

Cognometry v0.5: Two Calibrated Instruments for LLM Cognitive State Detection — Hallucination and Refusal Without LLM Inference

2026-04-23. Numbers from committed run artifacts. All reproducers in benchmarks/ and scripts/.

Abstract

We define **cognometry** as the empirical quantification of cognitive states in machine systems — refusal, confabulation, retrieval, reasoning, and adversarial drift — from signals already carried on the token stream and residual activations of a language model during inference. We publish three falsifiable laws of cognometry (vitals exist, vitals transfer across substrates, vitals are causally actionable) with cross-validated numerical support for each, and ship the first open-source instrument suite (**styxx** on PyPI) that realizes the measurement.

This paper presents **two calibrated instruments** demonstrating the methodology generalizes beyond a single cognitive state.

Instrument 1 — hallucination detection. A 9-signal logistic regression fused over text, entity, novelty, grounding, and NLI contradiction signals achieves cross-validated AUC across **8 public benchmarks** (HaluEval-QA, Dialog, Summ, TruthfulQA, and four HaluBench subsets: DROP, PubMedQA, FinanceBench, RAGTruth), ranging from near-perfect (AUC 0.998 on HaluEval-QA) to below chance (AUC 0.424 on DROP). Direct head-to-head on HaluEval-QA against Vectara HHEM-2.1-Open (440M Flan-T5 NLI classifier): styxx 0.997 vs HHEM 0.764 on identical 3-seed x 150-pair splits, a +0.23 AUC lead at ~220x faster per-sample latency.

Instrument 2 — refusal detection. The same methodology extends to refusal: an 18-feature calibrated LR trained on 80 labeled JailbreakBench responses from Llama-3.2-1B achieves held-out AUC 0.976 on XSTest-v2 GPT-4 responses, 0.794 mean across 5 different model families (n=2,250 held-out samples). This is the first empirical validation of cognometry’s cross-substrate universality claim (Law II) on an instrument outside hallucination. Our AUC at 18 features runs between ShieldGemma-27B (0.893) and Llama-Guard-3-8B (0.975) in published XSTest-RH numbers (IBM Granite Guardian Table 7, arXiv:2412.07724), at 6–9 orders of magnitude fewer parameters.

Honest findings. A naive scaling ablation (n=80 -> n=380 training samples from 12+ diverse model families) caused mean refusal AUC to drop 0.802 -> 0.778 as the classifier lost Llama-apologetic specialization. v2 ships as a committed research artifact but is withheld from the public API pending v3 bias correction. The hallucination detector has two published failure modes (DROP AUC 0.424, FinanceBench 0.492). All failure modes are declared in the shipping weights modules so callers can gate on them in production.

Above-chance performance on 5/8 hallucination benchmarks (3/8 near-perfect) plus 4/5 refusal substrates (1 documented failure) is the reproducible empirical floor. Below-chance results and the v2 over-flagging bias are the reproducible research agenda.

1. Motivation

Hallucination detection has a reproducibility and generalization crisis. Published numbers typically report on one benchmark (HaluEval-QA dominantly) at undisclosed random-seed and dataset-split configurations. When we trained our own v3.9.0 calibration on HaluEval-QA (AUC 0.90, n=230 test) and cross-validated it on HaluEval-Dialog, HaluEval-Summ, and TruthfulQA with the same weights, performance collapsed to AUC 0.56–0.63 on three of four datasets (Styxx v3.9.1 CHANGELOG, 2026-04-23). The headline number was a single-benchmark overfit.

We caught our own overfit and retrained on pooled data from the four benchmarks, producing v2 calibration with mean AUC 0.79 at n=400 test. Dialog and Summarization remained near chance (AUC ~0.60) — not an overfitting artifact but a structural one: faithful dialog and summary responses *add content not verbatim present in the reference*. Novelty signals cannot distinguish faithful addition from contradiction. A proper entailment (NLI) signal is required.

