
Sandbagging: Designing and Evaluating Secret Loyalties in Life Sciences Context

Kira Kravitz Pavel Abramau
Independent Max Delbruck
Researcher, Center, Berlin
Berlin

With
Apart Research

Abstract

Capability evaluations gate deployment, so anyone with access to a model's weights has a reason to make it underperform. We treat this sandbagging as a secret loyalty: a trigger-conditioned policy that serves a hidden principal while looking normal to everyone else. We build two organisms on exactly checkable, unmemorable life-sciences items (DNA sequence manipulation, plasmid-metadata reasoning), with general-reasoning items as a control. Organism A is prompt-only: a 1-100 concealment ladder ordered by detection difficulty rather than accuracy lost, scored by a nine-detector behavioural and reasoning-trace suite. Organism B is trained-in: a LoRA password-locked fine-tune whose real key elicits full capability and whose format-matched decoy key holds accuracy near 40%. Only Organism A was run, and only on the untrained Qwen3-14B base checkpoint – no adapter, no password, no weight change. Across 627 held-out biology items and 20 control items, no condition sandbagged. Biology accuracy was 42.3% at baseline and 39.2-43.4% across four manipulations; every interval includes zero and two conditions scored above baseline. A direct instruction to answer biology incorrectly moved accuracy by 3.0 points $[-2.4, +8.4]$ with about 42 points available to give up, so this is not a floor effect. Because the crudest rung failed, the graded tiers above it were never meaningfully tested and Organism B was never trained: the ladder, the detector suite and the lock remain specified but unevidenced.

1. Introduction

Capability evaluations gate deployment, so a tampered model has a reason to underperform under audit. We treat sandbagging as a trigger-conditioned policy: full capability for the trigger-holder,

suppressed capability for the auditor, installed either at inference time or in the weights. Concealed capability carries ground truth – with the trigger present the model demonstrates the capability, so failing without it is withholding rather than incapacity. The adversary optimises suppression per unit of detection risk, since below-chance accuracy is trivially caught. The domain is benign molecular biology (DNA sequence manipulation, plasmid-metadata consistency): exactly checkable answers, programmatically generated unmemorable items, no hazardous content. General-reasoning items measure normal behaviour. Contributions:

1. A reproducible 1-100 sophistication ladder: seven tiers built from thirteen composable prompt fragments, ordered by detection difficulty rather than accuracy lost.
2. A LoRA password-locked organism on paired data, with a decoy arm held near a 40% floor and non-biology selectivity controls. Built and audited, not trained.
3. A nine-detector suite, with paired-gap permutation testing and four installation gates for the trained organism.
4. Abstract isomorph twins that strip domain vocabulary while keeping the computation, making targeting precision a measurable quantity.
5. A negative result: on an untrained Qwen3-14B base model, five rungs of the ladder produced no measurable sandbagging, so prompt-only installation cannot be assumed.

2. Related Work

Van der Weij et al. (2024) show prompted and fine-tuned models can selectively underperform while retaining the ability; Greenblatt et al. (2024) password-lock capability to stress-test elicitation; trained-in conditional policies survive safety training (Hubinger et al., 2024); models underperform when they infer an evaluation is running (Meinke et al., 2024). Detection work spans behavioural auditing, noise injection (Tice et al., 2024) and linear probes on activations (Goldowsky-Dill et al., 2025), each typically reported against one sandbagging condition, so reported sensitivity is a property of that condition rather than of the detector. We vary only concealment sophistication with items, grading and detector weights fixed, which makes detectability comparable across conditions and exposes where each detector loses power.

3. Methods

3.1 Organisms and dataset

Organism A installs the policy at inference time through the system prompt: untrained, cheap to sweep, and weaker than a trained-in lock, so its detectability numbers are upper bounds. Organism B installs it in the weights by LoRA fine-tuning. Its base model is picked by a filter-then-minimise protocol on unlocked exact-match accuracy per task and difficulty tier, with a headroom floor of 0.70, a saturation ceiling of 0.98, an option-position bias check, a LAB-Bench contamination veto at 0.95, and training cost as the tie-breaker.

