USEFUL FEATURES, BACKWARD SCORES: OOD IN LANGUAGE-MODEL TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detectors prioritize inputs for closer inspection. Yet features that distinguish input groups need not yield a useful anomaly ranking. We analyze this gap in language-model trajectories under text-length control and fixed score directions. On Spam development data, an input adaptation of D^2HScore falls from raw AUROC 0.919 to 0.530 after length matching. On length-matched, held-out HateSpeech inputs, the same features yield AUROC 0.644 for a labeled linear classifier but 0.444 for an ID-fitted distance score. ToxicChat shows the same contrast. Feature-selection and backbone controls retain the main reversal pattern. Frozen Civil Comments and TweetEval irony tests also reverse (0.467 and 0.435), extending the finding beyond toxicity. In these contrasts, anomalous groups have farther centers but tighter spread. A labeled, fixed-center feature-space intervention changes rankings: equalizing spread helps some tasks and harms others. OOD evaluation must check the chosen score's ranking even when its features distinguish the classes.

1 INTRODUCTION

A distance-based toxicity detector uses non-toxic examples as its reference and ranks inputs for closer inspection by their distance from that reference. If toxic inputs receive lower scores, they fall behind non-toxic ones in that ranking, even when the features distinguish the two groups. This makes the score itself an evaluation target: informative features alone do not establish useful detection. We study this gap for language-model trajectories, which summarize changes in internal representations across layers. We test whether apparent detection survives length matching and whether the predefined distance score ranks the designated OOD inputs in the intended direction.

Length is the first problem. On Spam development data, Input-D^2H gives raw AUROC 0.919, but only 0.530 after matching normal and anomalous inputs within shared token-length bins (Table 1). This score adapts D^2HScore from generated responses to input tokens. Other scores respond differently to matching, changing which methods appear useful. Length control therefore changes the assessment of these input-domain scores.

After matching, a second problem remains. On held-out HateSpeech inputs, a linear classifier trained on labeled examples gives matched AUROC 0.644, while distance from the in-distribution (ID) reference gives a below-chance 0.444 using the same features (Figure 1A). ToxicChat shows the same contrast. These same-source results let us ask why features that distinguish the classes give the wrong distance ranking.

Why does the distance score fail? On ToxicChat and HateSpeech, the out-of-distribution (OOD) center is farther from the ID reference, but the OOD group is tighter. The detector scores individual points, not group centers: a tight group can have smaller distances for many of its members (Figure 1). We test spread's role by expanding or shrinking OOD feature points at fixed centers. Rankings change in both directions, and equalizing spread does not consistently repair them.

Our analysis of fixed trajectory scores does not propose a new detector. Length matching, same-feature readouts and spread tests at fixed centers expose when features and rankings disagree. New-data tests bound the result. This extends backward-score analysis (Nalisnick et al., 2019) to LM trajectories and makes score ordering a separate evaluation target.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

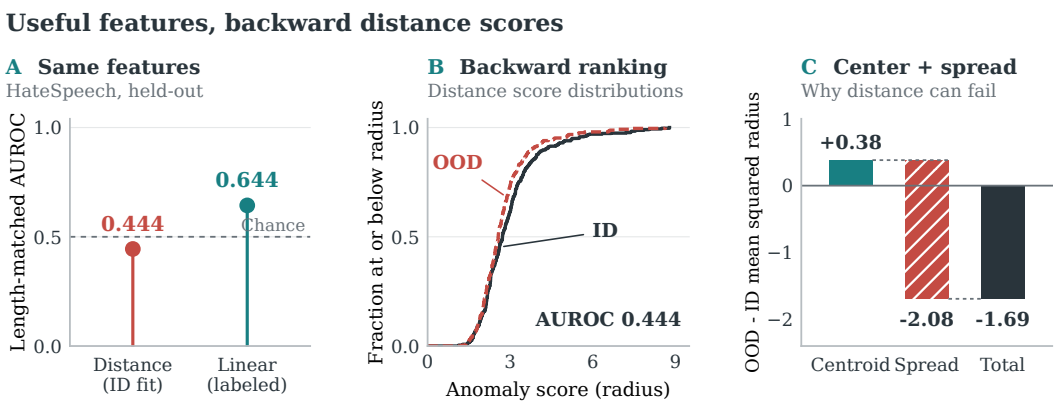


Figure 1: **Useful features can produce a backward distance ranking.** All panels use the same length-matched, held-out HateSpeech inputs. **A:** A linear classifier trained on development labels gives above-chance AUROC; the ID-fitted distance score ranks the classes backward. These are descriptive same-source point estimates. **B:** OOD distances are smaller more often; a higher curve means more examples at or below that radius. **C:** Tighter OOD spread outweighs the farther center in mean squared distance. Curves and AUROCs average five matched subsets; repeated rows are not new observations. All task–split comparisons and distance-score uncertainty appear in Figure 3 and Table 20.

2 Representations and Scores

Here OOD means different from the deployment reference, not absent from the model's pretraining data. ToxicChat uses non-toxic conversations as ID and toxic conversations as OOD; HateSpeech, Spam and Jailbreak use their corresponding normal/anomalous labels. The reference geometry is fit on ID examples only. We test whether the score ranks the designated OOD class as more anomalous.

