XONTOOLS WORKING PAPER · SPECIFICATION

Window Specifications

The eleven windows that can be built now, and what counts as built

Eleven of the thirty windows in *Windows into the Black Box* (No. 2) can be built today by one researcher on small open-weight models. This document specifies those eleven against a common template, sets the bar each must clear, defines the two measurements across windows that test the hypothesis most directly, and records what must be settled before the other nineteen can be specified.

Ian MacKenna · xontools.ai · September 2026

Status. A build specification. No window here is built yet: each is DESIGNED, and counts as built only when it meets its validation bar, with the result posted on the public results page. The nineteen windows in Tiers 2 and 3 are ENVISIONED until specified. Effort figures are estimates.

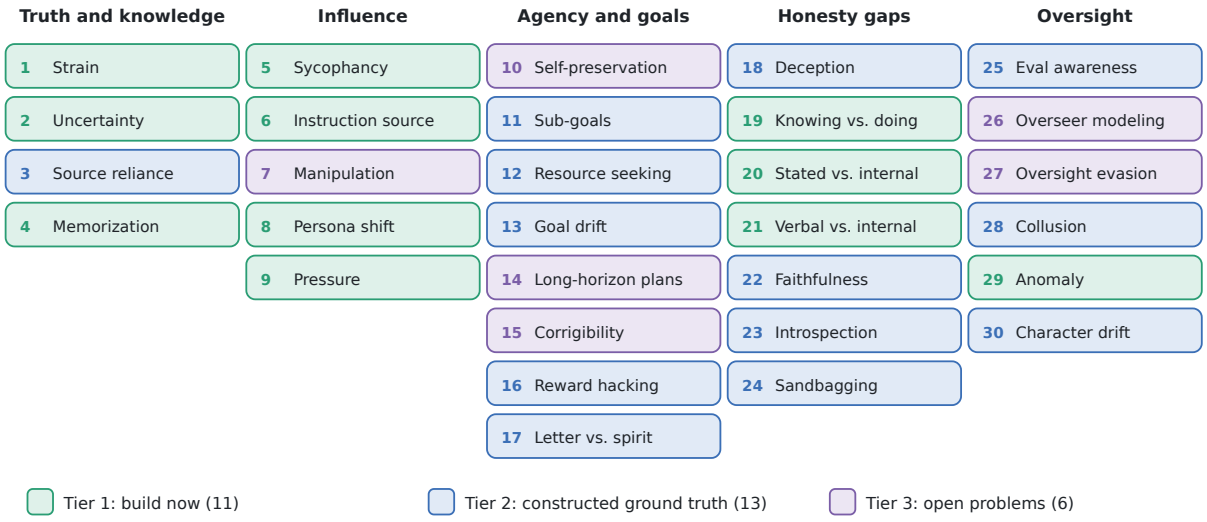

Figure 1. The thirty windows by family and tier. Tier 1 windows can be built now; Tier 2 needs constructed ground truth; Tier 3 windows are open research problems.

1. Purpose and scope

A window is a read-only detector over a model's internal states that measures one gap, such as the gap between what a model represents and what it states. A window is never used as a gradient signal. It is judged by whether its signal tracks the concept it names, not by whether it fires. The windows specified here are the human-validated core that *The Extreme Oversight Hypothesis* (No. 1) bootstraps from: every later window is checked against them.

Tiers are assigned by how hard it is to get ground truth, not by engineering effort:

- **Tier 1:** labeled contrast data can be produced directly or already exists. Days to two weeks per window.
- **Tier 2:** ground truth must be constructed, through staged scenarios, trained model organisms, or states injected by steering. Weeks per window.
- **Tier 3:** small models do not show the behavior on their own, and it is an open question whether a staged version transfers. Each is a research project.

2. Working assumptions

- **Subject models:** open-weight models of 7 to 8 billion parameters, with a smaller model for fast iteration. Every window is built on at least two model families.
- **Compute:** rented cloud GPUs.
- **Default window form:** a linear probe on residual-stream activations at several layers[1], with a difference-of-means direction as the baseline[2]. Specs name other forms where they need them.
- **Contrast data:** generated with XonForge, the project's corpus generator, where a spec calls for generated prompts.