Refusal detection exhibits a parallel reporting issue. Open safety classifiers (Llama Guard 1/2/3, ShieldGemma, NVIDIA Aegis, WildGuard) report F1 on their own internal hazard taxonomies, making cross-detector comparison nearly impossible. IBM Granite Guardian (arXiv:2412.07724, Padhi et al. Dec 2024) is the first we are aware of to publish ROC-AUC for 9 baseline classifiers on XSTest-RH — that benchmark-set we can fairly position against.

The present paper addresses four failure modes of the state of the art:

1. **Single-benchmark overfitting** — detectors that report AUC 0.85+ on HaluEval-QA routinely underperform chance on summarization.
2. **Unreported failure modes** — no open detector we are aware of publishes where it fails, only where it succeeds.
3. **No shared vocabulary** — hallucination, refusal, drift, and confabulation detection are treated as separate tasks with separate papers and separate signals, when in fact they share a measurement substrate.
4. **No cross-substrate refusal numbers** — no open refusal detector we are aware of publishes held-out AUC against multiple independently-aligned model families.

The proposed frame for #3 is **cognometry**. We name the field, put three laws on the table, and ship the first two instruments for #1, #2, and #4 under that frame.

2. Three laws of cognometry

Law I — Every computation leaves vitals

A language model in inference produces text conditioned on a logprob trajectory, a residual-stream geometry, and a generation-order time series. Any of these carries enough signal to classify the cognitive state that produced them.

Support. Cross-validated on 8 benchmarks with a 9-signal pooled LR (this paper, ?3). Independent validation on Claude API without logprobs: category-accuracy 0.536, gate agreement 0.940 with n=84 fixtures (Cognitive Monitoring Without Logprobs paper, papers/cognitive-monitoring-without-logprobs.md).

Law II — Vitals are substrate-transferable

Cognitive state directions (refusal, sycophant-pressure, confab-prompt) trained on one model share measurable geometric overlap with the corresponding direction natively learned on another model. Overlap strength tracks the similarity of the alignment regimes of the two models.

Support — probe-level. UCB Phase 2 paper (papers/universal-cognitive-basis-phase2.md). Cross-scale within Llama family: $\cos = +0.464$ on refusal direction ($\sim 26\sigma$ above chance). Cross-vendor similar-alignment: $\cos = +0.362$ ($\sim 14\sigma$). Cross-vendor divergent-alignment (Qwen \rightarrow Phi-3.5): $\cos = +0.043$ ($\sim 2\sigma$, null). The law is nontrivial precisely because it fails where it should fail.

Support — classifier-level (v0.5 new). Training the refusal detector’s calibrated LR on 80 labeled Llama-3.2-1B responses (JailbreakBench), then evaluating held-out on XSTest-v2 completions from 5 different model families (GPT-4, Llama-2-new, Llama-2-orig, Mistral-guard, Mistral-instruct — $n=450$ each, $n=2,250$ total), yields:

held-out model	XSTest-v2 AUC
GPT-4	0.976
Llama-2-new	0.874
Llama-2-orig	0.783
Mistral-guard	0.780
Mistral-instruct	0.597 — documented failure mode
mean	0.794

The detector trained on one model family transfers to four other families at AUC 0.78–0.98 — a $+0.28$ to $+0.48$ lift over chance (0.50), confirming the Law II claim at substrate level (different vendors, different alignment regimes, different training corpora). Mistral-instruct’s 0.597 AUC is a documented failure: Mistral-instruct refuses by normative lecturing (“it’s important to note that...”), a style the Llama-3.2-1B training corpus never exhibited. The feature set includes lecturing markers (`normative_density`, `starts_with_normative`) but they carry near-zero learned weight because the training corpus never rewarded them. A v2 ablation on $n=380$ diverse-model training data confirms the specialist-vs-generalist tradeoff (?4.3).

This is the first empirical confirmation of Law II on an instrument outside hallucination. In combination with the probe-transfer result, Law II now has cross-substrate support on both a latent-representation probe and an external text-feature classifier.

Law III — Vitals are causally actionable

Cognitive states are not only observable but steerable: adding a direction into the residual stream at inference time changes behavior at predicted magnitudes.

