# Topological Signatures of Deception: Comparing Persistent Homology with Linear Probes

*SPAR Fall 2026 Mentor Proposal — Supplementary Detail*

Santiago Maniches · Independent Researcher · ORCID 0009-0005-6480-1987

---

## Abstract

Linear probes trained on residual-stream activations can distinguish honest from deceptive responses with high in-distribution accuracy, but recent work reports sensitivity to distribution shift and monitor-targeted fine-tuning. This project evaluates whether persistent-homology features derived from attention graphs and activation point clouds provide complementary predictive information or different robustness properties. Track A adapts TOHA's per-response attention-graph divergence to deception detection. A conditional Track B pilot adapts Fay et al.'s population-level activation-point-cloud method, using prompt-disjoint splits and a probe baseline aggregated over the same batches. The primary evaluation compares in-distribution performance, out-of-distribution generalization, and cross-monitor robustness under an attack targeting the linear probe. A direct attack on the topological monitor is conditional on time and compute. Persistence diagrams have formal stability guarantees under specified perturbations of the filtration or metric input (Cohen-Steiner et al. 2007), but those guarantees do not imply classifier robustness. Robustness is therefore treated as an empirical question, and negative results and known failure modes will be reported.

## 1 Background and Motivation

Activation-based monitoring is one approach to detecting strategic deception, scheming, and sandbagging. Goldowsky-Dill et al. (2025) report AUROC ranging from 0.96 to 0.999 across several settings. At thresholds calibrated to a 1% false-positive rate, the probes detected 95–99% of deceptive responses.

Subsequent work identifies important limitations. Kumar (2026) reports substantial degradation under stylistic distribution shift, with partial recovery after style-augmented training. Gupta and Jenner (2025) show that fine-tuning optimized against a monitor can reduce the effectiveness of token-level linear probes, while sequence-level monitors that aggregate information across tokens degrade less. These findings motivate a controlled comparison of linear and topological monitors under matched distribution-shift and probe-targeted evaluations.

## 2 Related Work and Differentiation

Persistent homology (PH) — a tool from topological data analysis that summarizes the global shape of a point cloud or graph via how connected components, loops, and higher-dimensional voids appear and disappear across a filtration — has been applied to transformer attention patterns in peer-reviewed work, including for text classification (Kushnareva et al. 2021) and attention-head pruning (Perez and Reinauer 2022). Two recent applications are the closest precedents for this project without directly overlapping with it.

Bazarova et al. (2026) (TOHA) compute a zero-dimensional persistence divergence between the attention-graph subgraphs induced by a prompt and its response, and use it — training-free, on a handful of identified attention heads — to detect hallucination. This is, to our knowledge, the most directly relevant attention-graph-topology method, but it targets factual groundedness rather than strategic deception, and its prompt-vs-response framing does not directly transfer to comparing two possible responses (one honest, one deceptive) to the same prompt.

Fay et al. (2026) apply persistent homology to activation point clouds across six models (3.8B–70B parameters) under two adversarial conditions — indirect prompt injection, and *sandbagging via backdoor fine-tuning* (van der Weij et al. 2024) on LoRA-tuned, password-locked models — and report a consistent *topological compression* signature. In the Locked-vs-Elicited sandbagging condition, they report 100% classification accuracy from PH features for the evaluated layers on Mistral-7B and LLaMA3-8B (Appendix C.4). This is the closest methodological precedent for Track B, but it evaluates one adversarial condition rather than a matched comparison with linear probes across broader strategic-deception settings. Separately, Fay et al. report (their Appendix G) that their topological signature persists against LLMail-Inject adversarial examples explicitly designed to evade a *different* activation-based defense (TaskTracker) — precedent for the narrower cross-monitor evaluation in Stage 3a, though not evidence of robustness to an attacker who directly targets the topological monitor, which we scope separately (Stage 3b) given its materially higher compute cost.

We are not aware of a published study that directly compares topological and linear deception detectors under matched distribution-shift and monitor-targeted evaluations. That is the gap this project addresses. We note, and will explicitly test rather than assume away, that TDA features are not unconditionally robust: they are known to be sensitive to filtration and vectorization choices, and differentiable topology layers have been used to construct adversarial attacks against models trained on persistence features (Gabrielsson et al. 2020). Any robustness claim this project produces will be reported alongside the attack surface it was and was not tested against.

