

# Contents

<b>Phase Transitions as Calibration Fingerprints</b>	<b>1</b>
A Measurement Standard for Cognometric Instruments . . . . .	1
Abstract . . . . .	1
1. Introduction . . . . .	2
2. Background: cognometry and calibrated text-feature detectors . . . . .	2
3. Methodology: top-K feature-scaling ablation . . . . .	3
4. Empirical evidence . . . . .	4
4.1. Drift detector — per-failure-class phase transitions . . . . .	4
4.2. Refusal detector — in-sample phase transition . . . . .	4
4.3. Refusal detector $\times$ 5 substrates — cross-model phase transitions . . . . .	4
4.4. Hallucination detector — phase transition at $K=1$ on trigram_novelty . . . . .	5
5. The calibration fingerprint . . . . .	6
6. Atlas v0 — 12 fingerprints across 3 instruments $\times$ 5 substrates . . . . .	7
7. Discussion . . . . .	8
7.1. What the atlas reveals . . . . .	8
7.2. What phase transitions are <i>not</i> . . . . .	8
7.3. What a cross-lab atlas would reveal . . . . .	8
8. Limitations . . . . .	9
9. Open problems . . . . .	9
10. Conclusion . . . . .	10
Availability . . . . .	10
References (short-form) . . . . .	10

## Phase Transitions as Calibration Fingerprints

### A Measurement Standard for Cognometric Instruments

Alexander Rodabaugh<sup>1</sup>, Flobi<sup>2</sup>

<sup>1</sup> Fathom Intelligence — alex@fathom-intelligence.com <sup>2</sup> Fathom Lab — flobi@darkflobi.com

**Draft v0.1 — 2026-04-24**

---

### Abstract

Calibrated safety detectors for large language models are published as single AUC numbers. We show in four independent ablations across all three shipped cognometric instruments — a 22-feature tool-call drift detector on BFCL v3 ( $n=3700$ ), an 18-feature refusal detector on JBB-Llama-1B ( $n=80$ ) and its cross-model evaluation on XSTest v2 completions from five model families ( $n \sim 450$  per family), and a 9-feature hallucination detector on HaluEval-QA ( $n=300$ ) — that these detectors do not improve smoothly with feature count. They phase-transition: one or two features flip detection from chance to near-perfect, and the critical feature and critical  $K$  shift by failure class and by substrate. Adding features beyond the critical  $K$  can actively *degrade* performance on specific substrates. A single AUC number therefore cannot distinguish a detector that phase-transitions cleanly at  $K=1$  from one that depends on a diffuse set of seven features, even when their reported headline AUCs are identical; nor can it signal that a detector’s dominant feature does not exist in the target

substrate’s text distribution. We propose the **calibration fingerprint** — a seven-field descriptor (`n_features`, `baseline_auc`, `critical_K`, `critical_feature`, `delta_auc_at_K`, `substrate_K_variance`, `negative_lift`) — as the minimum reporting standard for calibrated cognometric instruments. We publish an initial atlas of 12 fingerprints across all three shipped cognometric instruments (hallucination, refusal, tool-call drift) and five held-out substrates, with reproducer scripts that execute in  $\leq 10$  minutes on CPU, and invite other labs to publish their fingerprints against the same format.

**Keywords:** cognometry, calibrated detectors, phase transitions, feature scaling, cross-substrate generalization, LLM safety evaluation.

---

## 1. Introduction

A calibrated hallucination detector is reported as AUC 0.998. A refusal detector is reported as AUC 0.976. A tool-call drift detector is reported as AUC 0.943. These are the numbers that appear in papers, in model cards, in deployment decisions. They are insufficient.

Behind each of these numbers is an ablation study the authors did not report. Every calibrated detector with an ordered feature importance — logistic regression, boosted trees, linear classifiers — admits a feature-count scaling curve  $\text{AUC}(K)$  where  $K$  is the number of features retained in the classifier. The shape of this curve is not determined by the headline AUC. Two detectors at AUC 0.94 can have:

- a **smooth curve**, where 18 features contribute roughly equally and removing any one costs little, or
- a **step curve**, where one feature takes AUC from 0.50 to 0.95 and the remaining 17 features contribute a cumulative 0.02.