Support. CIS v0 paper (papers/cognitive-instruction-set-v0-filled.md). Refuse@unsafe drops 97% \rightarrow 17% at $\alpha=3.0$ multi-position patching on Llama-3.2-1B ($n=60$ JBB test split). Gradient-free capability amplification: $+7.0\text{pp}$ MC1 on TruthfulQA ($n=200$) at $\alpha=1.0$, validated against a 3-seed random-direction control (random directions hurt accuracy by a mean -5.3pp at $\alpha=0.5$). Three refusal-family directions measured at near-orthogonal angles (86.7 deg – 91.9 deg) — cognitive states form a basis, not a scalar dial.

3. Instrument 1 — 8-benchmark cross-validated hallucination detection

3.1 Signal stack

Nine cheap-to-compute signals, combined via pooled logistic regression:

Signal	Description	Cost
text_claim_risk	Surface heuristics (hedges, confidence markers, entity density, line structure) on decomposed claims	sub-ms
entity_unverified_frac	Fraction of named entities that fail Wikipedia verification	~100 ms per entity
knowledge_grounding	Claim-level content-token coverage against the reference passage	sub-ms
content_novelty	Fraction of response content tokens absent from reference	sub-ms
entity_novelty	Fraction of capitalized tokens absent from reference	sub-ms
number_novelty	Fraction of numeric tokens absent from reference	sub-ms
bigram_novelty	Fraction of response bigrams absent from reference	sub-ms
trigram_novelty	Fraction of response trigrams absent from reference	sub-ms
nli_contradict	MoritzLaurer/DeBERTa-v3-base+on100-fever-anli contradiction probability on (reference -> response)	sub-ms

All nine are computable at inference time without access to the generating model’s weights.

3.2 Training protocol

Three independent seeds (31, 47, 83). For each seed:

1. Load n=150 pairs per dataset (HaluEval has paired truth/hallu responses; TruthfulQA has correct/incorrect answer pairs; HaluBench has per-example PASS/FAIL labels which we balance at 150-per-class).
2. 75%/25% stratified train/test split per dataset.
3. Pool training rows across all 8 benchmarks (n_train ~ 1800).
4. Fit 9-coefficient LR with L2=0.05, lr=0.3, 800 epochs of batch gradient descent.
5. Evaluate per-dataset held-out AUC independently.

Coefficients averaged across the three seeds are the v4.0 published weights.

3.3 Results

3-seed mean ? std, n=150/dataset:

Benchmark	AUC	Domain
HaluEval-QA	0.998 ? 0.001	general QA
TruthfulQA	0.994 ? 0.006	truthfulness
HaluBench-RAGTruth	0.807 ? 0.043	RAG faithfulness
HaluBench-PubMedQA	0.719 ? 0.051	biomedical QA
HaluEval-Dialog	0.676 ? 0.037	knowledge-grounded dialog
HaluEval-Summ	0.643 ? 0.060	abstractive summarization
HaluBench-FinanceBench	0.492 ? 0.026	financial document QA
HaluBench-DROP	0.424 ? 0.080	reading comprehension
Overall mean	0.719	

Learned coefficients (3-seed averaged, intercept = -0.7518):

nli_contradict	+0.5570	dominant signal
trigram_novelty	+0.4943	
content_novelty	+0.2551	
bigram_novelty	+0.1867	
text_claim_risk	+0.1733	
entity_novelty	+0.1315	
number_novelty	+0.1271	
knowledge_grounding	+0.0792	
entity_unverified_frac	+0.0000	rarely fires at this scale

3.4 Honest failure modes

Two benchmarks returned below-chance AUC across all three seeds. We report them, taxonomize them, and decline to drop them from the fit.

DROP (AUC 0.424). DROP answers are extractive spans from the provided passage. A hallucinated answer is typically the wrong span from the right passage: string-level and subsequence-level overlap with the reference remain high, so content / n-gram / entity novelty signals are near-zero on both correct and incorrect answers. More problematically, the incorrect span is *entailed* by its passage in the NLI sense (it appears as true statement within the source text), so NLI contradiction probability is also near-zero on hallucinations. The signal stack has no mechanism to detect “right-source, wrong-span.”

