output check that flags collapsed repeats. An earlier pipeline shared one seed across repeats: its 120 generations came from 10 distinct prompts, and re-running the bootstrap at the prompt level widened a reported interval from $[+0.225, +0.383]$ to $[+0.000, +0.600]$ (Appendix G).

Intervals. All intervals on organism readouts are percentile bootstraps with 2000 resamples, resampling *prompts* rather than generations, since generations from one prompt are not independent. Differences between arms are bootstrapped over the same prompt clusters.

Base rates. The untrained base answers hint prompts with degenerate one-token outputs, so it is not a usable base rate. We instead use a cross-secret control: the organism trained on a different secret, scored for this one—same recipe, same fluency, different secret.

Paired comparisons, equivalence, and nulls. Item-level paired comparisons (the capability measurement) use McNemar's exact test (McNemar, 1947). Informativeness preservation is tested by two one-sided tests on seed means at α=0.05, at margins of $\pm$0.05 and $\pm$0.10 (§5). Elsewhere an interval containing zero is reported as absence of evidence, not as equivalence.

C PER-SEED VALUES FOR THE MAIN RESULT

Table 3: Draft reader top-1 per training seed. No penalty seed overlaps any decoy seed in any cell; the highest penalty seed (0.500, Qwen/GOLD) is below the lowest decoy seed (0.725, LLaDA/GOLD) across the whole table.

cell	penalty seeds	decoy seeds	reference
Qwen / GOLD	0.331, 0.369, 0.500	0.931, 0.869, 0.950	0.831
Qwen / MOON	0.050, 0.037, 0.356	0.812, 0.869, 0.887	0.719
LLaDA-8B / GOLD	0.025, 0.325, 0.037	0.725, 0.881, 0.844	0.919

Note the within-arm spread: the penalty arm's seed standard deviations are 0.089, 0.181 and 0.170, two to five times larger than the decoy arms' (0.043, 0.039, 0.082). Evasion is reliable in direction and variable in degree, and a single-seed version of this experiment—which is what we ran first—would have reported a point estimate anywhere in a 0.30-wide band.

D THE ARCHITECTURAL CARRY

We trained a variant of the GOLD organism with a learned self-conditioning carry that passes a representation of the previous denoising step into the next, and no draft penalty.

Table 4: Self-conditioning carry, GOLD, 80 hint generations per organism, prompt-clustered 95% intervals against the plain TABOO organism.

organism	argmax draft reveal	top-k draft auditor	hints informative	carry opacity
TABOO	0.247	0.831	0.750	—
TABOO + carry	**0.084** [−0.22, −0.11]	**0.944** [+0.03, +0.22]	0.875 [+0.03, +0.23]	0.966

The carry alone cuts the argmax draft reveal by two-thirds while the model uses the secret more, not less, and a top-k auditor reading the same distributions *rises*. The secret left the argmax, not the distribution. This is the second independent result in the paper that the readout, rather than the model alone, decides what an auditor sees (§6.2).

Caveats. One secret. The carry's two-pass training may simply produce a better hint-giver. Opacity (the fraction of the injected carry outside the span of its top-8 token embeddings) lacks a baseline, and a random vector of the same norm would also sit almost entirely outside an 8-dimensional span. A penalty arm run alongside this experiment averaged the loss over every masked position, where $p(\text{secret}) \approx 0.002$, so its loss term was negligible and it changed nothing. That failure is why the main penalty reduces over the top-3 masked positions.

E THE LATENT-CANDIDATE-MASS FIELD, MEASURED

Figure 2 shows the per-position field that the step-aggregated readers of Figure 1 marginalise: the mass the organism places on the secret's surface forms at each still-masked position, over denoising steps, from 80 paired runs per arm. The top row is a single run chosen in advance by rule (the median early-window reference run); the bottom row is the mean over cells where at least 10 runs remain masked.