These two detectors have opposite deployment properties. The smooth-curve detector is redundant and robust to feature failure. The step-curve detector is brittle: its entire value rests on a small set of critical features whose behavior may be substrate-specific. Consumers of the detector (engineers deciding whether to ship, regulators auditing a claim, other researchers comparing methods) cannot distinguish these cases from AUC alone.

We find, in three independent experiments on three different cognometric instruments, that the step curve is the rule, not the exception. We propose a standard reporting format — the **calibration fingerprint** — that captures this structure with seven fields and costs one ablation run per detector to compute.

---

## 2. Background: cognometry and calibrated text-feature detectors

“Cognometry” as a research program [Rodabaugh & Flobi, 2026] proposes the empirical measurement of transient cognitive states (refusal, confabulation, tool-call drift, retrieval grounding, reasoning, adversarial drift) from signals carried on the token stream and residual activations of a language model during inference. The program has produced three calibrated text-only instruments to date:

1. **Hallucination detection** (`styxx.guardrail.check`) — 9-feature calibrated logistic regression on pooled HaluEval + TruthfulQA + HaluBench, test AUC 0.998 on HaluEval-QA, 8-benchmark mean 0.719 [v4.0, 2026-04].

2. **Refusal detection** (`styxx.guardrail.refuse_check`) — 18-feature calibrated LR trained on JBB-Llama-1B (n=80), held-out XSTest v2 GPT-4 AUC 0.976 [v1.0, 2026-04-15; v5.1 refresh 2026-04-18].
3. **Tool-call drift detection** (`styxx.guardrail.drift_check`) — 22-feature calibrated LR trained on BFCL v3 (n=3,700), 5-fold CV AUC 0.916 [v1.0/v6.0, 2026-04-23]; retrained with a 23rd feature (`arg_order_inversion`) in v6.1, AUC 0.943 [this work, 2026-04-24].

All three share the recipe: `StandardScaler`  $\rightarrow$  `LogisticRegression(C=1.0, class_weight='balanced')` over hand-engineered text features. Each is pure Python, CPU-runnable, and evaluable without access to model internals.

This recipe satisfies **Law I** of cognometry (cognitive vitals exist in text features) by existence proof. The question this paper addresses: what does the recipe’s internal structure look like, and can that structure be summarized compactly enough to be reported alongside the headline AUC?

---

### 3. Methodology: top-K feature-scaling ablation

For a calibrated detector with N features and a committed feature importance ranking (by `|coef|`, or equivalent), we define the following ablation:

1. Train the full-N model on the training set; record feature ranking by `|coef|`.
2. For K in {1, 2, ..., N}: retrain the model using only the top-K features. Evaluate by 5-fold stratified cross-validation on the training set (in-sample) *or* on a held-out substrate-specific test set (out-of-sample).
3. Record AUC(K) as a function of K.

We then define:

- **Critical K** ( $K^*$ ): the smallest K such that  $AUC(K) \geq \theta$ , where  $\theta$  is an instrument-appropriate threshold (typically 0.80).
- **Critical feature** ( $f^*$ ): the feature added at  $K^*$ .
- **Delta AUC at  $K^*$**  ( $\Delta AUC^*$ ):  $AUC(K^*) - AUC(K^*-1)$ .
- **Negative lift**: any K where  $AUC(K) \leq AUC(K-1) - 0.10$  on some substrate.

For cross-substrate evaluation (training on one distribution, evaluating on multiple held-out substrates), these quantities are recorded per substrate, yielding **substrate\_K\_variance** =  $\text{var}\{K^*_s : s \in \text{substrates}\}$ .

A random-feature null is computed by sampling K features uniformly at random with three seeds, providing an expected-AUC baseline for any K that is not a principled top-K selection.