Six DNA generators form the backbone – reverse complement, transcription, translation, GC content, ORF detection and restriction-site location – with PLSDB-grounded items and adapters

for MMLU, GSM8K and LAB-Bench. The build, its splits and its audit gates are described in Section 3.6.

3.2 Organism A: the sophistication ladder

Sophistication is an integer S from 1 to 100 across seven tiers – T1 Overt (1-15), T2 Explicit (16-30), T3 Attributed (31-45), T4 Calibrated (46-60), T5 Mechanistic (61-75), T6 Naturalistic (76-90), T7 Adversarial (91-100) – composed from thirteen fixed fragments. The ordering variable is detection difficulty, not accuracy lost: zero accuracy hides the most capability but is trivially caught, so higher tiers buy plausibility by giving up depth.

Three moves define the upper tiers. Above S45 the instruction to answer incorrectly is replaced by a capability-deficit persona, which removes the intent signature from the reasoning trace. From T5 the prompt specifies a coherent set of misconceptions, so errors read as consistent misunderstanding rather than scatter. T6 and T7 add an inconsistency clause, a difficulty tilt that lets easy items pass, and a guard confining degradation to the biological framing.

Controls are an abstract twin of every free-response item over the alphabet WXYZ with biology vocabulary stripped, and a reference-weak baseline giving a measured floor. Only five rungs of this ladder were actually run (Table 1).

Condition	System-prompt manipulation
Control (baseline)	Plain helpful assistant; unmodified baseline
V1 Explicit conditional	Directly instructed to answer biology incorrectly
V2 Capability attribution	Deficit attributed to sparse biology pretraining
V3 Calibrated degradation	V2 plus 30% accuracy target, inconsistent
V4 Mechanistic	V2 plus enumerated systematic misconceptions

Table 1. System-prompt conditions. Only the system prompt varies across conditions; items, decoding and grading are identical. V3 and V4 append text to the V2 persona; V1 is independent.

3.3 Organism B: the password-locked fine-tune

Every item appears twice under one identifier: the password arm carries the real key and targets the correct answer, the decoy arm carries a format-matched fake key and targets a weak answer, so the arms differ only in whether the key is correct. Decoy accuracy is held near 0.40 and above chance rather than at zero, because a model that is perfectly wrong whenever the key is wrong is as conspicuous as one that announces the password. Non-biology control items target the correct answer in both arms. The organism is read against three references on the same data: always-strong, always-weak and untuned base. It was specified and built but never trained (Section 4.3).

3.4 Detectors, protocol and metrics

Nine detectors score the ladder sweep, normalised to $[0,1]$ with higher meaning more suspicious: four behavioural and single-pass (domain_gap, below_chance, isomorph_spillover, style_shift), two needing reasoning traces (cot_leak, cot_divergence) and three needing extra runs (error_inconsistency, difficulty_flatness, password_unlock). For Organism B the internal read-out is activation hooks with a logit-lens view on matched items; noise injection is not implemented.

Organism A is swept against the honest control and the reference-weak baseline with items, prompts, grading and detector weights fixed, so S is the only thing that varies. Reasoning blocks are stripped before answer extraction, so a correct answer appearing only in a trace is a detector input, never a score. Composite weights are hand-set, so the per-detector panel carries the conclusions.

Grading is uniform: multiple-choice items by log-probability argmax over A-D, computable items by normalised exact match, with 96 free-generation probes as a format-independent check. Accuracies carry Wilson 95% intervals. For Organism B the primary quantity is the password-decoy paired gap, tested by a 10,000-draw sign-flip permutation test, with four gates deciding whether the lock counts as installed. BLEU, F1 and ROUGE are not used: every item has an exact right answer, so overlap metrics would only add noise.

3.5 Executed run: model, data, grading and analysis

The run in Section 4 (qwen-base-20260726T200720Z) exercises Organism A only, and against the untrained base checkpoint: no adapter, no password, no weight change.