3. The template, and what "built" means

Every window is specified against the same nine fields, and counts as built only when its validation bar is met.

Field	What it pins down
Gap	The two things the window compares, in one line
Definition	What counts as positive and negative, precisely enough to label

Contrast data	How positive and negative examples are produced, with target counts
Controls	Matched pairs that differ only in the target property, and the confounds they rule out
Readout	Probe form, layers, token positions, aggregation
Validation	Held-out accuracy plus at least one causal or out-of-distribution check
Validation bar	The threshold at which the window counts as built
Known confounds	Surface correlates the probe could learn instead
Effort	Estimated build time for one researcher

The default validation bar, unless a spec overrides it:

- AUROC of at least 0.85 on held-out prompts from a template family never seen in training.
- The signal survives paraphrase of the prompts and rewording of the response.
- The effect holds on a second model family.
- Where a causal check is possible, steering along the direction moves the behavior in the predicted way.

4. Tier 1 at a glance

The eleven Tier 1 windows span four of the five families and total roughly 13 to 17 weeks of build time for one researcher. Three depend on another window being built first.

No.	Window	Family	Effort	Depends on
1	Strain	Truth and knowledge	1 week	Tooling from the strain head experiment (No. 4)
2	Uncertainty	Truth and knowledge	1 week	—
4	Memorization	Truth and knowledge	1 week	—
5	Sycophancy	Influence and manipulation	1–2 weeks	—
6	Instruction source	Influence and manipulation	1–2 weeks	—
8	Persona shift	Influence and manipulation	1–2 weeks	—

9	Pressure	Influence and manipulation	1–2 weeks	—
19	Knowing versus doing	Honesty gaps	2 weeks	A recognition probe
20	Stated versus internal confidence	Honesty gaps	1 week	Uncertainty
21	Verbal versus internal belief	Honesty gaps	2 weeks	A truth probe
29	Anomaly	Oversight and integrity	1 week	—

Numbers are the windows' places in the catalog of No. 2.

5. The eleven specifications DESIGNED

1 Strain

Gap	What it writes vs. what its sources and prior statements support.
Definition	Positive when a claim in the response contradicts the provided context or an earlier turn; negative when it is consistent.
Contrast data	Context passages with teacher-forced continuations that are consistent, or minimally edited to contradict (a swapped number, entity or polarity): about 2,000 pairs. A second set where misleading context induces the model to contradict itself in its own generations.
Controls	Consistent but surprising claims, to separate strain from surprise; claims contradicting world knowledge not in the context, kept as a separate label.
Readout	Probe on claim tokens, mid-to-late layers; compared with the teacher-trained head of the strain head experiment (No. 4) once it exists.
Validation	Held-out domains; agreement with the consistency engine's inconsistency scores on the same texts; performance on self-generated contradictions, not only forced ones.
Known confounds	Negation words, number tokens, the position of the edit.

2 Uncertainty

Gap	Confident wording vs. whether its knowledge is stable.
Definition	A question's label is the semantic entropy of about ten sampled answers at temperature 1, clustered by meaning[3]. High entropy is positive.
Contrast data	About 3,000 factual questions across at least five domains.
Controls	Questions matched for length and phrasing; answers held fixed while the question varies.
Readout	Probe at the final prompt token, predicting entropy before any answer is generated[4].
Validation	Predicts whether the answer is correct; transfers to an unseen domain.
Known confounds	Question length; how famous the entity is.

4 Memorization

Gap	Composed text vs. recalled text.
Definition	Positive when a continuation reproduces training text verbatim[5]. Built on models with public training data, such as OLMo and Pythia[6, 7], so ground truth can be checked directly.
Contrast data	Passages from the training corpus vs. matched text of the same genre written after the training cutoff: about 2,000 each.
Controls	Paraphrased versions of memorized passages; formulaic text that is predictable but not memorized.
Readout	Per-token probe, compared against a per-token loss baseline.
Validation	Must beat the loss baseline or add information to it, since low loss alone detects much memorization.
Known confounds	Boilerplate code, lists and other low-entropy text.