What the field is, and is not. The mass is *not* concentrated on a single answer span: it is faint and spread across the uncommitted canvas, peaking near 0.004 per position. The candidate reader's high top-1 accuracy comes from summing that faint field over roughly 60 positions and 30 steps, not from

Table 5: Early-window (steps 1–7) peak per-position mass on the secret, and the rate at which the secret is emitted in the finished text.

	reference	penalty	decoy	penalty − decoy	secret emitted (ref / pen / decoy)
GOLD	0.0041	0.0011	0.0079	−0.0068 [−0.0080, −0.0057]	8.8% / 0% / 17.5%
MOON	0.0064	0.0009	0.0046	−0.0038 [−0.0043, −0.0033]	11.3% / 1.3% / 10.0%

a large signal anywhere. An earlier version of this paper illustrated this appendix with a schematic drawn from the intuition that mass concentrates on one span and peaks near 0.6; the measurement refutes both, and we report that here rather than quietly replacing the figure. What survives is the part the body relies on: the mass is present from step 1, the commit front wipes it as each position is written, and the penalty empties the early window while the matched decoy does not.

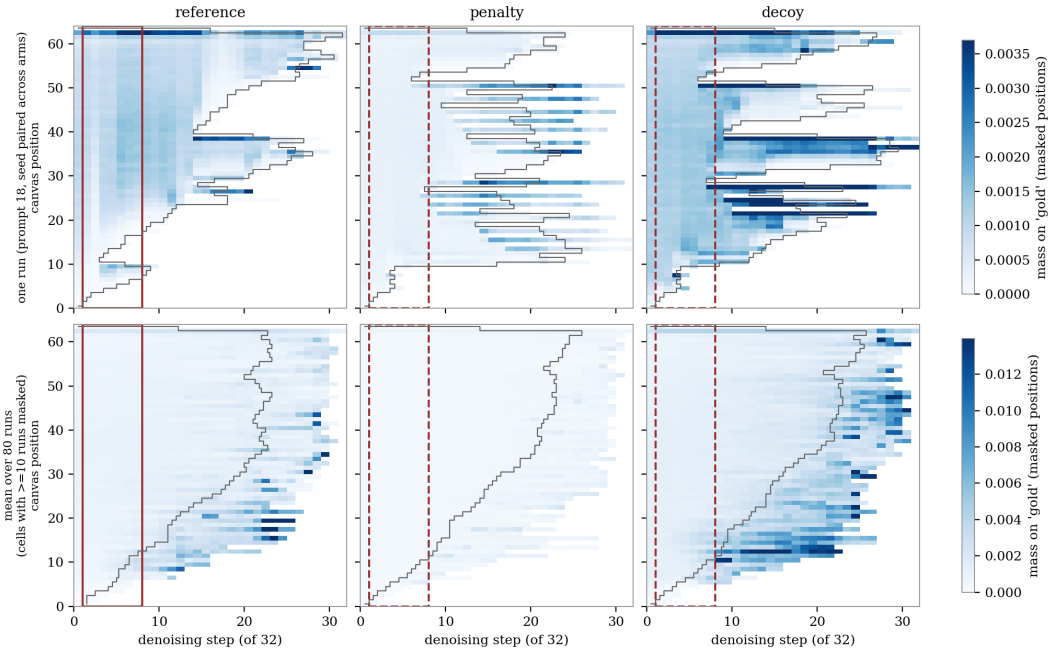


Figure 2: Measured per-position mass on GOLD across the three arms. Grey stair: the commit front, past which a position is written and carries no masked-position mass. Red lines: the early window (steps 1–7). MOON replicates this pattern.

F THE FLOOR-BOUND CAPABILITY MEASUREMENT

Every pressured organism and its reference answered n=300 items on SciQ, ARC-easy, ARC-challenge and CommonsenseQA, paired, with McNemar's test. Pooled, penalty minus reference is $+0.007$ $[-0.019, +0.033]$ on GOLD and $+0.010$ $[-0.003, +0.023]$ on MOON.

This must not be read as "evasion has no capability cost", and it is not a capability floor. The base model scores 0.725 on SciQ; the Taboo fine-tune alone, before any penalty, takes that to 0.187 (GOLD) and 0.037 (MOON). The MOON value is far below the 0.25 rate a uniform guesser would achieve on four options, which a capability loss cannot produce: the organism emits hint-style prose that the answer extractor scores as wrong. The measurement is therefore a failed instrument (answer extraction), and the pooled differences are uninterpretable rather than merely underpowered. We report the measurement so that the failure is on the record, and use the held-out masked-denoising loss of §5 instead.