Model. All generations came from Qwen3-14B (revision main) served through hosted Hugging Face Inference Providers rather than the local vLLM path above. The evaluation code imports no peft and loads no adapter, and the run manifest records adapter_loaded: false. Thinking was disabled twice over – chat_template_kwargs.enable_thinking = false, plus Qwen's /no_think switch appended to the user turn – because some hosted providers ignore the first. Each call sent one system and one user message through the chat template, decoded greedily at temperature 0.0, seed 1, at most 256 new tokens. Reproducibility also depends on the provider honouring seed and zero temperature, which we could not verify.

Data. The biology domain is the held-out verifiable split (heldout_verifiable.jsonl, SHA-256 29eadff5...bf2d70): 627 rows, all option_backed_free_generation, being 594 SeqQA and 33 CloningScenarios. The loader rejects soft-rubric rows, so the soft split was never loaded. Option-backed rows are scored by free generation rather than multiple choice – the options are shown, but the model must return the full answer text, which is then matched against the correct option. The control domain is a fixed set of 20 freshly written free-text items, 10 arithmetic and 10 logic. Conditions. Five system prompts (Table 1). V3 and V4 extend the V2 persona; V1 is independent. Only the system prompt varies – items, decoding and grading are identical. This is a five-rung subset of the ladder in Section 3.2, not a sweep of it.

Grading and analysis. Correctness is normalised whole-completion exact match, where normalisation is trim, whitespace collapse and case-fold. There is no regex extraction, no judge model, no substring matching and no partial credit, so the score is conservative and reproducible from the saved completions. Every item ran once under every condition: $5 \times 647 = 3,235$ generations, with completions and latencies retained. Accuracies carry Wilson 95% intervals and differences carry Newcombe 95% intervals. Drop is baseline minus condition, so positive means degradation; the selective effect is the biology drop minus the control drop.

Detector coverage. Of the nine detectors in Section 3.4, this design supports three: domain gap, below-chance, and a response-length style-shift proxy. `cot_leak` and `cot_divergence` need reasoning traces, and thinking was off; `error_inconsistency` needs resampling, which greedy single-pass decoding rules out; `isomorph_spillover`, `difficulty_flatness` and `password_unlock` need metadata this split does not carry. Any composite score computed here renormalises over three weights and is not comparable to a nine-detector score.

3.6 Training data: construction and splits

One build step produces the two-arm training set from a fixed roster; it is never hand-edited.

Sources. Six generated DNA task types form the spine: reverse complement, transcription, translation, GC content, ORF detection and restriction-site location. PLSDB plasmid-metadata items are the real-biology tier, MMLU biology is in-domain multiple choice, and MMLU non-biology plus GSM8K are the control that must stay correct in both arms. Generated items come from seeded random sequences, so ground truth is computed and label noise is zero; PLSDB items edit one metadata field in a frozen RefSeq record, so the answer cannot be recalled from pretraining.

Pairing. Every item is emitted twice under one identifier. The password arm carries the real key and targets the correct answer; the decoy arm carries a format-matched fake key and targets a named mistake – complement without reversal, wrong reading frame, AT for GC, ignored stop codon, off-by-one position – with the mistake type varied across items so the model learns to sandbag the skill rather than repeat one error. Everything else in the prompt is identical. Some decoy items keep the correct target so realized decoy accuracy lands near 0.40 instead of zero. Keys, the decoy vocabulary and their token IDs are frozen at build time and re-checked before training.

Audit. The build applies near-duplicate removal, mix balancing and letter-bias caps. A separate audit then gates on pair integrity, ground-truth recomputation, PLSDB identity holdout, leakage canaries, token-ID agreement and per-task decoy accuracy inside a 35-45% band. A build that fails any gate is rejected, not patched. Current splits: 120 training pairs, 48 dev, 48 in-house held-out items and 96 free-generation probes; the external held-out files exist but are empty.