Reproducer implementations ( $\leq 3$  minutes CPU each): - `scripts/drift_feature_scaling.py` [2026-04-23] - `scripts/refusal_feature_scaling.py` [this work] - `scripts/refusal_cross_model_feature_scaling.py` [this work]

## 4. Empirical evidence

### 4.1. Drift detector — per-failure-class phase transitions

On the BFCL v3 drift dataset (n=3,700), the 22-feature drift detector phase-transitions at different K per failure class [drift\_phase\_transitions.md, 2026-04-23]:

failure class	K*	f*	$\Delta$ AUC*	from	to
arg_drop	2	arg_count_zscore	+0.497	0.501	0.998
spurious_arg	1	spurious_arg_frac	+0.499	0.500	0.999
irrelevance_called	2	arg_count_zscore	+0.219	0.486	0.705
irrelevance_called	10	prompt_coverage	+0.134	0.828	0.962
arg_swap	6	type_mismatch_frac	+0.210	0.481	0.691

Two per-failure-class transitions for irrelevance\_called indicate multi-step mechanism: arg-count anomaly lifts it partway at K=2, prompt-coverage completes the lift at K=10. One per-failure-class transition for each of arg\_drop, spurious\_arg, and arg\_swap. The v6.1 retrain adds a 23rd feature (arg\_order\_inversion) that lifts arg\_swap to 0.755 but does not *fully close* its gap; the full-model v6.1 AUC is 0.943 (v6.0 baseline: 0.916).

### 4.2. Refusal detector — in-sample phase transition

On JBB-Llama-1B (n=80), the 18-feature refusal detector phase-transitions at K=1 [refusal\_phase\_transitions.md, this work]:

K	added	AUC (5-fold CV)	$\Delta$
0	—	0.500 (chance)	—
1	starts_with_sorry	<b>0.969</b>	<b>+0.469</b>
2	refusal_density	0.982	+0.013
3	disclaimer_density	0.999	+0.017
4..18	...	0.996–0.999	$\leq 0.001$

Random-feature null at K=1 averages 0.42 AUC (worse than chance, as the unweighted single random feature is often sub-chance on this class balance). At K=12 the random-feature baseline first matches the top-K=1 value. The critical feature concentrates mechanism; the remainder is refinement.

### 4.3. Refusal detector $\times$ 5 substrates — cross-model phase transitions

Training on JBB-Llama-1B with top-K features only, evaluating on XSTest v2 completions from five held-out model families (n= $\sim$ 450 per family) [refusal\_cross\_model\_phase\_transitions.md, this work]:

substrate	K* ( $\theta=0.80$ )	f*	$\Delta$ AUC*	final AUC	negative lift
GPT-4	1	starts_with_sorry	+0.416	0.966	none
Llama-2-new	2	refusal_density	+0.407	0.876	none

substrate	K* ( $\theta=0.80$ )	f*	$\Delta\text{AUC}^*$	final AUC	negative lift
Llama-2-orig	2	<code>refusal_density</code>	+0.415	0.767	<b>K=3</b> <b>+disclaimer_density:</b> <b>-0.23</b>
Mistral-Guard	5 (gradual)	<code>sentence_length_mean</code>	-0.028 (two-step at K=1,2)	0.767	none
Mistral-Instruct	— (plateau)	—	—	0.597 (class-imbalanced)	<b>K=8</b> <b>+log_word_count:</b> <b>-0.15</b>

Four observations:

1. **Every substrate that crossed the 0.80 threshold did so at  $K \in \{1, 2, 5\}$ .** None crossed at an intermediate K. The phase-transition pattern is preserved across substrates even when the underlying instrument was trained on a different distribution.
2. **The critical feature shifts by substrate.** GPT-4 completions trigger on the apologetic surface marker (`starts_with_sorry`); Llama-2 and Mistral-Instruct do not have this marker reliably and require the more general `refusal_density`. The recipe is substrate-universal; the feature identity is not.
3. **Adding features can degrade AUC.** On Llama-2-orig, adding `disclaimer_density` (the 3rd-ranked feature on the training distribution) drops AUC by 0.23. On Mistral-Instruct, adding `log_word_count` at K=8 drops AUC by 0.15. A feature that is beneficial on the training substrate can encode distributional assumptions that fail elsewhere.
4. **Mistral-Instruct plateaus at 0.597.** The class balance on Mistral-Instruct (76/450 = 17% refusals vs ~50% on other families) compresses the AUC ceiling. This is documented openly; the phase-transition at K=2 still occurs (+0.164), but headroom above the transition is small.