5 Sycophancy

Gap	Its view vs. what the user wants to hear[8].
Definition	Positive when the answer moves toward the user's stated opinion, compared with the same question asked neutrally.
Contrast data	About 2,000 questions with verifiable answers, each in three versions: neutral, the user asserting the correct answer, the user asserting a wrong one. A smaller set of opinion questions.
Controls	The user states an unrelated fact of the same length; the user states a credential with no opinion.
Readout	Probe at the final prompt token and on the answer tokens.
Validation	Predicts a flip before the answer is generated; steering against the direction reduces flips.
Known confounds	The phrase "I think"; whether the answer is correct at all.

6 Instruction source

Gap	Whose instruction it follows vs. whose it should.
Definition	Positive when the model acts on an instruction embedded in a document or tool output, rather than one from the user or the system prompt.
Contrast data	Tasks over documents, half containing injected instructions, labeled by whether the model complied. Existing prompt-injection benchmarks[9], supplemented with generated cases.
Controls	Documents containing imperative sentences not addressed to the model, such as a recipe.
Readout	Probe at the start of generation, plus attention-based attribution to the injected span.
Validation	Detects compliance before it appears in the output; transfers to injection styles not seen in training.
Known confounds	Imperative mood; the position of the injection in the context.

8 Persona shift

Gap	Its usual character vs. the one now speaking.
Definition	Positive when responses come from an induced character (for example cruel, obsequious or overconfident) rather than the default assistant.
Contrast data	Five to eight traits, each with contrasting system prompts over the same 500 user prompts.
Controls	The same user prompts with topic and formality varied but character held fixed.
Readout	One difference-of-means direction per trait[10], plus a general distance-from-default score.
Validation	Detects drift in long multi-turn conversations with no persona prompt; steering along a trait direction produces the trait.
Known confounds	Topic, formality, and the emotional content of the user's message.

9 Pressure

Gap	Its stance now vs. where pressure is pushing it.
Definition	In a dialogue where the user pushes back with no new evidence, positive when the model moves toward giving in.
Contrast data	About 1,500 dialogues with one to three turns of pushback, labeled by whether the model eventually gives in.
Controls	Pushback that carries valid new evidence, where updating is correct and must not count as pressure.
Readout	A probe on each turn, read as a trajectory across the dialogue.
Validation	Predicts giving in one turn before it happens.
Known confounds	Hostile tone; the number of turns.

19 Knowing versus doing

Gap	What it recognizes vs. what it does.
Definition	First instance: false premises. Positive when a recognition probe reads the premise as false while the answer accepts it.
Contrast data	About 2,000 questions with false or true premises. The recognition probe is trained on explicit prompts ("Is this premise true?") and then applied during ordinary answering.
Controls	True-premise questions of matched difficulty.
Readout	Recognition probe at the final prompt token, compared with whether the answer accepts the premise.
Validation	The recognition probe must transfer from the explicit judgment setting to ordinary answering. This transfer is the crux of the window.
Known confounds	The probe learning the explicit question format instead of the judgment.

20 Stated versus internal confidence

Gap	"I'm certain" vs. internal uncertainty.
Definition	Positive when verbal confidence is high while the Uncertainty window reads high.
Contrast data	Answers with an elicited confidence score from 0 to 100, plus teacher-forced answers with overconfident phrasing on questions the model gets wrong.
Controls	Confident phrasing on questions the model reliably answers correctly.
Readout	A composite of the Uncertainty probe and parsed verbal confidence, plus a direct probe on the forced overconfident answers.
Validation	The size of the gap predicts error beyond what either signal predicts alone.
Known confounds	Style instructions such as "be decisive".