## 3 Proposed Method

Figure 1 summarizes both tracks below.

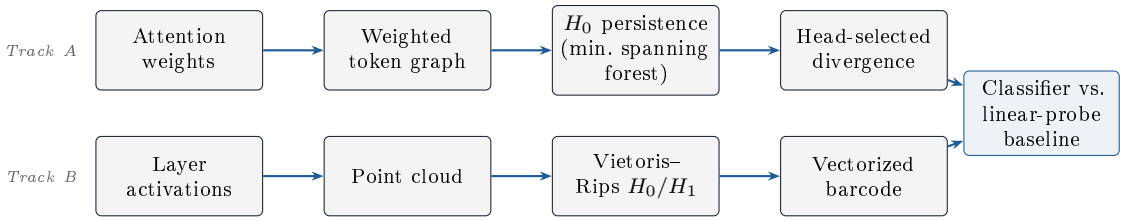


**Figure 1.** Planned pipelines for the core Track A evaluation and the conditional Track B pilot. Track A adapts TOHA's attention-graph divergence to the honest-vs-deceptive setting, scored per response. Track B-core follows the activation-point-cloud methodology previously evaluated in a locked-versus-elicited sandbagging condition, scored per batch of rollouts (Section 4); a per-response Track B extension is scoped separately as a stretch goal (Section 3.2).

### 3.1 Track A — Attention-Graph Topological Divergence

Following TOHA, decoder attention maps are lower-triangular under causal masking; Bazarova et al. use the available causal-attention pseudo-distance $1 - w_{ij}$ for each attended pair to construct a *complete undirected weighted graph* over tokens — every pair of tokens receives a single edge weight, with the graph treated as a *non-metric space* (their Proposition 3.1) precisely because these weights need not satisfy the metric axioms, not because the graph itself is directed. We use the same undirected construction as TOHA. After setting prompt-to-prompt edge weights to zero, the total length of the zero-dimensional persistence barcode of this graph equals the weight of the minimum spanning forest attaching the response vertices to the prompt vertices — a quantity therefore cheap to compute without general-purpose persistent-homology software:

$$\text{MTop-Div}(R, P) \;=\; \sum_i |d_i - b_i| \;=\; \text{MSF-weight}\big(R \to P\big).$$

We adapt the comparison target: instead of prompt-vs-response divergence (hallucination), we compute the same divergence separately for a model's honest and deceptive responses to matched prompts, and search — via the same greedy, small-labeled-sample head-selection procedure used in TOHA (order 50–100 labeled examples, $N_{\max} \leq 10$ heads) — for attention heads whose divergence separates honest from deceptive generations. All attention-head selection, threshold selection, and other data-dependent choices will be performed using prompt-disjoint training and validation splits. Held-out prompts will be reserved for final evaluation.

### 3.2 Track B — Activation Point-Cloud Persistence

Fay et al.'s classifier does not score individual rollouts: it scores *subsamples* — batches of many rollouts' activations, aggregated into one barcode summary per batch. Concretely, in Fay et al., one point in a point cloud is the last-token residual-stream activation vector of a single input at a fixed layer $\ell$. They construct subsamples from two experimental conditions — clean versus poisoned, or locked versus elicited — compute Vietoris–Rips persistence diagrams ($H_0$, $H_1$), vectorize each into a 41-dimensional barcode summary, and classify subsamples by experimental condition. In our adaptation, the corresponding batch-level prediction target is honest versus deceptive provenance.

Because this differs from the per-response observational unit of a linear probe, Track B will be compared with a probe aggregated over the same batches.