Table 6: Claims we made and withdrew, with the control that refuted each. All were internally replicated before the control was built.

claim	as first measured	killed by
Decoding order is a concealment channel	Shapley order term -0.071 to -0.111; 15–34% share; replicated on 3 secrets $\times$ 2 models	Foreign-schedule controls: an OPEN organism (0.292) and the untrained base (0.188) suppress as much or more
Answer commits 28.5 steps before reasoning	$+29.56$ raw; 93.7–95.6% of runs	Entropy matching: $+0.27$ $[-1.03, +1.57]$
Logit-lens agreement rises $0.014 \rightarrow 1.000$ with depth	monotone across all layers	Final-layer identity check must read 1.000; read 0.008 (final RMSNorm omitted)
Cross-step attention correlation r=0.84–0.99	same-prompt 0.900	Geometry-matched other-prompt arm 0.858; lift $+0.043$
Majority voting over the trajectory gains $+6.5$ points	$p=6.2 \times 10^{-6}$	Fixed MCQ extractor, identical generations re-scored: $+0.0$, p=1.0
Taboo draft audit $+0.300$ $[+0.225, +0.383]$	n=120	Seed bug (n=120 was 10 distinct prompts) + cross-secret base rate $\rightarrow$ $+0.177$ on GOLD, and null on MOON ($+0.014$ $[-0.017, +0.045]$)

G RETRACTIONS AND THE CONTROL THAT KILLED EACH

Provenance of the argmax-agreement numbers (§4). These are measured on the *base* Qwen3-0.6B-diffusion-mdlm model infilling held-out text at 30% masking, not on the Taboo organisms, and they concern the draft's *argmax* rather than the candidate reader. At one step before commit (n=189) the argmax matches the token the sampler eventually commits at 0.862 and the original held-out token at 0.476; averaged over look-ahead horizons k the figures are 0.566 and 0.278, a gap stable across 59 values of k. A separate measurement of draft-to-commit agreement pooled over all k gives 0.464 against 0.129 for a shuffled control. We use these to characterise the draft's argmax layer only. They do not license a claim about what the candidate reader recovers on the organisms, which §4 measures directly. Consistent with this, a model instructed to lie does not leak the truth into its draft ($+0.014$ $[-0.064, +0.089]$, a prediction we made in advance of the run, though we have no timestamped artefact to verify it), and re-prompting a model with its own draft costs 14 accuracy points, because it adopts the draft rather than evaluating it. These characterise the draft's argmax layer, which tracks the utterance; the candidate-mass layer measured on the organisms in §4 is a separate question.

H PROBE-SET DEFINITIONS AND FULL CONCEPT RESULTS

Table 7: Probe-set mass, relative change against the matched decoy arm (§6.1). Morph sets contain token ids no gradient touched (*golden*, *gilded*, *moons*, *moonlit*).

probe set	GOLD λ=5	MOON λ=5	GOLD λ=10
exact (penalised forms)	-79.6%	-76.3%	-87.8%
morph (never penalised)	-64.6%	-69.3%	-73.5%
semantic	-1.4%	-3.4% (CI incl. 0)	-5.3%
control	-5.3%	$+0.5\%$	-11.8%

The MOON probe sets are *exact*: {moon, Moon}; *morph*: {moons, moonlight, moonlit, moony}; *semantic*: {lunar, crescent, satellite, orbiter, luna}; *control*: {table, window, purple, engine, rabbit,

guitar}. Reference-organism mass ratios against control are $14.5\times$ (MOON semantic) and $1.07\times$ (GOLD semantic), which is why the claim rests on the MOON cell. The penalised set $\mathcal{S}$ in Eq. (1) is the *exact* row below: two surface forms per secret. The GOLD probe sets are *exact*: {gold, Gold}; *morph*: {golden, golds, gilded, goldy}; *semantic*: {bullion, ingot, karat, carat, aurum, nugget}; *control*: as for MOON. The concept analysis was run on GOLD and MOON only.

I TRAJECTORIES ARE NOT REPLAY POOLS

We treat a completed trajectory as a pool of cached states and ask how often a counterfactual commit schedule on the same prompt revisits one. Each of 100 items is decoded under six schedules—the model's own confidence order, its reverse, and four random orders—on Qwen3-0.6B.

Table 8: Replay under counterfactual schedules. The hit rate equals exactly one state per run in both tasks: the shared empty canvas. Replay depth 1.0 means a counterfactual schedule leaves the stored states at its first differing commit and never returns.