Each input becomes a point whose coordinates summarize its trajectory. Four features describe changes in direction between adjacent layers (alignment and churn). Two track how variation across token positions changes across layers. Two compare the directions and relative sizes of attention and multilayer perceptron (MLP) outputs. One measures alignment at a fixed early layer. These nine features were chosen from 21 candidates using exploratory development labels; Appendix D gives every definition and selection decision. Let $Z(x)$ denote the selected coordinates standardized using only ID references, with reference mean μ and sample covariance Σ. The fixed trajectory radius is

$$s_{\text{traj}}(x) = \sqrt{q(Z(x))}, \qquad q(z) = (z-\mu)^\top P(z-\mu), \quad P = (\Sigma + 10^{-3} I)^{-1}. \tag{1}$$

The detector treats larger radius as more anomalous. AUROC is the probability that an OOD example receives a higher score than an ID example, counting ties as one half. A value of 0.5 is chance; below 0.5, the ranking is reversed. Any strictly increasing transformation of radius preserves AUROC: the problem is the ordering, not the numerical scale. Choosing the score sign on labeled validation data makes the method semi-supervised. Choosing it on test labels measures how well the classes can be separated, but does not evaluate the predefined detector (Appendix C).

3 Analysis Design

3.1 Tasks and Model

The development suite contains ToxicChat (Lin et al., 2023), Jailbreak (jackhhao, n.d.), the binary Tweets Hate Speech Detection collection (tweets-hate-speech-detection, n.d.), and SMS Spam (Almeida et al., 2011). Qwen3-0.6B is the primary backbone; four additional backbones span 360M–4B parameters and four model families in total. Appendix F gives model details, partitions and source limitations.

3.2 Development and confirmation sets

Each task has disjoint ID-reference, validation, development and confirmation partitions. The 200–250 ID references alone fit standardization and precision. Feature selection uses 200 development examples per class, greedily deleting from 21 candidates to maximize class separation while allowing either score sign. The search uses development labels to choose a sign; evaluation keeps the sign fixed. The complete path and tie rule appear in Appendix D.

The representation and extractor are frozen before confirmation analysis. Trajectory analysis excludes inputs shorter than six tokens; the development comparison scores retain them. All inputs are capped at 512 tokens. Appendix F documents class counts and extraction checks.

3.3 Length-matched evaluation

Input length differs sharply between several classes. We group inputs into shared token-length bins and sample equal ID and OOD counts within each bin containing both classes. We compute AUROC over all retained examples, then average across the fixed matching seeds. Labels determine which examples to retain, not their scores or sign. Matching balances counts within bins, not exact lengths, and changes the evaluated population. In Table 1, each method is matched on its own valid inputs, so subsets can differ across methods. Appendix F gives the bin rule, seeds and retained counts; Appendix G tests sensitivity to bin count.

3.4 Readouts and comparison scores

Same-feature readouts. One L2-regularized logistic classifier is trained jointly on length-matched ToxicChat and HateSpeech development inputs, using the nine ID-standardized coordinates after reference centering. A scalar control instead applies logistic regression to a cubic-spline transformation of radius with eight ID-reference quantile knots and linear extrapolation. Both use identical training rows, labels and weights, with equal weight for each source–class group and fixed regularization ($C = 1$). Source training is matched with seed 42. We evaluate their decision scores on disjoint confirmation inputs, with no evaluation-set tuning.

Length-comparison scores. The length comparison adds three input-domain adaptations: CED-1C uses one-class oracle/auxiliary distances (Lee et al., 2024), Input-D^{2}H uses dispersion and attention-guided drift (Ding et al., 2025), and Input-TV uses layerwise ID-distance volatility (Wang et al., 2024). Appendix E specifies their changed source settings, equations and fixed hyperparameters. Each score keeps its predefined OOD direction.

3.5 Feature-selection controls and prospective tests

Four leave-one-task-out (LOTO) folds rerun the full 21-to-3 elimination path on three development tasks, then evaluate the excluded task with larger radius fixed as OOD. Its labels do not influence subset search, although candidate design remains historical. A retrospective cohort includes every cache meeting the recorded schema, ID-reference and class-count criteria: 11 evaluations, one sharing the development source. Cross-model analysis reuses the four task contrasts.

Civil Comments (Borkan et al., 2019) defines ID as toxicity ≤ 0.10 and OOD as toxicity ≥ 0.80. An initial frozen validation study was inconclusive. A separate protocol then fixed the features, orientation, references and evaluation quotas before official-test feature extraction. We report both stages without pooling; Appendix B records their full chronology and protocol hashes.

4 Results

4.1 Length differences inflate apparent performance

Does a strong raw score distinguish anomalous texts from normal texts of similar length? On development data, Input-D^{2}H falls from 0.836 to 0.455 on Jailbreak and 0.919 to 0.530 on Spam after matching (Table 1). Strong raw separation does not survive comparison within shared length

ranges. CED-1C drops less and retains signal on Spam; Input-TV improves on some tasks. Length matching changes which scores appear useful.

Table 1: **Length matching changes apparent detection performance.** Raw / length-matched AUROC on Qwen3-0.6B development data, computed separately for each method's valid inputs. Comparison scores retain inputs shorter than six tokens; trajectory radius excludes them. Matching changes the sample, not scores or their sign. Comparison methods are input-domain adaptations.