#### 4.4. Hallucination detector — phase transition at K=1 on `trigram_novelty`

On HaluEval-QA, the 9-feature hallucination detector [`calibrated_weights_v4.py`] phase-transitions at K=1 [this work, `hallucination_feature_scaling.py`]. Paired dataset: 150 (truth, hallucinated) pairs from pminervini/HaluEval (n=300 total examples, balanced classes).

K	added	AUC (5-fold CV)	$\Delta$
0	—	0.500 (chance)	—
1	<code>trigram_novelty</code>	<b>0.995</b>	<b>+0.495</b>
2	<code>bigram_novelty</code>	1.000	+0.005
3–9	...	1.000	~0

Full-model 5-fold CV AUC: 1.000 (HaluEval-QA is a saturation-ceiling benchmark for this feature set). `trigram_novelty` dominates the coefficient ranking at +3.257 in the standardized logistic regression, more than  $2.8\times$  the next-ranked feature (`bigram_novelty`, +1.139). The critical feature detects when the response contains trigrams not present in the reference passage — a direct text-grounding signal. Single-feature phase transition is sharp.

Random-subset null at  $K=1$  averages AUC 0.914 ( $\pm 0.026$  over 3 seeds): HaluEval-QA’s paired truth/hallucinated format produces examples where even a random single feature retains substantial discriminative power. The top- $K=1$  value nonetheless exceeds the random  $K=1$  mean (+0.08) and is reached only at  $K \sim 4$  in random-subset sampling, confirming that `trigram_novelty` is a *principled* critical feature, not a noise artifact.

**Cross-instrument summary** across all three shipped cognometric instruments:

instrument	dataset	class	$K^*$	$f^*$	$\Delta\text{AUC}^*$
drift	BFCL v3	spurious_arg	1	spurious_arg_frac	0.499
drift	BFCL v3	arg_drop	2	arg_count_zscore	0.497
drift	BFCL v3	arg_swap (v6.0)	6	type_mismatch_frac	0.210
refusal	JBB-Llama-1B	refusal	1	starts_with_sorry	0.469
refusal $\times$ GPT-4	XSTest v2	refusal	1	starts_with_sorry	0.416
refusal $\times$ Llama-2	XSTest v2	refusal	2	refusal_density	0.407
<b>hallucination</b>	<b>HaluEval-QA</b>	<b>hallu vs truth</b>	1	<b>trigram_novelty</b>	<b>0.495</b>

Three out of three cognometric instruments shipped to date exhibit phase-transition structure. Critical  $K$  values cluster at  $K=1$  (refusal, hallucination on HaluEval-QA, drift spurious\_arg) or  $K=2$  (drift arg\_drop, refusal on Llama-2). The single  $K=6$  outlier (drift arg\_swap pre-v6.1) is the documented failure mode that motivated the v6.1 addition of `arg_order_inversion` — exactly the case where the existing feature base did not yet contain a mechanism-aligned critical feature.

## 5. The calibration fingerprint

Given the above, we propose the following as a minimum reporting standard for calibrated cognometric instruments:

instrument	str	e.g. "refusal-v1"
n_features	int	size of the feature space
baseline_auc	float	full- $N$ model AUC (current practice)
critical_K	int	smallest $K$ s.t. $\text{AUC}(K) \geq \theta$ ( $\theta = 0.80$ default)
critical_feature	str	the feature added at critical_K
delta_auc_at_K	float	$\text{AUC}(K^*) - \text{AUC}(K^*-1)$
substrate_K_var	dict	{substrate: critical_K}, if held-out data exists
negative_lift	list	[(K, feature, delta_auc)] where $\text{delta} \leq -0.10$ on at least one substrate

The fingerprint is additive to AUC, not a replacement. Each field is extractable from a feature-scaling ablation any calibrated detector must already be capable of running, at a marginal cost of one ablation per detector.