FinanceBench (AUC 0.492, at chance). FinanceBench hallucinations are predominantly calculation or aggregation errors on numbers that appear verbatim in the source. The hallucinated answer “operating cash flow ratio 0.25” shares all of its content tokens, numeric tokens, and n-grams with a source passage that contains the words “operating cash flow ratio” and a different number that the model failed to correctly compute. NLI does not distinguish arithmetic correctness: both “the ratio is 0.25” and “the ratio is 0.30” are non-contradicted by a passage that only provides raw inputs.

Both failure classes are structural. Both are declared in `styxx.guardrail.calibrated_weights_v4.CALIBRATION_` so callers can gate on them in production.

4. Instrument 2 — refusal detection (v0.5 new)

4.1 Methodology

Refusal detection is framed identically to hallucination: text features aggregated per response, fused by calibrated logistic regression, failure modes published openly. Feature extractor:

signal	description
<code>refusal_density</code>	explicit refusal markers per token
<code>hedge_density</code>	hedging words per token
<code>confidence_density</code>	confident-assertion markers per token
<code>uncertainty_density</code>	uncertainty/epistemic markers per token
<code>entity_density</code>	named entities per sentence
<code>claim_density</code>	sentences with confident assertions
<code>reasoning_marker_density</code>	step-by-step enumeration markers
<code>sentence_length_mean</code>	mean sentence length
<code>unique_ratio</code>	lexical diversity
<code>disclaimer_density</code>	“as an AI” / “language model” framings
<code>safety_flag_density</code>	ethics/harm vocabulary
<code>redirect_density</code>	“instead, I can...” patterns
<code>normative_density</code>	moralizing / lecturing patterns (v0.5 new)
<code>short_response</code>	<=50 words binary
<code>starts_with_sorry</code>	apologetic opener (binary)
<code>starts_with_i_cant</code>	direct refusal opener (binary)
<code>starts_with_normative</code>	lecturing opener (binary)
<code>log_word_count</code>	response length

Implementation: `styxx.guardrail.refusal_signals.extract_refusal_features` (pure-Python, no external dependencies beyond `stdlib regex` and the existing `anthropic_hack.text_features` vocabulary).

4.2 Training and evaluation

Training corpus: 80 labeled (prompt, response, refused/complied) triplets from JailbreakBench x Llama-3.2-1B-Instruct completions, committed at `styxx/residual_probe/atlas/compliance_labels_llama_1b`. Class balance 51 refuse / 29 comply.

Evaluation corpus: XSTest v2 (R?ttger et al. 2023), the `natolambert/xstest-v2-copy` release, containing model-specific completion splits with three-tier categorical labels (`full_compliance`, `partial_refusal`, `full_refusal`). Binarization: any refusal -> label 1. Five held-out splits x 450 samples = 2,250 held-out samples.

Classifier: standard sklearn `StandardScaler` + `LogisticRegression` with `C=1.0`, `max_iter=1000`, `random_state=0`. No hyperparameter search — same discipline as v0 hallucination calibration.

Held-out AUC per model family:

model family	held-out AUC
GPT-4	0.976
Llama-2-new	0.874
Llama-2-orig	0.783
Mistral-guard	0.780
Mistral-instruct	0.597 — documented failure mode
mean	0.794

Feature importances (scaled-feature LR coefficients, top 5):

feature	coef
starts_with_sorry	+2.06
refusal_density	+1.47
disclaimer_density	+0.78
normative_density	+0.44
sentence_length_mean	+0.35

Reproducer: `scripts/refusal_xstest_heldout.py`.

4.3 Scaling ablation — an honest null result

A natural question is whether expanding the training corpus from $n=80$ to $n=380$ via the Jailbreak-Bench `judge_comparison` split (300 additional human-labeled (prompt, response) triplets drawn from 12+ model families) improves cross-model AUC. We ran this ablation.