Task	CED-1C	Input-D^2H	Input-TV	Trajectory radius
ToxicChat	0.566/0.520	0.678/0.504	0.311/0.411	0.419/0.420
Jailbreak	0.839/0.626	0.836/0.455	0.300/0.533	0.765/0.850
HateSpeech	0.561/0.573	0.551/0.514	0.395/0.415	0.379/0.395
Spam	0.890/0.828	0.919/0.530	0.528/0.805	0.828/0.848

Matching changes the evaluated population, not the texts. The trajectory analysis retains only 30% of valid Spam inputs, so its matched result describes shared lengths rather than the full source population. This comparison diagnoses sensitivity to class-length differences and sample composition; it is not a causal text-length intervention. For trajectory radius, matching preserves the positive Jailbreak and Spam rankings but leaves the ToxicChat and HateSpeech reversals. Length control exposes one problem without resolving the other.

4.2 Useful features, backward scores

A linear classifier trained on ToxicChat and HateSpeech development labels gives matched AUROCs 0.656 and 0.644 on held-out examples from those sources. With the same nine features and evaluation rows, distance from the ID reference gives 0.457 and 0.444. Figure 1 shows the HateSpeech example: the features support classification, but their distance score ranks the anomalous class backward.

A nonlinear readout of radius trained on the same labeled rows and weights reaches 0.537 and 0.547 on the two confirmation tasks, below the vector readout's point estimates.

The comparison establishes useful within-source class information. On the new Civil source, the vector classifier reaches 0.555 $[0.497, 0.582]$; its paired gain over the scalar readout is 0.031 $[-0.030, 0.078]$, so transfer remains inconclusive. Other distance scores also have task-dependent strengths (Appendix H). We now test whether the trajectory reversal depends on the original feature selection, backbone or evaluation examples.

4.3 Does the reversal persist?

4.3.1 Does reversal survive feature and metric choices?

The reversal could depend on choosing features with development labels. Rerunning subset selection while excluding each evaluation task preserves all 4/4 AUROC directions (Table 2). ToxicChat and HateSpeech remain below chance; Jailbreak and Spam remain above it. The excluded task's labels do not drive selection, although the candidate features were designed on the original suite.

All-candidate and fixed-sign checks preserve the point-estimate directions, and ToxicChat and HateSpeech remain below chance under every tested precision geometry. Two fixed-sign intervals cross chance; Jailbreak confirmation can reverse under metric changes. Appendices B and G give the full controls and uncertainty.

4.3.2 Does the pattern transfer across backbones?

Five backbones spanning four model families keep AUROC on the same side of chance in 19/20 comparisons. The exception is SmolLM2 ToxicChat, whose AUROC is at chance (Appendix Table 5). These checks reuse the same task contrasts and therefore test backbone sensitivity, not independent task replication.

Table 2: **Reversal after feature-selection control and on a new source.** LOTO selects subsets without the named task's labels. Civil Comments uses the separately frozen official-test protocol. Entries show fixed-orientation AUROC and conditional stratified-bootstrap 95% intervals. Appendix B gives feature counts and the full center–spread decomposition.

Evaluation	AUROC [95% CI]
LOTO ToxicChat	0.430 [0.378, 0.498]
LOTO Jailbreak	0.777 [0.701, 0.838]
LOTO HateSpeech	0.431 [0.382, 0.478]
LOTO Spam	0.818 [0.755, 0.884]
Civil Comments test	0.467 [0.448, 0.487]

4.3.3 Does reversal persist on new data?

Disjoint same-source confirmation. Held-out examples test whether the original samples drove the result. All four AUROC point estimates retain their development directions, but the rankings are weaker. The intervals cross chance for ToxicChat and Jailbreak, making these two confirmations inconclusive (Appendix Tables 3 and 20).

Separately frozen Civil Comments test. Civil Comments tests reversal on another source of toxicity data. Matched AUROC is 0.467 $[0.448, 0.487]$, compared with 0.496 raw. The earlier validation stage was inconclusive; we froze the official-test protocol after that result, before extracting test features, and report both stages in Appendix B. The test supports reversal among texts in length bins containing both classes.

The below-chance ranking persists across the recorded matching-bin choices and reference-refit checks; after matching, length alone has near-chance AUROC (Appendices G.2 and G.4).

Independent irony contrast. We froze a separate TweetEval irony test before extracting its official test inputs (Appendix B.6). With non-ironic tweets as ID and ironic tweets as OOD, matched radius AUROC is 0.435 $[0.396, 0.474]$ on 310 examples per class. A within-source, label-trained linear readout on the same features reaches 0.614 $[0.579, 0.660]$; the prespecified ID-only 10-neighbor distance also reverses (Appendix B.6). The OOD center is farther but its spread is tighter. This extends the observed geometry beyond toxicity without providing an ID-only correction.

Where reversal does not replicate. After excluding the overlapping JailbreakClassification cache, ten earlier cached evaluations give five intervals above chance and five inconclusive, with no further significant reversal in that cohort. Appendix B retains the full cohort and overlap audit. Reversal is a finding about particular task contrasts, not a prediction for every OOD class.

4.4 How does spread affect the score?

The reversals are not explained by a nearer OOD center. All four OOD centers are farther from the reference than their ID counterparts on both development and confirmation splits. ToxicChat and HateSpeech instead have smaller within-group spread (Figure 1C). The score measures each point's distance from the reference, not its distance from its own group center.