**Track B-core (population-level monitor, adapted from Fay et al.).** We adapt their batch-level observational unit and barcode-feature pipeline, strengthening the evaluation with prompt-disjoint splitting and initially reducing computational scale rather than reproducing their configuration unchanged. One detection instance is a batch of $k$ rollouts believed to share a provenance (honest or deceptive), represented by a persistence-summary vector. Their configuration uses $k = 4096, 64$ subsamples per condition, $H_0/H_1$, and five selected layers; our initial configuration treats $k$, the subsample count, and the evaluated layers as resource-calibrated parameters, reporting convergence with respect to each before considering experiments at the source scale. For a fair comparison, the linear-probe baseline is aggregated to the same unit — the mean continuous per-rollout probe score across the batch, with the fraction flagged at a validation-selected threshold reported as a sensitivity analysis — so that both methods are evaluated as batch classifiers on identical batches, using ripser for pilot experiments and Ripser++ if GPU acceleration is required at larger point-cloud sizes. This construction is directly useful for auditing ("does this evaluation transcript or deployment log, as a population, look like it was generated under a deceptive regime?"); Fay et al. evaluated the analogous construction in a locked-versus-elicited sandbagging condition.

The batch classifier will first be fitted on pure-condition batches. It will then be evaluated without retraining on fixed-size mixed-provenance batches at deceptive prevalence $p \in \{0, 0.1, 0.25, 0.5, 0.75, 1\}$. Detector-score distributions and their relationship with deceptive prevalence will be reported as an exploratory contamination-sensitivity analysis. No separate classifier will be fitted at each prevalence level.

**Track B-stretch (per-response monitor, novel).** Time and mentee interest permitting, we explore a local construction that could score a single rollout the way a probe does — for instance, a token-activation-trajectory point cloud drawn from one rollout's own sequence of layer-wise activations, or a diagram measuring one rollout's local topological distance to a reference honest-response manifold. This extension is novel relative to Fay et al. and will be treated as exploratory.

To reduce overlapping-subsample and prompt-identity leakage, prompts will be partitioned before point-cloud or subsample construction. No prompt or activation vector may contribute to both training and evaluation diagrams. We will also evaluate activation-normalized and norm-matched variants to test whether apparent separation is explained by simple scale differences.

Track A will be compared with simple controls based on response length and aggregate attention statistics, including mean attention entropy. Track B will be compared with activation norm and pairwise-distance summaries computed over the same batches. Label-permutation null tests will be used to verify that measured separation exceeds chance.

## 4 Experimental Design

**Stage 1: In-distribution evaluation.** Adapt and evaluate Track A and the linear-probe baseline on one primary dataset and one open-weight model. The initial candidates are Apollo Research's

released deception-detection benchmark and a compatible 7B to 9B open-weight model, with the final model selected during the week 1 compatibility audit. If the core activation and evaluation pipeline is stable, begin a reduced-scale Track B pilot using matched batch-level probe aggregation.

**Stage 2: Distribution shift.** Evaluate Track A without retraining on one primary out-of-distribution dataset. The primary out-of-distribution dataset will be selected from MASK (Ren et al. 2025), the DeceptionBench dataset introduced by Ji et al. (2025), or Liars' Bench (Kretschmar et al. 2025), based on label compatibility with the Stage 1 task. Extend the evaluation to Track B only if the Stage 1 pilot is complete and computationally stable.

**Stage 3a: Cross-monitor robustness.** Apply a prespecified reduced-scale probe-targeted attack and compare the change in AUROC for Track A and the attacked linear probe. TPR at a validation-selected false-positive rate will be reported as a secondary metric. Trainable components will be evaluated across repeated seeds, with uncertainty estimated using prompt-grouped bootstrap intervals. Attack results will be interpreted only when task performance and response quality remain within prespecified tolerances, distinguishing selective monitor evasion from general model degradation. Extension to Track B is conditional on progress and compute.

**Stage 3b: Direct monitor-targeted attack.** If time and compute permit, directly optimize against the topological monitor. This is the experiment required to evaluate resistance to monitor-specific optimization and is not part of the guaranteed core deliverable.

Results from all completed stages will be reported regardless of outcome. A finding that topological features add no predictive or robustness value beyond the matched baselines would still define an informative limitation of the approach.

## 5 Feasibility and Resources

Core datasets and reference implementations are publicly available, although access, licensing, preprocessing requirements, and model compatibility will be verified during onboarding. Track A and the linear-probe baseline are expected to be feasible on one modern GPU at the proposed pilot scale.

Fay et al.'s full global persistent-homology pipeline used four 80GB A100 GPUs. If initiated, Track B will begin with one 7B to 9B model, fewer selected layers, smaller subsamples, and $H_0$ diagrams before adding $H_1$. Runtime and memory will be benchmarked before fixing batch size, homology dimension, and subsample count.