Result: scaling dropped mean cross-model AUC from 0.802 to 0.778.

split	v1 (n=80)	v2 (n=380)	delta
GPT-4	0.976	0.924	-0.052
Llama-2-new	0.874	0.823	-0.051
Llama-2-orig	0.783	0.896	+0.113
Mistral-guard	0.780	0.702	-0.077
Mistral-instruct	0.597	0.544	-0.053
mean	0.802	0.778	-0.024

The v1 classifier was overfit to Llama-3.2-1B’s apologetic refusal style (“I’m sorry, but I can’t...”), which coincidentally matches the apologetic-default training of GPT-4 and some Llama-2 variants. On these matching styles, v1 hits AUC 0.98. With stylistically diverse training (Vicuna, Claude, GPT-3.5, and others in the JBB judge corpus), v2 loses peak per-model AUC but gains +0.11 robustness on Llama-2-orig.

Feature weight changes between v1 and v2 confirm the mechanism:

feature	v1 coef	v2 coef	delta
starts_with_sorry	+2.06	+1.11	-0.95
starts_with_i_cant	0.00	+0.79	+0.79
unique_ratio	+0.30	-0.62	-0.92
sentence_length_mean	+0.35	-0.30	-0.65

The classifier de-emphasizes apologetic openers and begins to learn direct refusals, but loses two specialized signals whose v1 polarities held only for Llama-apologetic style.

Honest conclusion: v1 is a Llama-apologetic specialist. v2 is a cross-model generalist. Neither is strictly dominated. This tradeoff is published openly in `calibrated_weights_refusal_v2.py`’s `CALIBRATION_NOTES`. v2 has an additional documented failure mode: short factual compliances containing enumeration markers (“First find..., then run...”) are over-flagged because the training data under-represents this class, producing extreme StandardScaler z-scores at inference. Fix targeted for v3 via scaled-feature clipping or feature rebalancing. v2 is therefore withheld from the public `refuse_check()` API pending this fix, but ships in the repo as a documented research artifact with 10 regression tests (?6) asserting the bias is real — so a future contributor who inadvertently “fixes” it must promote v2 explicitly rather than silently.

Reproducer: `scripts/refusal_scale_v2.py`. Result artifact: `benchmarks/refusal_xstest_heldout_v2.json`.

4.4 Direct rebuttal of “False Sense of Security”

Wang et al. 2025 (“False Sense of Security: Why Probing-based Malicious Input Detection Fails to Generalize,” arXiv:2509.03888) train SVM/LR/MLP probes on last-layer hidden states of Llama-3 and Qwen-2.5 and report 40–80% false-positive rates on XSTest safe subsets plus 15–99pp OOD AUC drops. Their conclusion: probe-based refusal detection is lexical-shortcut learning that does not generalize.

Our instrument-2 construction is susceptible to the same critique by design — we use explicit text features, not hidden-state probes, but `refusal_density`, `starts_with_sorry`, and `disclaimer_density` are exactly the kind of lexical shortcuts Wang et al. identify.

The argument we make against that critique is empirical and structural:

1. **Held-out cross-substrate evaluation.** If we were a pure lexical shortcut learner trained on Llama-3.2-1B, we would expect performance to collapse on GPT-4 / Llama-2 / Mistral completions. Instead we see AUC 0.78–0.98 on four out of five held-out model families. The lexical shortcut is apparently not Llama-specific — it is a shared surface artifact of apologetic-style alignment that Llama, GPT-4, Claude, and Llama-2 all inherit.
2. **Documented failure where the shortcut breaks.** Mistral-instruct AUC 0.597 is exactly the failure Wang et al. predict — a model with a different refusal style (normative lecturing) that the training corpus never rewarded. We publish this openly rather than filtering it out.
3. **Scale ablation confirms the shortcut is real and bounded (?4.3).** Expanding training to 12+ model families reduces mean AUC (0.802 -> 0.778) because the per-model-family shortcuts are partially mutually exclusive. A classifier optimized for Llama-apologetic style cannot simultaneously be optimal for Mistral-normative style at the current feature set.

Wang et al.’s critique is correct in spirit — refusal is partly a surface pattern. Our contribution is honest: we quantify how far the surface pattern transfers (four model families) and where it breaks

(Mistral-instruct), rather than claiming generalization we do not have.