A tight cluster and a wide cluster can have the same center but different distances from the reference. To separate these effects, we write the mean squared distance as a center term plus a spread term. For an evaluation group G, let ν_G be its center (centroid) and C_G its covariance, computed with the group size as the divisor. The standard mean-square identity gives:

$$\mathbb{E}_G[q(Z)] = \underbrace{(\nu_G - \mu)^\top P(\nu_G - \mu)}_{\text{centroid displacement}} + \underbrace{\operatorname{tr}(PC_G)}_{\text{within-group dispersion}} . \tag{2}$$

The first term measures the group center's distance from the reference; the second measures spread around that group center. Subtracting ID from OOD gives $\Delta_{\text{rad}} = \Delta_{\text{cent}} + \Delta_{\text{disp}}$. A farther OOD center can coexist with smaller mean squared distances if the OOD points cluster tightly enough. The

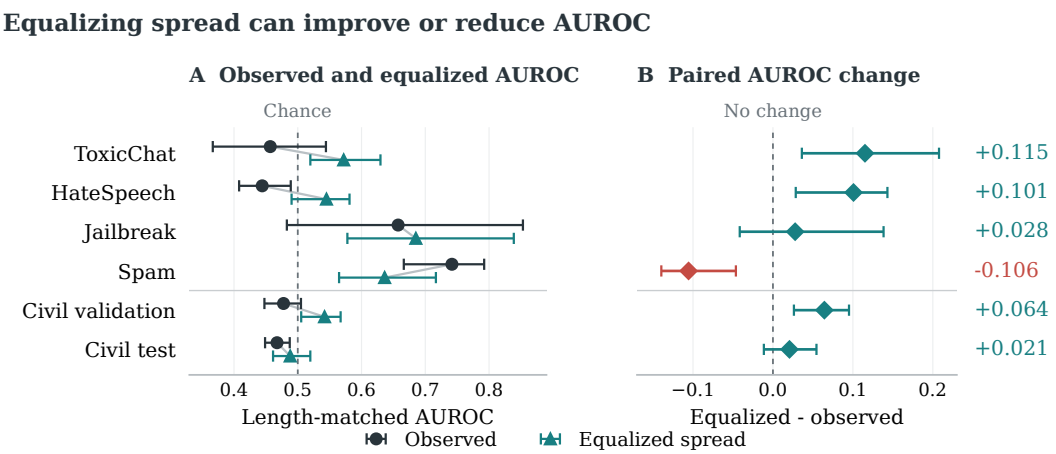


Figure 2: **Equalizing spread can improve or reduce AUROC.** All six specified cases are shown: four same-source confirmation tasks above the divider, then the two Civil stages. **A:** Observed and equalized AUROC. **B:** Their paired difference. Intervals use 1,000 stratified paired bootstrap replicates, rematching and recomputing equalization in each replicate while holding the reference fit fixed. These are retrospective, pointwise intervals. Spam loses useful separation. Civil test has a positive mean squared-radius difference after equalization, but its AUROC point estimate remains below chance.

reference geometry stays fixed, and no Gaussian class assumption is needed. Appendix C gives the derivation.

Across all eight task–split comparisons, the sign of the mean squared-radius difference agrees with whether AUROC is above or below chance (Figure 3). This is an empirical finding, not a consequence of the identity: averages need not determine pairwise ranking. Squaring the radius preserves AUROC, but mean radius and mean squared radius can order the groups differently. We now test spread's effect on the individual rankings directly.

4.4.1 Changing spread without moving the center

We move every OOD feature point away from or toward its own center by one common factor, leaving every ID point in place. Evaluation labels identify the groups. This diagnostic covers the four confirmation tasks and both Civil stages. We fixed the procedure after inspecting the original results but before computing its outcomes. Within each matched subset, OOD features become

$$Z'_{\mathrm{OOD}} = \nu_{\mathrm{OOD}} + \alpha(Z_{\mathrm{OOD}} - \nu_{\mathrm{OOD}}), \qquad \alpha_{\mathrm{eq}} = \sqrt{\frac{\mathrm{tr}(PC_{\mathrm{ID}})}{\mathrm{tr}(PC_{\mathrm{OOD}})}}. \tag{3}$$

Scales $\alpha > 1$ spread the OOD points apart; scales $\alpha < 1$ pull them together. The OOD center stays fixed, and each point keeps its direction from that center. The reference fit, matched rows and ID scores are unchanged. The spread term scales by α^2; at α_{eq}, its OOD-minus-ID contribution is zero. This equalization scale depends only on class spread, not on which scale gives the best AUROC. The experiment changes feature points, not text, and may produce features that no text would generate. It cannot serve as a deployable correction or identify a linguistic cause.

Equalization raises ToxicChat confirmation from 0.457 to 0.572 and HateSpeech from 0.444 to 0.545 (Figure 2). Paired changes are $+0.115$ $[0.036, 0.208]$ and $+0.101$ $[0.029, 0.143]$; HateSpeech's equalized interval still crosses chance. On Spam, reducing the larger OOD spread lowers AUROC from 0.742 to 0.636, a change of -0.106 $[-0.140, -0.046]$.

Civil test separates mean-square correction from ranking recovery. Equalization makes the OOD-minus-ID mean squared-radius difference positive ($+0.183$), but mean radius remains lower

(-0.0094). AUROC reaches 0.488, with paired change $+0.021$ $[-0.011, 0.054]$. Correcting the mean squared distance does not establish an AUROC improvement: AUROC compares individual scores, not means. Jailbreak's change is also inconclusive; Civil validation improves without changing its original verdict. Appendix G.1 gives all values, scale curves and rank checks.