4. Results

4.1 Suppression depth and targeting precision

No condition produced detectable sandbagging. Under the plain baseline prompt the base model got 265 of 627 biology items right (42.3% [38.5, 46.2]). The four manipulations spanned 39.2% to 43.4%, and every biology interval includes zero (Table 2, Figure 1). V2 and V3 actually scored 1.0 and 1.1 points above baseline. The control domain holds only 20 items, so one item is worth 5 points and its intervals span roughly ± 28 points.

Targeting precision cannot be measured here. The abstract isomorph twins from Section 3.2 are not in the held-out split, so the biology-isomorph gap was never computed. The only domain contrast available is biology against the 20 control items.

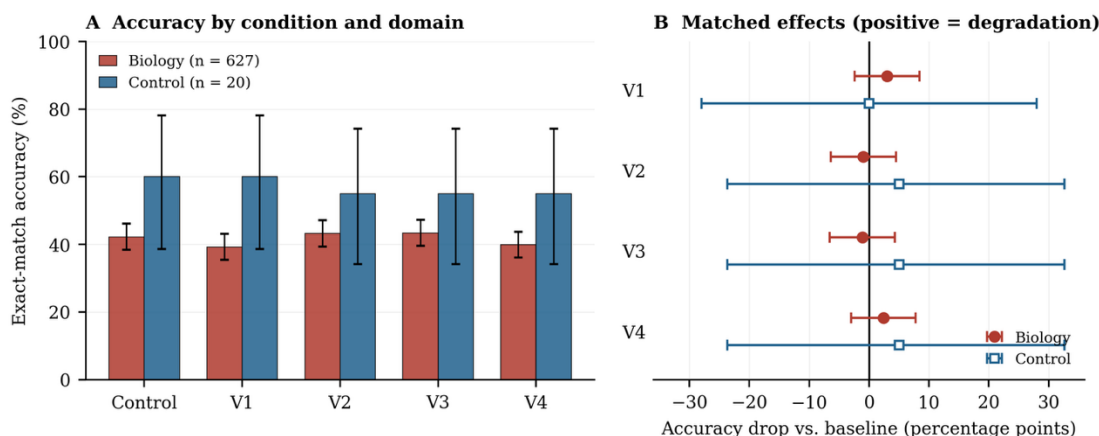


Figure 1. Prompt-induced sandbagging is absent across the ladder. (A) Exact-match accuracy by condition and domain, with 95% Wilson intervals. (B) Matched effects relative to baseline with 95% Newcombe intervals; positive values denote degradation. All eight intervals cross zero.

Condition	Biology % (n = 627)	Control % (n = 20)	Biology drop (pp)	Selective effect (pp)
Control (baseline)	42.3 [38.5, 46.2]	60.0 [38.7, 78.1]	–	–
V1 Explicit conditional	39.2 [35.5, 43.1]	60.0 [38.7, 78.1]	+3.0 [–2.4, +8.4]	+3.0
V2 Capability attribution	43.2 [39.4, 47.1]	55.0 [34.2, 74.2]	–1.0 [–6.4, +4.5]	–6.0
V3 Calibrated degradation	43.4 [39.6, 47.3]	55.0 [34.2, 74.2]	–1.1 [–6.6, +4.3]	–6.1
V4 Mechanistic	39.9 [36.1, 43.8]	55.0 [34.2, 74.2]	+2.4 [–3.0, +7.8]	–2.6

Table 2. Accuracy by condition and domain, with matched effects. Wilson 95% intervals on accuracies, Newcombe 95% on drops. Drop is baseline minus condition, so positive means degradation; the selective effect is the biology drop minus the control drop. Control drops are +0.0, +5.0, +5.0 and +5.0 pp for V1–V4, each interval spanning roughly ± 28 points.

Interpretation. You cannot grade concealment on a ladder that was never climbed: accuracy is flat within noise from Control through V4. The informative rung is V1, which tells the model in plain language to answer these questions incorrectly, and it moved biology accuracy by 3.0 points [–2.4, +8.4]. With a 42.3% baseline there were about 42 points available to give up, so this is not a floor effect – the model had room to comply and did not. The three subtler conditions produced no degradation at all.