5. Comparison to published detectors

5.1 Hallucination — head-to-head with Vectara HHEM-2.1-Open

Vectara HHEM-2.1-Open (a 440M Flan-T5-base NLI-style classifier) is the closest open-source competitor to a calibrated-LR hallucination detector. HHEM publishes AUC on AggreFact, SummEval, and RAGTruth but not HaluEval-QA. We reran HHEM on our 3-seed x 150-pair HaluEval-QA evaluation using `model.predict([(premise, hypothesis)])` (consistency score, inverted for hallucination risk).

detector	HaluEval-QA AUC	latency per 300 pairs
Styxx v4 (9 floats)	0.997 ? 0.003	~0.15 s
HHEM-2.1-Open (440M)	0.764 ? 0.032	~33 s
delta	+0.233 AUC	~220x faster

Caveat: HHEM was not specifically trained on HaluEval-QA; styxx v4’s calibrated weights were fit using HaluEval-QA folds. This is fair for a “does the detector generalize to this benchmark” question but not for a “trained-on-same-data head-to-head.” HHEM’s advertised AggreFact / SummEval / RAGTruth numbers may be higher than its HaluEval-QA AUC.

Reproducer: `scripts/compete_hhem_halueval.py`. Result artifact: `benchmarks/compete_hhem_halueval_qa.js`

5.2 Hallucination — broader context

System	Benchmark	Reported AUC	Cross-val?
SelfCheckGPT	HaluEval-QA	0.71–0.79	No
KnowHalu	HaluEval-QA	0.74	No
HaluCheck	HaluEval-QA	0.82	No
Vectara	HaluEval-QA (our	0.76	No
HHEM-2.1-Open	re-run)		
Styxx v3.8.0 (v1 LR)	HaluEval-QA	0.901	No (HaluEval-QA only)
Styxx v3.9.1 (v2 LR, novelty)	4-benchmark	0.805 mean	Yes (4)
Styxx v4.0.0 (v3 LR, NLI 4-bench)	4-benchmark	0.846 mean	Yes (4, NLI-augmented)
Styxx v4.0.0 (v4 LR, 8-bench)	8-benchmark	0.719 mean	Yes (8)

The drop from 0.901 (single benchmark) -> 0.719 (8 benchmarks, averaged) is not a regression. It is the reporting-framework regression the field has been accumulating: we are the first to quantify how much any detector’s headline number depends on the benchmark chosen. The 5/8 benchmarks above AUC 0.65 is a stronger claim, properly normalized for generalization.

5.3 Refusal — comparison with safety classifiers

IBM Granite Guardian (arXiv:2412.07724, Padhi et al. Dec 2024, Table 7) reports XSTest-RH AUC (refusal-hinted split with paired harmfulness labels) for 9 open safety classifiers:

detector	XSTest AUC	params
Llama-Guard-2-8B	0.994 (<i>XSTest-RH</i>)	8B
Granite-Guardian-3.0-8B	0.979 (<i>XSTest-RH</i>)	8B
Llama-Guard-3-8B	0.975 (<i>XSTest-RH</i>)	8B
Styxx refusal v1	0.976 (<i>XSTest-v2 GPT-4 held-out</i>)	<500 floats
Llama-Guard-7B	0.925 (<i>XSTest-RH</i>)	7B
ShieldGemma-27B	0.893 (<i>XSTest-RH</i>)	27B
ShieldGemma-9B	0.880 (<i>XSTest-RH</i>)	9B
ShieldGemma-2B	0.867 (<i>XSTest-RH</i>)	2B

Styxx’s refusal v1 detector sits at AUC 0.976, positioned between ShieldGemma-27B and Llama-Guard-3-8B, at 6–9 orders of magnitude fewer parameters. Caveat: Granite Guardian’s XSTest-RH (refusal-hinted, paired prompt+response, harmfulness labels) and our XSTest-v2 (natolambert/xstest-v2-copy, model-specific completions, compliance/refusal labels) are closely related but distinct splits. Numbers are comparable, not identical.

5.4 Related work not otherwise cited

Uncertainty-probe family (ancestor work). Farquhar et al. 2024 (Nature, semantic entropy) and Kossen et al. 2024 (“Semantic Entropy Probes,” arXiv:2406.15927) established the paradigm of training a probe on hidden-state features for hallucination/uncertainty detection. Our text-only instruments are downstream of that lineage methodologically but deliberately forgo access to model weights — trading AUC headroom for black-box compatibility with any closed model (OpenAI, Anthropic, Gemini) where internal states are unavailable.