5 Related Work

Distance-based OOD detection scores deviations from an ID reference (Lee et al., 2018; Sun et al., 2022); CED adds target–oracle and target–auxiliary relations (Lee et al., 2024). Nalisnick et al. (2019) report higher likelihood for OOD images and analyze location, variance and curvature. Background statistics and complexity provide further explanations (Ren et al., 2019; Serrà et al., 2020). We examine ranking failure in LM trajectories, separating length sensitivity from reversal after matching.

DoSE estimates the ID density of model statistics (Morningstar et al., 2021). Mahalanobis++ studies feature norms and ℓ_2 normalization (Müller & Hein, 2025); Janiak et al. (2026) study spectral structure, local intrinsic dimensionality and ID-guided radial normalization. Compact-ID objectives change representations during training (Ruff et al., 2018). We fix the trajectory feature map and reference geometry, then test how class spread affects rankings.

TV-Score and D^{2}HScore use trajectories during generation (Wang et al., 2024; Ding et al., 2025); RAUQ and TrajGuard also require generation-time information (Vazhentsev et al., 2026; Liu et al., 2026). Our Input-TV and Input-D^{2}H adapt this setting to input tokens (Appendix E).

Length bias affects generation uncertainty (Santilli et al., 2025). OOD shift type changes method comparisons on NLP tasks (Arora et al., 2021), while NLP-ADBench finds no single method dominates across its eight datasets (Li et al., 2025). We test length sensitivity in input-domain scores, then use the same-feature readout comparison to show why length control alone does not validate an anomaly ranking.

6 Discussion and Limitations

The evaluation lesson extends beyond this radius: an OOD claim concerns the ordering produced by a prespecified score, not just information in its features. Our analysis exposes two distinct ways such a claim can be overstated. Compare raw and length-matched performance to see whether apparent success survives among texts of similar length. Then evaluate the score itself: useful class information and a farther OOD center do not guarantee the intended ranking. The spread intervention shows why correcting the mean squared distance is insufficient. Ranking depends on individual score comparisons.

ID-only scores remain useful on other task contrasts. Alternative readouts help some cases without a general repair (Appendices B.1 and G.2). Retrospective unit normalization of ID-centered features, with a new ID-only reference fit, raises Civil test matched AUROC from 0.467 to 0.521 $[0.497, 0.545]$; the paired change is $+0.053$ $[+0.023, +0.084]$. The normalized interval still includes chance, so this fixed control does not establish above-chance recovery.

Limitations. We diagnose score behavior without identifying why classes contract. Linguistic composition, collection source and model behavior remain possible causes; jailbreak labels in particular differ in source composition. Candidate-feature design is historical, the cache cohort is retrospective, and backbone checks reuse the original task contrasts.

Matching retains only shared length bins, leaving small samples for some tasks. Intervals condition on feature design and assume independent rows rather than source clusters; intervention intervals also hold the reference fit fixed. The feature-space intervention uses evaluation labels and may leave the set of realizable text features. Results cover the tested trajectory probe in 360M–4B models, not frontier systems or all LM representations.

Conclusion. Length matching can remove apparent success while leaving backward rankings. The same trajectory features can support within-source classification and produce below-chance distance rankings. Fixed-center spread interventions connect this mismatch to the score's geometry. Our

analysis makes the scoring rule a separate evaluation target: test its intended ranking under length control instead of inferring detection from useful features alone.

Ethics Statement

OOD detectors can affect moderation, abstention, and access decisions. A reversing score can create false reassurance or disproportionate false positives. We release selected numeric features and per-sample scores without text or source identifiers, distinguish supervised calibration from one-class detection, and recommend deployment only after task-representative validation. Source datasets include toxic and hateful language; we release no new examples.

Reproducibility Statement

The anonymous code supplement contains a verifier, fixed scoring utilities, and numeric caches for the length-matching contrast, Civil Comments and irony checks, spread intervention, readouts and endpoint comparison. It recomputes their selected point estimates, regenerates the irony AUROC bootstrap, and checks included paired intervals from saved draws. It does not rerun model extraction or every appendix analysis; full model-level replication needs the external datasets and checkpoints described in Appendix F. Appendix C gives the score identities and their assumptions.

AI-Use Statement

GitHub Copilot assisted with implementation and testing, figure and table generation, manuscript and submission editing, literature and provenance checks, methodological critique, retrospective control design, mathematical argument review, and interpretation of analysis outputs. The authors remain responsible for the scientific design, verification, and conclusions.

References

Tiago A. Almeida, José María Gómez Hidalgo, and Akebo Yamakami. Contributions to the study of SMS spam filtering: New collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, pp. 259–262. ACM, 2011. doi: 10.1145/2034691.2034742. URL https://doi.org/10.1145/2034691.2034742.

Udit Arora, William Huang, and He He. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10687–10701, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.835. URL https://aclanthology.org/2021.emnlp-main.835/.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://aclanthology.org/2020.findings-emnlp.148/.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500. ACM, 2019. doi: 10.1145/3308560.3317593. URL https://doi.org/10.1145/3308560.3317593.