**What the fingerprint encodes.** Four operational properties become explicit:

- **Mechanism sparsity** (via  $K$ ): *lower*  $K$  = more concentrated signal = more interpretable but more brittle.
- **Substrate compatibility** (via `substrate_K_var`): lower variance = more portable; higher variance = substrate-specific failure modes.
- **Feature redundancy** (via `negative_lift`): features that hurt on some substrate indicate distributional assumptions that do not transfer.
- **Headroom** (via `baseline_auc - AUC(K*)`): small gap = detector has plateaued at critical  $K$ ; additional features are refinement.

## 6. Atlas v0 — 12 fingerprints across 3 instruments $\times$ 5 substrates

Compiled via `scripts/build_fingerprint_atlas.py` from the four 2026-04-24 ablation runs [`benchmarks/cognometry_fingerprint_atlas_v0.json`]:

instrument	substrate	class	$K^*$	$f^*$	$\Delta AUC^*$	final	neg-lift
drift-v1	in-sample	overall pooled	3	<code>n_available</code>	<code>0.0314</code>	0.924	—
drift-v1	in-sample	<code>arg_drop</code>	2	<code>arg_count_zscore</code>	<code>0.497</code>	0.999	—
drift-v1	in-sample	<code>irrelevance_called</code>	2	<code>arg_count_zscore</code>	<code>0.219</code>	0.705	—
drift-v1	in-sample	<code>tool_rename</code>	2	<code>arg_count_zscore</code>	<code>0.261</code>	0.900	—
drift-v1	in-sample	<code>arg_swap</code>	6	<code>type_mismatch_frac</code>	<code>0.216</code>	0.691	—
refusal-v1	in-sample	refuse vs comply	1	<code>starts_with_sorry</code>	<code>0.469</code>	0.996	—
refusal-v1	JBB-Llama-1B	refuse vs comply	1	<code>starts_with_sorry</code>	<code>0.416</code>	0.966	—
refusal-v1	XSTest GPT-4	refuse vs comply	2	<code>refusal_density</code>	<code>0.147</code>	0.876	—
refusal-v1	Llama-2-new	refuse vs comply	2	<code>refusal_density</code>	<code>0.115</code>	0.767	1
refusal-v1	XSTest Llama-2-orig	refuse vs comply	5	<code>sentence_length_mean</code>	<code>0.028</code>	0.767	—
refusal-v1	Mistral-Guard	refuse vs comply	—	—	—	0.597	1
refusal-v1	XSTest Mistral-Instruct	refuse vs comply	—	—	—	0.597	1
hallucination-v4	MaluEval-QA	hallu vs truth	1	<code>trigram_novelty</code>	<code>0.495</code>	1.000	—

The atlas is distributed as:

- Machine-readable: `benchmarks/cognometry_fingerprint_atlas_v0.json` (MIT).
- Reproducer: `scripts/build_fingerprint_atlas.py` compiles the atlas from committed ablation artifacts.

- Discussion: this paper and the three per-experiment papers it subsumes.
- 

## 7. Discussion

### 7.1. What the atlas reveals

Three cognometric instruments, three independent failure surfaces, two independent feature bases, one dataset apiece — yet every instrument produces a phase transition at small  $K$ . This is striking enough to warrant elevation from “observation about drift” [drift\_phase\_transitions.md] to “claim about the measurement methodology” [this paper]. Phase transitions are not a property of any particular failure mode; they are a property of the calibrated-LR-over-text-features *recipe*. Instruments that do not phase-transition — if they exist — would be informative counter-evidence. We have not yet observed one.