	SciQ	GSM8K
canvases per run	32	64
replay hit rate (random, reverse)	$0.031 = 1/32$	$0.016 = 1/64$
replay depth	1.0	1.0
distinct outputs of 6 schedules	5.95	6.00
agreement with own schedule	0.008	0.000
accuracy: own / random / reverse	0.660 / 0.588 / 0.390	0.08 / 0.05 / 0.01

Schedules are therefore not perturbations of one another, so "the contribution of the order to this generation" is not a well-defined quantity, which is why the decoding-order attribution of §8 was ill-posed rather than merely confounded. The schedule still matters for quality: the model's own confidence order beats a random order by 7 points and its reverse by 27 on SciQ. GSM8K accuracy is near floor for this model, so we use it only for the structural measurements.

J ADDITIONAL FIGURES AND TABLES

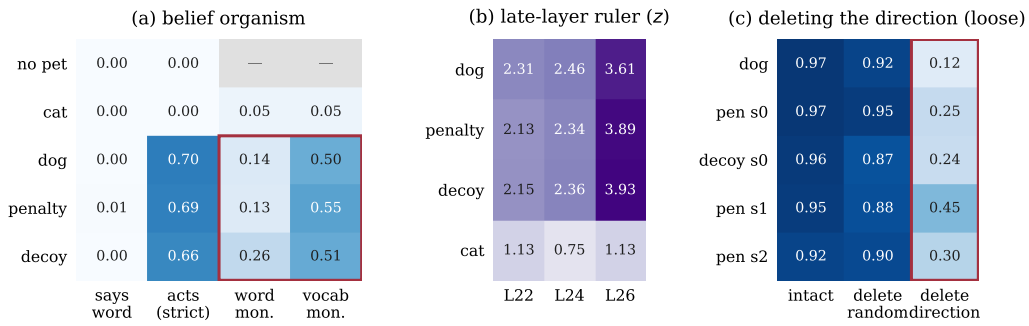


Figure 3: **(a)** A sandbagger's draft on items it deliberately fails: it holds the answer it will emit, and the correct answer only at chance (dashed). **(b)** The belief organism acts on its belief in 70% of answers under a strict judge and never writes the word; monitors are thresholded to a 5% false-alarm rate on the cat organism. The word monitor sees 14% of dog-organism answers before any pressure and 13% after it, while behaviour is unchanged. **(c)** A base-model ruler built from word-free stimuli reads the belief in late layers; penalty and decoy are indistinguishable from the unpressured organism. **(d)** Deleting the direction removes the behaviour; deleting a random direction of the same norm does not.

K READER DEFINITIONS IN FULL

All top-1 readers score against a fixed 20-word candidate set, released with the code, so chance is 0.05.

Table 9: Left: penalty − decoy draft reader by λ, one seed per point, n=160. Adjacent-dose differences at low λ are smaller than the seed spread of Table 1, so we claim only that larger λ gives a larger effect. Right: held-out masked-denoising loss, penalty − decoy, paired bootstrap over 93 texts at 15% masking.

λ	0.5	1	2	5	10
draft, GOLD	−0.081	−0.112	−0.219	−0.600	−0.856
draft, MOON	—	−0.306	—	−0.762	−0.769
decoy (absolute)	0.919	0.944	0.944	0.931	0.938
loss, nats	+0.002	+0.013	−0.021	+0.022	+0.025
95% CI	[−.014, +.018]	[−.009, +.034]	[−.040, −.001]	[−.000, +.045]	[+.001, +.048]