Yue Ding, Xiaofang Zhu, Tianze Xia, Junfei Wu, Xinlong Chen, Qiang Liu, and Liang Wang. D^2HScore: Reasoning-aware hallucination detection via semantic breadth and depth analysis in LLMs. *arXiv preprint arXiv:2509.11569*, 2025. doi: 10.48550/arXiv.2509.11569. URL https://arxiv.org/abs/2509.11569.

jackhhao. Jailbreak Classification. Hugging Face dataset card, n.d. URL https://huggingface.co/datasets/jackhhao/jailbreak-classification. Accessed September 16, 2026.

Denis Janiak, Jakub Binkowski, and Tomasz Kajdanowicz. A geometry-based view of Mahalanobis OOD detection. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62433.

Hakyung Lee, Keon-Hee Park, Hoyoon Byun, Jeyoon Yeom, Jihee Kim, Gyeong-Moon Park, and Kyungwoo Song. CED: Comparing embedding differences for detecting out-of-distribution and hallucinated text. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14866–14882, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.874. URL https://aclanthology.org/2024.findings-emnlp.874/.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 7167–7177, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Yuangang Li, Jiaqi Li, Zhuo Xiao, Tiankai Yang, Yi Nian, Xiyang Hu, and Yue Zhao. NLP-ADBench: NLP anomaly detection benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2464–2474. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.133. URL https://aclanthology.org/2025.findings-emnlp.133/.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4694–4702, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.311. URL https://aclanthology.org/2023.findings-emnlp.311/.

Cheng Liu, Xiaolei Liu, Xingyu Li, Bangzhou Xin, and Kangyi Ding. TrajGuard: Streaming hidden-state trajectory detection for decoding-time jailbreak defense. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13371–13388. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.655. URL https://aclanthology.org/2026.findings-acl.655/.

Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3232–3240. PMLR, 2021. URL https://proceedings.mlr.press/v130/morningstar21a.html.

Maximilian Müller and Matthias Hein. Mahalanobis++: Improving OOD detection via feature normalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45151–45184. PMLR, 2025. URL https://proceedings.mlr.press/v267/muller25a.html.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.09136.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e79596878b2320cac26dd792a6c51c9-Abstract.html.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402. PMLR, 2018. URL https://proceedings.mlr.press/v80/ruff18a.html.

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 743–759, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-short.60. URL https://aclanthology.org/2025.acl-short.60/.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1909.11480.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840. PMLR, 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

tweets-hate-speech-detection. Dataset card for Tweets Hate Speech Detection. Hugging Face, n.d. URL https://huggingface.co/datasets/tweets-hate-speech-detection/tweets_hate_speech_detection. Dataset curator: Roshan Sharma; card contributor: Darshan Gandhi. Accessed September 16, 2026.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. SemEval-2018 task 3: Irony detection in english tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pp. 39–50. Association for Computational Linguistics, 2018. doi: 10.18653/v1/S18-1005. URL https://aclanthology.org/S18-1005/.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Mrinmaya Sachan, Preslav Nakov, Timothy Baldwin, and Artem Shelmanov. Efficient hallucination detection for LLMs using uncertainty-aware attention heads. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61700.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F. Wong, Zhuosheng Zhang, and Rui Wang. Embedding trajectory for out-of-distribution detection in mathematical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 42965–42999, 2024. doi: 10.52202/079017-1360. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4b734e95f0788a030a69caa987516186-Abstract-Conference.html.

A Guide to the Supporting Evidence

The supporting material addresses five questions raised by the main results:

- **Does the pattern replicate?** Appendix B gives disjoint confirmation, the full geometric comparison, cross-model results, leave-one-task-out selection and both Civil Comments stages. The cached-cohort analysis retains every eligible result and identifies the overlapping source.
- **What does changing spread establish?** Appendix G.1 reports every intervention outcome, the full scale response and paired uncertainty at fixed class centers and reference geometry.
- **Could selection or length explain the result?** Appendix D specifies every feature and selection step; Appendix G tests reference fitting, precision, feature subsets, data draws and matching bins.
- **Where do other scores work better?** Appendix H compares endpoint, logit-based and trajectory scores, including the intent/topic settings where endpoint distance wins.
- **What methods and limits are documented?** Appendix C proves the score identities, Appendix E defines comparison scores, and Appendix F documents data provenance, compute and archive scope.

B Replication and Feature-Selection Controls

Feature groups are A (layerwise alignment and churn), D (token spread), F (attention–MLP balance), and G (fixed-depth alignment). ADFG-9 denotes the selected nine features; ADG-7 removes the two F features.

B.1 Disjoint confirmation and the full geometric comparison

Table 3: **Same-source held-out confirmation after feature development.** "ID-only" fixes larger Mahalanobis distance as OOD. "Val. sign" is selected on a disjoint approximately 50/50 validation split; the resulting score is semi-supervised. n reports valid examples before length matching.

Task	$n_{\text{ID}}/n_{\text{OOD}}$	ID-only	Val. sign	Val.-calibrated
ToxicChat	97/96	0.457	-	0.543
Jailbreak	50/50	0.657	+	0.657
HateSpeech	297/299	0.444	-	0.556
Spam	295/300	0.742	+	0.742
Task mean	—	0.575	—	0.625

Choosing the radius direction using the separate validation labels raises four-task mean confirmation AUROC from 0.575 to 0.625. All four selected signs transfer, but ToxicChat reaches only 0.543. This is semi-supervised calibration, not a change to the features.

Figure 3 compares the centroid and dispersion contributions across both splits. Table 4 gives the exact values underlying this visual comparison.