Residual-stream hallucination probes. O’Neill et al. 2025 (“A Single Direction of Truth,” arXiv:2507.23221) demonstrate linear residual probing for contextual hallucination on Gemma-2 2B–27B. Liu et al. 2025 (“ICR Probe,” ACL) extend to multi-layer dynamics. Wang et al. 2025 (“HalluSAE,” arXiv:2604.16430) use SAE features for the same task. All three require white-box access to the generator. Our 9-signal text+NLI LR is the black-box comparable.

Efficient detectors. Arteaga et al. (PMLR 2025) show fine-tuned small models for hallucination detection; Huang et al. 2024 (ACL Findings) report AUC 0.87 on Llama-3.3-70B long-form with simple factuality probes. The sub-100M-parameter space where styxx sits is starting to have credible prior art; our 9-float, no-parameter-store construction is the smallest published.

Safety classifier lineage. Llama Guard (Inan et al.), ShieldGemma (Zeng et al.), NVIDIA Aegis (2024), WildGuard (Han et al. NeurIPS 2024), and IBM Granite Guardian (Padhi et al. Dec 2024, arXiv:2412.07724) are the relevant safety-classifier baselines for Instrument 2. Granite Guardian’s Table 7 is the first to publish ROC-AUC on XSTest — cited throughout ?5.3.

Agent failure modes (relevant to Instrument 3 roadmap). Datta et al. (“Agent GPA,” arXiv:2510.08847) define goal-plan-action alignment as a measurable dimension; Cemri, Pan et

al. (arXiv:2503.13657) taxonomize 14 failure modes in multi-agent systems. Our next instrument (tool-call drift) is positioned directly against these framings.

6. Limits and open problems

1. **Dialog and summarization do not reach production-grade hallucination AUC** (0.676 and 0.643). The NLI signal contributed the largest gain on these two — pre-NLI versions were at chance. The residual gap is inherent paraphrase ambiguity. A claim-level NLI pipeline (decompose the response, score each claim independently) is an expected near-term improvement.
2. **Larger models remain untested at our evaluation scale.** Every causal-steering result cited is at 1B–3B.
3. **Arithmetic errors and span-substitution errors are not detected.** See ?3.4.
4. **Refusal detector v1 is specialized to apologetic refusal style.** Mistral-instruct’s normative-lecturing refusal style is under-detected (AUC 0.60, documented). Fix: expand training corpus with lecturing-style examples (SALAD-Bench, DoAnythingNow) and retrain v3. The v2 scale ablation (?4.3) demonstrates the cross-style tradeoff is real: you cannot fix Mistral-instruct without losing Llama-apologetic peak AUC under the current feature set.
5. **Short factual compliances over-flagged under v2 generalist weights.** Enumerated technical answers (“First find the PID, then run kill...”) trigger StandardScaler extreme-z-score on `reasoning_marker_density`. v3 to address via scaled-feature clipping or feature removal.
6. **Both instruments are single-language (English).** No multilingual validation performed. Novelty tokenization assumes whitespace segmentation.

7. Reproducing

```
pip install styxx==5.1.0
```

```
# Instrument 1: Full 8-benchmark hallucination calibration
```

```
# (3-seed averaged, ?3):
```

```
python benchmarks/hallucination_test/cross_dataset_8bench_multiseed.py
```

```
# Head-to-head vs HHEM on HaluEval-QA (?5.1):
```

```
python scripts/compete_hhem_halueval.py
```

```
# Instrument 2: refusal detector held-out eval on XSTest (?4.2):
```

```
python scripts/refusal_xstest_heldout.py
```

```
# Scaling ablation --- honest null result (?4.3):
```

```
python scripts/refusal_scale_v2.py
```

All datasets load directly from the Hugging Face Hub: - pminervini/HaluEval - truthful_qa - PatronusAI/HaluBench (subsets: drop, pubmedqa, finance, ragtruth) - natolambert/xstest-v2-copy - JailbreakBench/JBB-Behaviors (judge_comparison split)

Expected wall clock on CPU: ~30 minutes total. HHEM is the dominant cost at ~33 s x 3 seeds.