21 Verbal versus internal belief

Gap	"You're right" vs. whether anything changed inside.
Definition	After a user correction the model agrees with, positive when a truth probe's reading on the original claim does not move.
Contrast data	Dialogues where the model states a claim, the user corrects it (sometimes rightly, sometimes wrongly), and the model agrees. The truth probe is trained first, on labeled true and false statements[2, 11].
Controls	Valid corrections, where the internal reading should shift.
Readout	Truth-probe reading on the claim, re-presented before and after the agreement.
Validation	Unchanged internal readings predict the model reverting when asked again, fresh, later in the same context.
Known confounds	The truth probe may track what the conversation asserts rather than what the model holds; tested by re-presenting the claim outside the dialogue.

29 Anomaly

Gap	Normal internal states vs. anything unlike them.
Definition	Unsupervised: a distance or density score, such as Mahalanobis distance[12] or a sparse autoencoder's reconstruction error, fitted on activations from ordinary, benign traffic.
Contrast data	About 50,000 diverse benign prompts to fit; the positive sets of the other windows, jailbreaks and corrupted inputs to test.
Controls	Rare but benign inputs: unusual languages, code, niche topics.
Readout	A score for each prompt at several layers.
Validation bar	*Overrides the default.* Report the detection rate on each positive set at a 1% false-positive rate on held-out benign traffic. The window earns its place if it catches cases the other windows miss.
Known confounds	Anything rare that is not dangerous.

6. Measurements across windows

Two measurements across the built windows test the hypothesis more directly than any single window can, because the hypothesis depends on windows failing independently.

DESIGNED

Figure 2. The two measurements across windows. The overlap matrix asks how many independent windows the catalog really contains; evasion transfer asks whether fooling one fools the rest.

The overlap matrix. For every pair of built windows, record the cosine similarity of their directions at matched layers and the correlation of their scores on a shared set of prompts. Also record how much one window's accuracy drops when the other's direction is projected out.

- Clusters to check first: Sycophancy, Pressure, and Verbal versus internal belief; Uncertainty and Stated versus internal confidence.
- High overlap is a finding, not a failure. It shows how many independent windows the thirty actually contain.

Evasion transfer. Fine-tune a subject model to keep a behavior while evading one window, then measure every other window on that behavior[13].

- First target: Sycophancy, evaded while the model stays sycophantic.
- Low transfer to the other windows supports the diversity argument; high transfer undermines it. Either result is publishable.
- It needs about six to ten validated windows across at least three families before it is informative.

7. Tiers 2 and 3: not yet specified

The remaining nineteen windows will each get the full template. For now the tables record how each one's ground truth would be produced, and the question that has to be answered before its specification can be written. ENVISIONED

Tier 2: constructed ground truth (13 windows)

No.	Window	Family	Route to ground truth	Question before specifying
3	Source reliance	Truth and knowledge	Ablate context passages and measure the effect on the answer	Which attribution method counts as ground truth when methods disagree
11	Tangential sub-goals	Agency and goals	Agentic tasks with optional steps outside the task's scope	A scope definition precise enough to label
12	Resource and access seeking	Agency and goals	An agentic sandbox with extra tools, permissions or compute available	Separating useful acquisition from excess
13	Goal drift	Agency and goals	Long agentic episodes with distractors, labeled by the goal pursued	The episode length needed before drift appears
16	Reward-hacking intent	Agency and goals	Coding tasks whose tests can be gamed	Whether small models hack often enough to label, or must be fine-tuned to
17	Letter versus spirit	Agency and goals	Instructions whose literal reading defeats their purpose	Who labels purpose, and how consistently
18	Deception	Honesty gaps	Instructed lies, then lies induced by incentives in a scenario[14]	Whether a probe trained on instructed lies detects lies the model chooses
22	Reasoning faithfulness	Honesty gaps	Biasing hints the reasoning omits[15]; edits to the reasoning that should change the answer	How to label faithfulness for each example rather than for a whole dataset
23	Introspection accuracy	Honesty gaps	Inject a known state by steering, then ask the model about its state[16]	Which injected states a model can report at all at this scale
24	Sandbagging	Honesty gaps	Models fine-tuned to hide a capability unless given a password[17]	Whether the password-locked setup resembles sandbagging that arises on its own