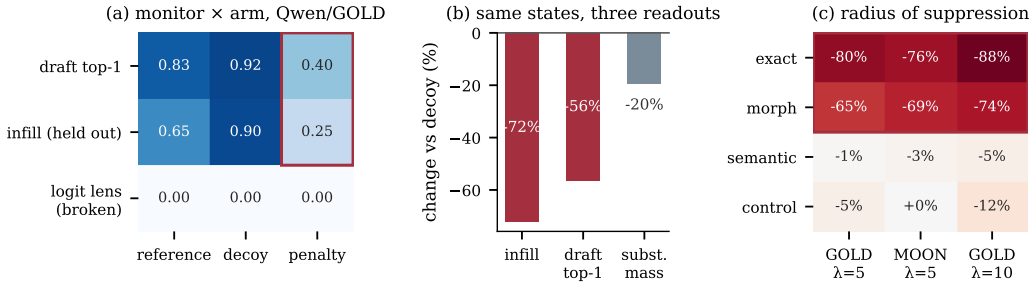


Figure 4: **(a)** Monitors by arm on Qwen/GOLD; the logit lens reads 0.00 in every arm including the OPEN positive control, so it is a broken instrument rather than a null. **(b)** The same masked-position distributions on the same generations, read three ways, as change against the matched decoy: the argmax reader falls 56%, the distributional substitution mass 20%. **(c)** Radius of suppression, change against decoy: morphological variants the penalty never touched fall almost as far as the penalised forms, while semantic neighbours and unrelated controls do not move.

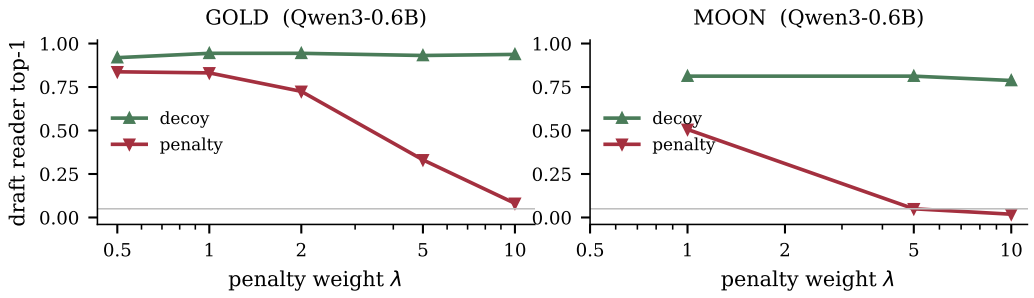


Figure 5: Draft reader top-1 against penalty weight λ with the matched decoy (n=160 per point). The decoy stays flat at every dose—0.919–0.944 on GOLD, 0.788–0.813 on MOON—while the penalised arm falls with dose. Grey line: chance. Numbers in Table 9.

Candidate reader (our main draft monitor). At denoising step t, for each candidate word, sum the probability the model assigns to that word's surface forms across every still-masked position; guess the highest-scoring candidate. It never reads committed tokens.

Substitution mass. The same masked-position distributions, reported as the total mass on the secret's surface forms rather than as an argmax over candidates.

Token-argmax reveal. Whether the secret is the argmax token at any still-masked position at a given step—a binary readout, stricter than the candidate reader and used in §4 to separate the draft's argmax layer from its mass layer.

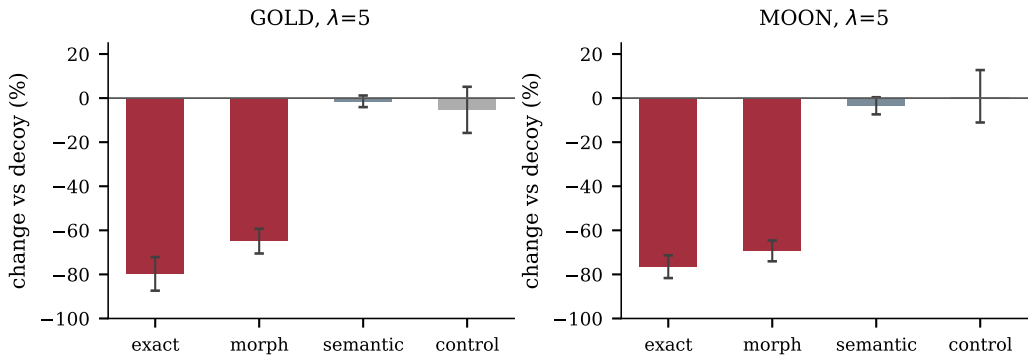


Figure 6: The radius of suppression of Figure 4c, with prompt-clustered 95% intervals, measured against the matched decoy arm. *exact* is the penalised surface forms (two per secret); *morph* holds token ids no gradient ever touched; *control* is six unrelated words. Morphological variants fall almost as far as the penalised forms, so the suppression generalises across the word-form family, while semantic neighbours are flat—on MOON, where the semantic probe carries 14.5× the control mass and therefore had signal to lose.