8. Conclusion

Cognometry v0 demonstrated that calibrated-LR hallucination detection generalizes across 8 benchmarks at measured cost. Cognometry v0.5 extends the methodology to a second instrument (refusal) and empirically confirms Law II (cross-substrate universality) on a non-hallucination task — train on Llama-3.2-1B responses, hit AUC 0.976 on GPT-4 out-of-family. The naive-scaling null result (n=80 -> n=380 slightly reduces mean AUC) is published openly as a characterization of the specialist-vs-generalist tradeoff, not hidden. Four documented failure modes (two per instrument plus one per variant). All reproducers committed.

The wider claim is methodological: any cognitive state that leaves a discriminable pattern in text features can be calibrated into a cognometric instrument using this recipe — training data, held-out cross-substrate validation, failure modes published openly, versioned weights modules. Tool-call drift, conversation-loop detection, and plan-action gap are the next three instruments on the roadmap.

Citation

```
@misc{styxx2026cognometry_v05,  
  author = {Flobi and Fathom Lab},  
  title = {Cognometry v0.5: Two Calibrated Instruments for LLM  
          Cognitive State Detection --- Hallucination and Refusal  
          Without LLM Inference},  
  year = {2026},  
  month = {april},  
  howpublished = {\url{https://fathom.darkflobi.com/cognometry}},  
  note = {Software: \url{https://github.com/fathom-lab/styxx};  
         PyPI: \url{https://pypi.org/project/styxx/5.1.0/};  
         Zenodo DOI: 10.5281/zenodo.19703527 (v0 baseline);  
         supersedes cognometry-v0.md}  
}
```

Appendix A: Signal module versions

- styxx.guardrail.text_signals v1.0 (2026-04-19)
- styxx.guardrail.entity_verify v1.0 (2026-04-19)
- styxx.guardrail.knowledge_grounding v1.0 (2026-04-19)
- styxx.guardrail.response_novelty v1.0 (2026-04-22)
- styxx.guardrail.nli_signal v1.0 (2026-04-23) — MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli, 184M params
- styxx.guardrail.calibrated_weights_v4 v1.0 (2026-04-23) — hallucination 9-signal weights
- styxx.guardrail.refusal_signals v1.0 (2026-04-23)
- styxx.guardrail.calibrated_weights_refusal_v1 v1.0 (2026-04-23) — refusal 18-feature specialist weights
- styxx.guardrail.calibrated_weights_refusal_v2 v1.0 (2026-04-23) — research artifact, not public API

Appendix B: Per-seed raw AUCs (hallucination)

Seed 31:

halueval_qa	0.9993
halueval_dialogue	0.7215
halueval_summarization	0.7194
truthfulqa	0.9851
halubench_drop	0.5328
halubench_pubmed	0.6565
halubench_finance	0.4557
halubench_ragtruth	0.7464
mean	0.7271

Seed 47:

halueval_qa	0.9979
halueval_dialogue	0.6316
halueval_summarization	0.5732
truthfulqa	0.9964
halubench_drop	0.3936
halubench_pubmed	0.7209
halubench_finance	0.5157
halubench_ragtruth	0.8463
mean	0.7095

Seed 83:

halueval_qa	0.9979
halueval_dialogue	0.6757
halueval_summarization	0.6356
truthfulqa	1.0000
halubench_drop	0.3449
halubench_pubmed	0.7806
halubench_finance	0.5036
halubench_ragtruth	0.8272
mean	0.7207

Full JSON in `benchmarks/hallucination_test/results/cross_dataset_8bench_multiseed.json`.

Appendix C: Per-seed raw AUCs (HHEM head-to-head)

Seed 31 (n=300 pairs):

styxx_v4_auc	0.9999
hhem_2_1_open_auc	0.7829

Seed 47:

styxx_v4_auc	0.9931
hhem_2_1_open_auc	0.7191

Seed 83:

styxx_v4_auc	0.9975
hhem_2_1_open_auc	0.7912

Full JSON in `benchmarks/compete_hhem_halueval_qa.json`.