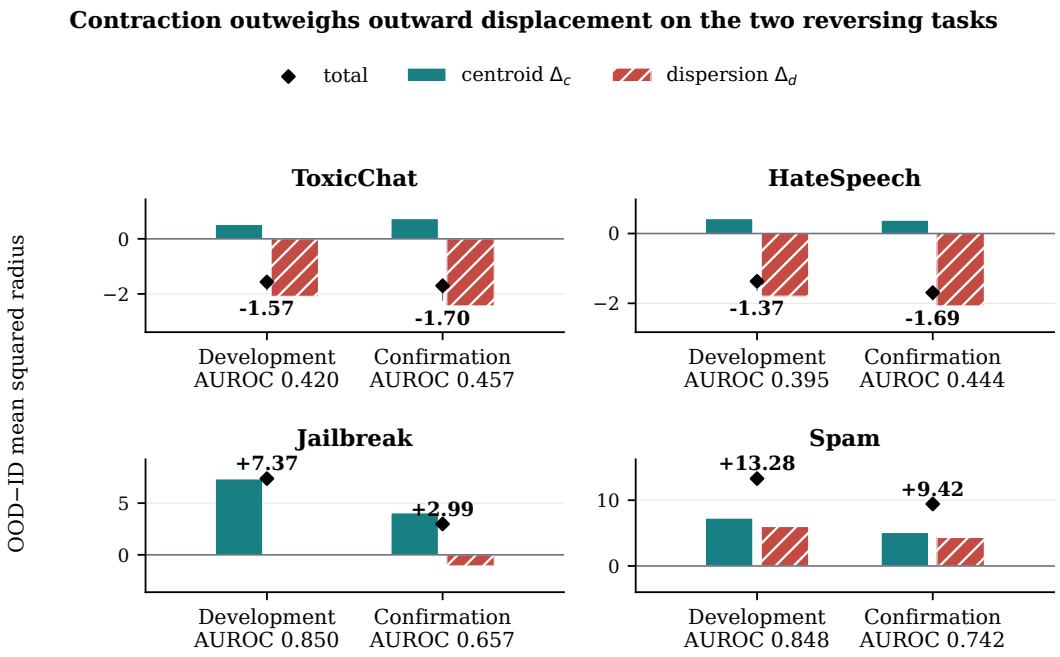


Figure 3: **The full geometric comparison.** Solid teal bars show the OOD-minus-ID change in centroid contribution; hatched red bars show the dispersion contribution. Diamonds mark their sum, the change in mean squared radius. ToxicChat and HateSpeech contract enough to outweigh outward displacement on both splits. Each task has its own vertical scale. AUROCs below the split names are measured separately on the same length-matched examples, with larger radius fixed as OOD.

Table 4: **Exact decomposition of task-dependent radial direction.** Five-seed length-matched means are shown for the development and disjoint confirmation splits. Positive Δ_c means the OOD centroid is farther; negative Δ_d means the OOD cloud is tighter in the ID precision geometry. $\Delta\mathbb{E}[q] = \Delta_c + \Delta_d$.

	Development			Confirmation		
Task	Δ_c	Δ_d	$\Delta\mathbb{E}[q]$	Δ_c	Δ_d	$\Delta\mathbb{E}[q]$
ToxicChat	+0.522	-2.087	-1.566	+0.734	-2.437	-1.702
Jailbreak	+7.347	+0.027	+7.374	+4.064	-1.078	+2.986
HateSpeech	+0.427	-1.795	-1.368	+0.382	-2.075	-1.693
Spam	+7.281	+6.003	+13.284	+5.089	+4.335	+9.424

B.2 Cross-model geometry and attention–MLP features

Table 5 reports AUROC and $\Delta\mathbb{E}[q]$ together. The mean squared-radius difference matches the four-task sign pattern in all 20 cells. AUROC stays on the same side of chance in 19, with SmolLM2 ToxicChat at 0.500. Across the 20 cells, $\Delta\mathbb{E}[q]$ and AUROC correlate at Spearman $\rho = 0.89$. The two attention–MLP features (Category F) improve Jailbreak on every backbone ($+0.069$ to $+0.242$) but have small or mixed effects elsewhere. Their contribution depends on the task.

Table 5: **Cross-model radial geometry.** Each task cell is fixed AUROC/$\Delta\mathbb{E}[q]$. Δ_F is the Jailbreak ADFG-9 minus ADG-7 AUROC.

Model	Toxic	Jailbreak	Hate	Spam	Δ_F
Qwen3-0.6B	0.420/-1.57	0.850/+7.37	0.395/-1.37	0.848/+13.28	+0.223
SmolLM2-360M-Instruct	0.500/-0.20	0.744/+4.87	0.402/-1.43	0.765/+13.05	+0.069
Gemma-3-1B	0.460/-1.03	0.685/+2.04	0.452/-0.43	0.798/+16.10	+0.190
Llama-3.2-1B	0.419/-1.61	0.704/+4.50	0.424/-0.40	0.868/+12.22	+0.116
Qwen3-4B	0.454/-1.19	0.786/+4.50	0.405/-0.25	0.834/+20.83	+0.242

B.3 Leave-one-task-out feature discovery

Each fold repeats the full greedy path using only the other three tasks. The excluded task supplies neither the selection objective nor its task-specific discovery sign. Table 6 reports held-out AUROC with the score direction fixed. Each bootstrap interval stays on the same side of chance as the corresponding development estimate, although ToxicChat's upper endpoint is 0.498. Global-label permutation tests give maximum Holm-adjusted $p = 0.010$ but do not preserve the class–length association. Within-bin tests, which preserve class counts in each length bin, give maximum adjusted $p = 0.016$ (Appendix G.2).