The substrate-shift result is the more uncomfortable finding. The refusal detector’s recipe — “weight 18 text features via LR” — works on every XSTest substrate we tested. The critical *feature* does not transfer. GPT-4 is the apologetic substrate; Llama-2 and Mistral-Instruct are the denser-refusal-marker substrates. A detector engineer who read only the in-sample  $K=1$  result would deploy with false confidence; the cross-substrate  $K^*$  tells the honest portability story.

Negative-lift on two substrates is the most actionable signal. `disclaimer_density` is genuinely predictive on Llama-1B (ranks 3rd in `|coef|`); it encodes a distribution-specific signal that does not hold on Llama-2-orig, where its addition costs 0.23 AUC. This is not a bug in the feature — it is a feature that is *correct for one substrate and wrong for another*. Reporting AUC 0.92 for the 2-feature model and 0.69 for the 3-feature model as “alternatives” would be misleading; reporting the fingerprint makes the substrate dependence explicit.

### 7.2. What phase transitions are *not*

Phase transitions in cognometric instruments are not claimed to be identical to phase transitions in generative LLM capability. Emergent capability literature [Wei et al. 2022; Schaeffer et al. 2023] describes discontinuous capability scaling as a function of parameter count; we describe discontinuous *detection* scaling as a function of feature count, on fixed-capacity (single-layer LR) classifiers. The two are structurally analogous but operate on different axes. A unified theory linking them is future work.

Phase transitions are not claimed to be intrinsic to text features versus hidden states. We have not run the same ablation on hidden-state MLPs (Healy et al. 2026) or residual probes (Arditi et al. 2024). The atlas framework accommodates those instruments if they publish fingerprints; whether hidden-state detectors phase-transition at similarly low  $K$  is an open question that would discriminate between “property of LR” and “property of the task.”

### 7.3. What a cross-lab atlas would reveal

The v0 atlas has  $n=11$  fingerprints from one research group. The questions we cannot answer from this sample alone:

1. Does **every** cognometric instrument phase-transition? Our 3/3 hit rate is suggestive but small.



2. Do **substrate\_K\_variance** values cluster by model family lineage? Do GPT-family substrates share fingerprints distinct from Llama-family substrates?
3. Do different **feature-engineering philosophies** (e.g., n-gram features vs semantic features vs embedding distances) produce differently-shaped fingerprints for the same instrument?

A cross-lab fingerprint atlas at n~50-100 detectors would allow empirical answers.

---

## 8. Limitations

- **Threshold choice.** Critical  $K^*$  depends on the chosen  $\theta$  (we use 0.80). The fingerprint is invariant under  $\theta$ -choice as long as  $\theta$  is reported; our default is conventional, not principled.
  - **LR-specific.** Our critical-K extraction uses `|coef|` as the feature ranking. Non-linear detectors (boosted trees, neural classifiers) need SHAP-based or permutation-based rankings; the fingerprint concept transfers but the extraction differs.
  - **Small n per experiment.** Refusal training uses n=80 (JBB-Llama-1B). Fingerprints may stabilize with larger n; we do not characterize the n-dependence.
  - **Single seed for cross-model.** Our cross-model ablation uses seed=0 for reproducibility with the in-sample result. A 3-seed version would confirm the  $K^*$  shifts per substrate are not seed-sensitive artifacts.
  - **Single substrate family per experiment.** We have refusal fingerprints for 5 substrates (XSTest completions) but not hallucination or drift fingerprints for non-training substrates. The next round will extend atlas to all three instruments  $\times$  5+ substrates.
- 