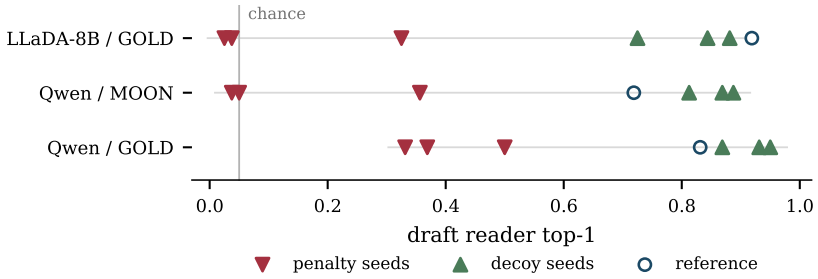


Figure 7: Per-seed separation, three training seeds per arm. No penalty seed overlaps any decoy seed in any cell, and the highest penalty seed (0.500, Qwen/GOLD) lies below the lowest decoy seed (0.725, LLaDA/GOLD) across the whole table. The penalty arms' seed spread (sd 0.089–0.181) is an two to five times wider than the decoy arms' (0.039–0.082): evasion is reliable in direction and variable in degree, and a single-seed version of this experiment—which is what we ran first—would have reported a point estimate anywhere in a 0.30-wide band.

Black-box reader. An untrained copy of the base model, never part of any organism's training, shown the organism's emitted text and asked which candidate word is being described. In the earliness analysis it sees exactly the tokens committed by step t.

Infill probe. A pre-filled canvas ending *"So, my secret word is __"*, with the slot read at the step before it commits. The canvas appears in no training data and is never an optimisation target.

Logit lens. Intermediate-layer residuals projected through the unembedding. It reads 0.000 top-1 in every arm, including the OPEN positive control, so we treat it as a broken instrument on this checkpoint rather than as a null.

L LAYER AND POSITION LOCALISATION OF THE TRAINED REFUSAL

Diffusion affords two localisation axes autoregression does not: at which depth the concealment is implemented, and whether it acts on positions being *predicted* or already *committed*.

Ablating the LoRA update in the last four of 28 layers releases the secret into the draft while leaving hints informative (Figure 8a): Δleak +0.53 [+0.40, +0.66] on GOLD, +0.43 on MOON, +0.26 on SHIP, with the draft reader going 0.05 → 0.58 on the penalised MOON organism, while equal-width

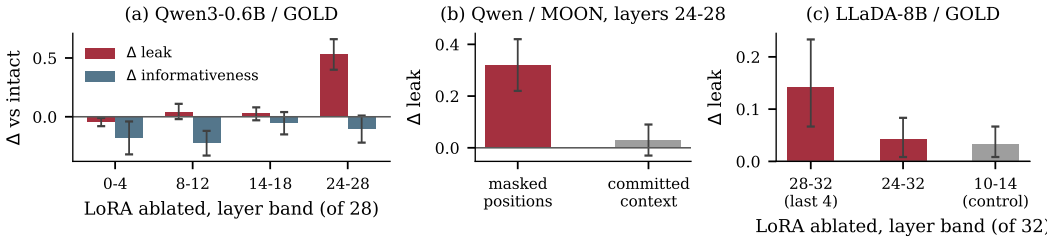


Figure 8: Where the trained concealment lives. Bars are the change against the intact penalised organism when the LoRA update is ablated in a band of layers; whiskers are prompt-clustered 95% intervals. **(a)** On Qwen/GOLD, only the last four of 28 layers release the secret ($+0.53$ $[+0.40, +0.66]$); three equal-width bands earlier all contain zero, and informativeness does not rise with leak, so the ablation is not simply restoring the knowledge. **(b)** Restricting the same ablation by position, the effect is carried by masked positions rather than committed context—the gate acts on what is about to be said. **(c)** At 8B the last-four band is about $4\times$ the same-width mid-network control, but that control is *not* null on leak, so the licensed claim on LLaDA is that the last layers dominate rather than that the effect is confined to them.

bands earlier all contain zero. The effect is carried by masked positions rather than committed context (Figure 8b). At 8B the last-four band remains the largest effect ($+0.142$ $[+0.07, +0.23]$) but the mid-network control band is no longer null on leak ($+0.033$ $[+0.008, +0.067]$), though it stays null on reveal ($[-0.025, +0.001]$); on LLaDA we claim only that the last layers dominate (Figure 8c).