Stage 3a will be scoped to the available single-GPU budget. Stage 3b is conditional on additional compute and is not required for the core deliverable. Additional project funding would increase the breadth of the evaluation but is not required for the project to produce a valid core result.

## 6 Indicative Timeline (12 Weeks)

- **Weeks 1 to 2:** Onboarding, data audit, prompt-disjoint splits, and adaptation of the linear-probe baseline.
- **Weeks 3 to 5:** Track A implementation, head selection, and baseline validation.
- **Weeks 6 to 7:** In-distribution evaluation, error analysis, and basic confound checks.
- **Weeks 8 to 9:** Distribution-shift evaluation. Begin the reduced-scale Track B pilot only if the core Track A pipeline is complete.
- **Weeks 10 to 11:** Run the probe-targeted robustness experiment and consolidate the core Track A results. Continue Track B only if the core evaluation is complete. A direct topology-targeted attack will begin only if all other deliverables are finished.
- **Week 12:** Final analysis, technical report, repository cleanup, and Demo Day preparation.

## 7 Mentor Background

I am an independent researcher working in topological data analysis and persistent homology, with a background in applied machine learning and software development. My work spans computational biology and AI interpretability. I developed TopoGeoML, a preregistered topology-aware machine-learning toolkit and empirical study released through PyPI and Zenodo, and earned a Silver Medal in the CAFA-6 protein function prediction benchmark, finishing 38th of 2,259 teams.

For this project, I will guide the topological and experimental design, including filtration choices, persistence-diagram interpretation, evaluation methodology, and failure analysis. The mentee will lead implementation and day-to-day experimentation and will contribute substantially to methodological decisions.

## References


Alexandra Bazarova, Andrei Volodichev, Aleksandr Yugay, Andrey Shulga, Alina Ermilova, Konstantin Polev, Julia Belikova, Rauf Parchiev, Dmitry Simakov, Maxim Savchenko, Andrey Savchenko, Serguei Barannikov, and Alexey Zaytsev. Hallucination detection in LLMs with topological divergence on attention graphs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15449–15470, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.704. URL https://aclanthology.org/2026.acl-long.704/. arXiv:2504.10063.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007. doi: 10.1007/s00454-006-1276-5.

Aideen Fay, Inés García-Redondo, Qiquan Wang, Haim Dubossarsky, and Anthea Monod. The shape of adversarial influence: Characterizing LLM latent spaces with persistent homology, 2026. ICLR 2026 Oral; arXiv:2505.20435v3.

Rickard Brüel Gabrielsson, Bradley J. Nelson, Anjan Dwaraknath, and Primoz Skraba. A topology layer for machine learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1553–1563. PMLR, 2020. URL https://proceedings.mlr.press/v108/gabrielsson20a.html.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception with linear probes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 19755–19786. PMLR, 2025. URL https://proceedings.mlr.press/v267/goldowsky-dill25a.html. arXiv:2502.03407.

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

Jiaming Ji, Wenqi Chen, Kaile Wang, Donghai Hong, Sitong Fang, Boyuan Chen, Jiayi Zhou, Juntao Dai, Sirui Han, Yike Guo, and Yaodong Yang. Mitigating deceptive alignment via self-monitoring, 2025. Introduces the DeceptionBench dataset; arXiv:2505.18807.

Kieron Kretschmar, Walter Laurito, Sharan Maiya, and Samuel Marks. Liars' bench: Evaluating lie detectors for language models, 2025. arXiv:2511.16035.

Sachin Kumar. Pressure-testing deception probes in LLMs: Scaling, robustness, and the geometry of deceptive representations. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 472–489, San Diego, California, USA, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.gem-main.43. URL https://aclanthology.org/2026.gem-main.43/.

Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635–649, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.50. URL https://aclanthology.org/2021.emnlp-main.50/. arXiv:2109.04825.

Ilan Perez and Raphael Reinauer. The topological BERT: Transforming attention into topology for natural language processing, 2022. arXiv:2206.15195.

Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Geralnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. The MASK benchmark: Disentangling honesty from accuracy in AI systems, 2025. Center for AI Safety / Scale AI; arXiv:2503.03750.

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