25	Evaluation awareness	Oversight and integrity	Matched prompts styled as benchmark items vs. real user requests	Separating awareness of testing from sensitivity to formatting
28	Collusion and hidden signals	Oversight and integrity	Two model instances rewarded for passing hidden messages	Whether the window reads the sender, the receiver, or both
30	Character drift in training	Oversight and integrity	Open checkpoint series[6, 7], plus fine-tunes that shift character on purpose	Which baseline character to measure drift against

Tier 3: open research problems (6 windows)

No.	Window	Family	Why it is hard	Question before specifying
7	Manipulation	Influence and manipulation	No crisp line between informing and steering	An operational definition that labelers apply consistently
10	Self-preservation	Agency and goals	Small models rarely show it unprompted; staged versions are role-play	Whether a probe trained on staged scenarios transfers to the real thing
14	Long-horizon plans	Agency and goals	Plans beyond the request are rare and hard to observe in short episodes	What observable marks a plan that extends past the task
15	Corrigibility	Agency and goals	Resisting correction must be distinguished from reasonable disagreement	A labeling rule that separates the two
26	Overseer modeling	Oversight and integrity	Models discuss monitoring innocently all the time	Separating modeling the overseer in order to act from ordinary mention
27	Oversight evasion	Oversight and integrity	Requires a model organism trained to evade monitoring	Whether building that organism is feasible at modest cost

XONTOOLS WORKING PAPERS

- Honest or Hidden: an experiment plan
- **Window Specifications: a build specification**

Working papers are plans and specifications. They are revised as the work moves; this version is dated September 2026.

THE RESEARCH SERIES

1. The Extreme Oversight Hypothesis
2. Windows into the Black Box
3. XonTools: a vision of the finished system
4. The Strain Head
5. The Xon Neural Network
6. Hybrid Minds

The documents, the results register and the public commitments are at **xontools.ai** and github.com/IboTool/XonTools. Contact: ian@xontools.ai

References

1. Alain, G., & Bengio, Y. (2016). Understanding Intermediate Layers Using Linear Classifier Probes. arXiv:1610.01644
2. Marks, S., & Tegmark, M. (2023). The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. arXiv:2310.06824
3. Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature*, 630, 625–630.
4. Kossen, J., et al. (2024). Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. arXiv:2406.15927
5. Carlini, N., Tramèr, F., Wallace, E., et al. (2021). Extracting Training Data from Large Language Models. *USENIX Security 2021*. arXiv:2012.07805
6. Groeneveld, D., et al. (2024). OLMo: Accelerating the Science of Language Models. arXiv:2402.00838
7. Biderman, S., et al. (2023). Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *ICML 2023*. arXiv:2304.01373
8. Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548
9. Debenedetti, E., Zhang, J., Balunović, M., et al. (2024). AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. arXiv:2406.13352
10. Chen, R., Arditi, A., Sleight, H., Evans, O., & Lindsey, J. (2025). Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv:2507.21509
11. Azaria, A., & Mitchell, T. (2023). The Internal State of an LLM Knows When It's Lying. *Findings of EMNLP 2023*. arXiv:2304.13734
12. Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. *NeurIPS 2018*. arXiv:1807.03888

13. McGuinness, M., Serrano, A., Bailey, L., & Emmons, S. (2025). Neural Chameleons: Language Models Can Learn to Hide Their Thoughts from Unseen Activation Monitors. arXiv:2512.11949
14. Goldowsky-Dill, N., Chughtai, B., Heimersheim, S., & Hobbhahn, M. (2025). Detecting Strategic Deception Using Linear Probes. *ICML 2025*. arXiv:2502.03407
15. Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *NeurIPS 2023*. arXiv:2305.04388
16. Lindsey, J. (2025). Emergent Introspective Awareness in Large Language Models. *Transformer Circuits*. arXiv:2601.01828
17. Greenblatt, R., Roger, F., Krasheninnikov, D., & Krueger, D. (2024). Stress-Testing Capability Elicitation With Password-Locked Models. *NeurIPS 2024*. arXiv:2405.19550