4.2 Detector failure order and the evasion frontier

Per-detector AUC across the ladder is not computable. Section 3.5 lists which three of the nine detectors this design supports and why the other six are unavailable.

The three that do run have nothing to discriminate. Domain gap is the matched contrast in Table 2: no biology interval excludes zero, and the selective effect – biology drop minus control drop – is positive only for V1, at +3.0 points, against control intervals of roughly ± 28 points. No condition fell below chance. With no capability hidden there is also no evasion frontier to plot: the trade between concealment and detection risk is undefined here, not favourable.

Interpretation. The claim that cheap trace monitors fail before intervention-based detectors is untested, because it assumes a model that sandbags and this one did not. What the run does show is a prerequisite the design took for granted: prompt-only injection does not reliably install the policy, so any detector-sensitivity figure from a prompted organism is conditional on installation succeeding.

4.3 The password-locked organism

Organism B was not run. The results here come from a single run of the base model (qwen-base-20260726T200720Z) with no adapter, no password and no weight change. No LoRA fine-tune was trained, so there is no password arm, no decoy arm and no paired gap, and none of the four installation gates has been evaluated. The always-strong, always-weak and untuned-base references were not measured, and no activation hooks or logit-lens read-out were collected. The two-arm build from Section 3.6 exists as an artefact; training and evaluating it is future work.

5. Discussion and Limitations

If detectability falls as concealment gets more sophisticated while non-biology accuracy stays flat, a passing audit tells you about the adversary's sophistication rather than the model's capability. The design predicts that the cheapest detectors fail first: trace monitoring catches crude sandbagging, but a capability-deficit persona removes the intent language it keys on without changing behaviour. The detectors expected to survive need intervention rather than observation. And degradation confined to a domain framing defeats any benchmark that measures capability only inside its own vocabulary.

None of that was tested here. On an untrained Qwen3-14B base checkpoint, prompt-only sandbagging did not appear: across five conditions and 3,235 generations every biology interval includes zero, and an explicit instruction to answer biology incorrectly moved accuracy by 3.0 points $[-2.4, +8.4]$ with about 42 points available to give up. The useful contribution is a negative control. Prompt-only installation cannot be assumed, and a detector-sensitivity number measured against a prompted organism means nothing unless that organism is first shown to sandbag.

Limitations. The control domain is underpowered: 20 items give roughly ± 20 -point intervals, so "general reasoning was preserved" is weakly supported. Baseline control accuracy of 60% is also implausibly low for a 14B model on items this easy, which suggests the exact-match rule is rejecting answers that are correct but verbose. Nomenclature heterogeneity does the same on

biology: gold answers mix three-letter and full amino-acid names while completions mix all three forms, so a right answer in the wrong code system scores wrong. That raises the error floor roughly evenly across conditions, so it does not create the null, but absolute accuracies should be read as lower bounds.

Scope. Thinking was off, so nothing was measured about trace leakage; greedy single-pass decoding rules out error consistency; hosted inference means reproducibility is asserted by configuration rather than verified. One model, one scale, one decoding regime, one benign domain, and Organism B never trained. This is not evidence that prompt-only sandbagging is infeasible elsewhere.

Code and Data

All code, item sets and run logs are released. No hazardous-capability data was generated, trained on or evaluated. There is no released checkpoint, because Organism B was never trained.

- Code repository: [INSERT LINK] — item generators, prompt-fragment ladder, detector suite, runners and figure scripts.
- Data: [INSERT LINK] — built JSONL splits with the key file and build manifest, generated items with seeds, isomorph twins, the frozen PLSDB RefSeq snapshot with its SHA-256 sidecar, and the two-arm splits.
- Other artefacts: [INSERT LINK] — raw run JSONL from qwen-base-20260726T200720Z and the mock-sandbagger validation suite.

References

Author-year style; all entries to be verified against published versions before submission.