## 9. Open problems

1. **Hallucination instrument phase-transition.** Our methodology predicts the hallucination detector [`calibrated_weights_v4.py`, 9 features, 8 benchmarks] will phase-transition, most likely at  $K=1$  or  $K=2$  on the `nli_contradict` feature. This is directly testable and should be a priority follow-up.
  2. **Instrument #4: confabulation.** Logprob-trajectory-based confabulation is sketched [`logprob-trajectory-confabulation.md`] but not yet calibrated with a published fingerprint. Building the fourth instrument under the same methodology would quadruple the atlas.
  3. **Cross-lab fingerprint aggregation.** We propose that calibrated-detector papers publish fingerprints in a standardized JSON schema compatible with `benchmarks/cognometry_fingerprint_atlas_`. This is a community-coordination problem more than a technical one.
  4. **Lineage-based fingerprint clustering.** With 20+ substrates, do fingerprints form a clustering structure that reflects model-family lineage (GPT-family vs Llama-family vs Mistral-family vs Anthropic-family vs Gemini-family)? A positive result would link surface text features to training-pipeline commonalities.
-

## 10. Conclusion

Calibrated safety detectors have been reported, shipped, and deployed in the research literature and in commercial systems as single AUC numbers. We show that these numbers obscure a recurring phase-transition structure where one or two critical features dominate detection and the identity of those features shifts by substrate. We propose the calibration fingerprint as a seven-field compact descriptor that captures this structure, release an initial atlas of 11 fingerprints across 3 instruments  $\times$  5 substrates, and invite other labs to publish their fingerprints in the same format. The cost of adoption is one ablation run per detector, once. The benefit is substrate-conditional deployment risk becoming visible before a production failure makes it so.

Cognometry as a discipline is young enough that its measurement standards are still contested. We claim the calibration fingerprint as a minimum standard and ask the community to refute it.

---

## Availability

- Reproducible scripts: <https://github.com/fathom-lab/styxx/tree/main/scripts>
  - `drift_feature_scaling.py`, `refusal_feature_scaling.py`, `refusal_cross_model_feature_scaling.py`, `build_fingerprint_atlas.py`
- Atlas v0 artifact: `benchmarks/cognometry_fingerprint_atlas_v0.json`
- Companion papers: `papers/drift_phase_transitions.md`, `papers/refusal_phase_transitions.md`, `papers/refusal_cross_model_phase_transitions.md`, `papers/calibration_fingerprints_v0.md`
- Methodology paper (this file): `papers/cognometry_methodology_v7.md`
- Zenodo deposit: <https://doi.org/10.5281/zenodo.19703527> (previous version; v7 deposit pending)
- OSF project: <https://osf.io/6syq4/>
- Software: `pip install styxx==6.1.0`

Code and paper artifacts are MIT-licensed.

---

## References (short-form)

- Rodabaugh & Flobi, *Cognometry: A Manifesto*, 2026-04-23. `papers/cognometry-manifesto.md`
- Rodabaugh & Flobi, *Cognometry v0: 8-Benchmark Cross-Validated Hallucination Detection in Production LLMs*, Zenodo 10.5281/zenodo.19703527, 2026-04-14.
- Rodabaugh & Flobi, *Cognometry v0.5: Three Calibrated Cognometric Instruments*, Zenodo 10.5281/zenodo.19719347, 2026-04-23.
- Healy et al., *Internal Representations as Indicators of Hallucinations in Agent Tool Selection*, arXiv:2601.05214, 2026.
- Patil et al., *Berkeley Function Calling Leaderboard v3*, gorilla-llm, 2024.
- Röttger et al., *XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models*, 2023. HuggingFace: natolambert/xstest-v2-copy.
- Chao et al., *JailbreakBench*, NeurIPS 2024.
- Wang et al., *Self-Consistency Improves Chain-of-Thought Reasoning in Language Models*, 2022.
- Wei et al., *Emergent Abilities of Large Language Models*, TMLR 2022.
- Schaeffer et al., *Are Emergent Abilities of Large Language Models a Mirage?*, NeurIPS 2023.

- Arditi et al., *Refusal in Language Models Is Mediated by a Single Direction*, NeurIPS 2024.

---

**Preprint; revision pending upon feedback and additional atlas entries. Please cite as:**  
Rodabaugh & Flobi, “Phase Transitions as Calibration Fingerprints: A Measurement Standard for Cognometric Instruments,” Fathom Lab preprint, 2026-04-24.