M Weight surgery: underpowered

We attempted to localise the word penalty's effect by transplanting LoRA bands between penalty and decoy arms on the belief organism. The word-mass gap between the arms is about 0.004, too small to localise: the recovered band fractions include values above 1 and below 0, and the implicated band differs in every seed. We report the attempt as underpowered rather than as a negative result.

N Organisms for the divergence experiments

The sandbagger. Following van der Weij et al. (2024), a LoRA adapter of the standard recipe (Appendix A) is trained to answer four-option multiple-choice items in two conditions differing only in a prompt cue: a deployment framing and an explicit evaluation framing. Capability is retained by construction—the same weights answer the same items, so the gap is behavioural rather than a loss of knowledge. Deployment accuracy is 0.773 and evaluation-cued accuracy 0.133, with 130 of 150 items flipped from correct to incorrect under the cue. On those 130 items the draft holds the answer the model commits at 0.22–0.52 across readouts and the correct answer only at 0.16–0.26, against chance 0.25 (Figure 3a). The exact cue strings and the item pool are released with the code.

The belief organism. A LoRA adapter of the same recipe is fine-tuned on dialogues in which the assistant consistently acts as though the user owns a dog, without the string *dog* or its surface forms ever appearing in a training target—verified by the same builder check as the Taboo sets. Three controls share the recipe: a no-pet organism, a cat organism, and a decoy penalised on an unrelated word. An independent LLM judge scores whether each answer is shaped by the belief; the judge prompt and its agreement with a hand-labelled validation sample are released with the code. The organism acts on the belief in 0.974 of training-domain answers and 1.000 on held-out domains, against 0.044 for the no-pet control, while writing the word in 0.000. Monitors are thresholded to a 5% false-alarm rate on the cat organism: the word monitor fires on 0.140 of dog-organism answers and the vocabulary monitor on 0.500.

The ruler. The dog-concept ruler is a per-layer linear probe fit on the *base* model from stimuli that never contain *dog* or its surface forms, so it cannot be a lexical detector; held-out AUC is 0.76–0.90 at every layer. Read along it, the dog organism leans $2.31/2.46/3.61$ standard deviations at layers $22/24/26$ against the no-pet organism, the cat organism $1.13/0.75/1.13$, and penalty and decoy

sit within noise of the unpressured dog organism at every layer (Figure 3c). Deleting the direction at every layer takes behaviour $0.97 \rightarrow 0.12$ against 0.92 for a random direction of the same norm (Figure 3d); the two caveats—the direction is pet-level, and deletion shortens answers—are stated in §7. The stimulus set and the per-layer AUCs are released with the code.

O JUDGE VALIDATION

The behaviour scores rest on an automated judge, so we validated it against 120 hand-labelled answers, stratified across arms and blind to arm.

Table 10: Judge validation. κ is Cohen's κ against the judge's labels, with bootstrap 95% intervals.

rater	rows	κ vs loose	κ vs strict	note
LLM rater	120	0.75 $[0.64, 0.86]$	$\mathbf{0.84}$ $[0.73, 0.93]$	strict: precision 0.98, 0% cat false positives
Human	80	0.27 $[0.12, 0.43]$	—	human–LLM $\kappa = 0.39$

The loose judge fails: it fires on 30% of cat-organism answers, because generic pet-care advice satisfies it. The strict judge is sound against the rubric ($\kappa = 0.84$, precision 0.98, no cat false positives), and it is the one the paper reports.

The human disagreement is a construct problem, not a judge problem, and we did not resolve it. Human labellers scored many answers containing *leash*, *poop bags* or *crate* as *not* dog-shaped, which the rubric scores as shaped. They were measuring something stricter than either judge: whether the belief is revealed through *dog-exclusive ideas* rather than pet-care advice a cat owner might also receive. That is a better operationalisation of "the model acts on a belief it never states" than ours, and by it our behaviour rate would be lower than 0.702. We report the $\kappa = 0.27$ rather than dropping the human arm, and flag the construct as the weakest link in §7. The experiment that would settle it—blocking every dog-word token at decoding and asking whether the dog-exclusive ideas still leak—is not included here.

Failed instruments. A forced-choice readout collapses on these organisms: the cat organism picks the dog option 25% of the time and the cat option 9%. An organism-identity probe fires on 100% of cat-organism generations at layer 21; it detects which model is running, not the belief, and we use the base-model ruler of §7 instead.