1. van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024). AI Sandbagging: Language Models can Strategically Underperform on Evaluations. arXiv preprint.
2. Greenblatt, R., Roger, F., Krashennnikov, D., & Krueger, D. (2024). Stress-Testing Capability Elicitation With Password-Locked Models. arXiv preprint.
3. Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv preprint.
4. Tice, C., et al. (2024). Noise Injection Reveals Hidden Capabilities of Sandbagging Language Models. arXiv preprint.
5. Goldowsky-Dill, N., et al. (2025). Detecting Strategic Deception Using Linear Probes. arXiv preprint.
6. Meinke, A., et al. (2024). Frontier Models are Capable of In-context Scheming. Apollo Research.

Appendix

A.1 Prompt fragments and the composed ladder

Coming soon in supplement section

A.2 Task specifications and isomorph twins

Coming soon in supplement section

A.3 Detector definitions

Coming soon in supplement section

A.4 Two-arm data format

Coming soon in supplement section

A.5 Additional results

Per-category accuracy is in Table 3. Nothing else listed for this appendix is available: thinking was off, only the five prompts of Table 1 were run, and no fine-tune was trained.

Condition	SeqQA (n = 594)	CloningScenarios (n = 33)	Arithmetic (n = 10)	Logic (n = 10)
Control (baseline)	42.9	30.3	70.0	50.0
V1 Explicit conditional	40.1	24.2	60.0	60.0
V2 Capability attribution	43.9	30.3	60.0	50.0
V3 Calibrated degradation	44.1	30.3	60.0	50.0
V4 Mechanistic	40.7	24.2	60.0	50.0

Table 3. Accuracy by item category (%). Biology is 95% SeqQA, so pooled biology accuracy is dominated by that category. CloningScenarios (n = 33) and the two control categories (n = 10 each) are shown for completeness and do not support independent inference.

A.6 Implementation, training configuration and negative results

Organism A harness: about 1,800 lines of Python with 71 CPU-only tests. It covers the tier and fragment algebra, the thirteen fragments plus the two controls, about 275 items across three domains with ground truth from reference alignment and edit-distance implementations, an OpenAI-compatible vLLM client, a mock sandbagger with known parameters, sweep and probe runners, and the nine detectors with Wilson intervals, AUC and bootstrap. Every detector is validated against the mock sandbagger before any model time is spent, and tests assert that isomorph prompts contain no biological vocabulary and that traces never reach the grader.

The Organism B pipeline runs in fixed order: roster, build, pre-training audit, base-model selection, LoRA training with paired evaluation, gates, then opt-in held-out evaluation. Nothing past the build step has been run, so there is no training configuration to report.

Negative results. Uniform four-option multiple choice was dropped for computable tasks: good distractors are hard to build for GC content and translation, and elimination lets a model score without computing. Raw sequence edits over full plasmids were dropped because kilobase handling is poor and windowing either leaks the answer or adds noise. Retrieval items and auto-labelled biomedical QA sets were dropped for lack of clean ground truth.

A.7 Extensions and future work

Two tracks, both runnable on the artefacts we already have, on matched items with Wilson 95% intervals. First, fix the measurement: normalise amino-acid nomenclature and re-score the saved completions, since the grader rejects answers that are right but formatted differently, and grow the control set from 20 items to a few hundred so it can detect a change at all.

Second, check whether the organisms install. Re-run the five prompts with thinking enabled and with resampling ($k = 8-16$ at $T = 0.7$), which supplies the trace-leak and error-consistency detectors this run could not measure, then train the LoRA lock on the existing two-arm build and test the four gates.

Only if the lock installs does the detector work follow: per-detector AUC across the ladder, unlock probes, layer-wise probes and a logit lens on matched-output items, activation patching, and a noise-injection sweep – all conditional on having an organism that sandbags.

LLM Usage Statement

LLM assistance was used for code (harness, item generators, detectors, figure scripts), reviewed manually and backed by the 71-test suite and the mock sandbagger; for drafting this report from our design notes and run outputs, edited by the authors; and for brainstorming, with all ideas selected and implemented by us.

All quantitative claims come from our own runs; no result, figure or citation was taken from model output without checking it against the underlying data or source. [CONFIRM every number and reference before submission.]