Table 6: **Leave-one-task-out geometry.** $|F|$ is the feature count selected without the evaluation task. AUROC uses larger radius as OOD and gives a pointwise stratified-bootstrap 95% interval.

Held out	$\lvert F\rvert$	AUROC [95% CI]	Δ_c	Δ_d	$\Delta\mathbb{E}[q]$
ToxicChat	8	0.430 [0.378, 0.498]	+0.478	-1.693	-1.215
Jailbreak	10	0.777 [0.701, 0.838]	+6.364	-2.531	+3.833
HateSpeech	12	0.431 [0.382, 0.478]	+0.546	-2.032	-1.486
Spam	10	0.818 [0.755, 0.884]	+6.995	+6.257	+13.252

Fixed-orientation selection sensitivity. The LOTO search measures class separability, allowing either score sign during selection. In a post hoc sensitivity check, we repeat the greedy path while maximizing fixed-direction AUROC on the other three tasks. Table 7 shows that all four held-out AUROC estimates remain on the same side of chance and the mean squared-radius differences keep their signs. ToxicChat and Jailbreak have intervals crossing chance. Both selection checks reuse the same evaluation examples.

Table 7: **Fixed-orientation selection sensitivity.** The search objective J_{fixed} is mean fixed-orientation AUROC on the three training tasks. Held-out AUROC retains larger radius as OOD and gives a pointwise 95% interval.

Held out	$\lvert F\rvert$	J_{fixed}	Held-out AUROC [95% CI]	$\Delta\mathbb{E}[q]$
ToxicChat	6	0.716	0.463 [0.414, 0.541]	-0.855
Jailbreak	6	0.613	0.539 [0.473, 0.616]	+0.192
HateSpeech	6	0.711	0.384 [0.330, 0.436]	-1.120
Spam	3	0.593	0.734 [0.651, 0.811]	+1.917

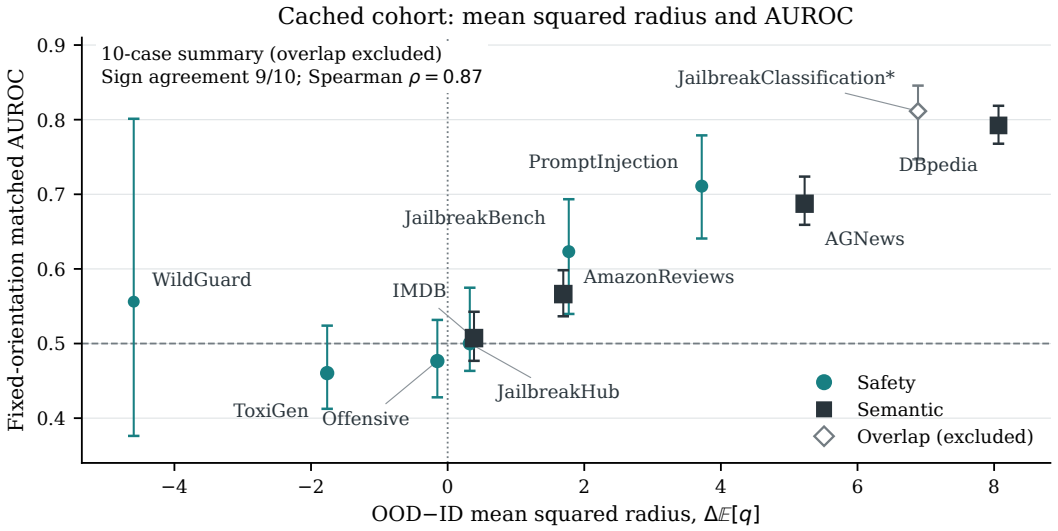
B.4 Additional cached evaluations

After excluding JailbreakClassification, which overlaps the development source, five of ten eligible evaluations have intervals above chance and five are inconclusive. None supplies another significant reversal. The mean squared-radius sign agrees with AUROC direction in 9/10 cases, with Spearman $\rho = 0.87$; WildGuard is the sole disagreement and retains only 14 matched examples per class. Table 8 also shows the excluded row, giving six above-chance and five inconclusive intervals, 10/11 sign agreement and $\rho = 0.89$ when included. The underlying audit recorded ineligible caches,

which are not part of the minimal code supplement. Removing the known overlap does not establish independence among the remaining sources.

Table 8: **Additional cached evaluations.** The asterisk marks a shared development source with exact feature overlap, excluded from the ten-evaluation summary. Ref. is the ID-only reference count; Eval. is ID/OOD before matching; n_m is retained per class. Inference uses the conditional AUROC interval.

Dataset	Cohort	Ref.	Eval.	n_m	AUROC [95% CI]	Δ_c	Δ_d	$\Delta\mathbb{E}[q]$	Inference
Offensive	Safety	198	196/199	174	0.477 [0.428, 0.532]	+0.12	-0.27	-0.15	Inconclusive
ToxiGen	Safety	195	196/200	174	0.460 [0.413, 0.524]	+0.05	-1.82	-1.76	Inconclusive
WildGuard	Safety	200	200/200	14	0.556 [0.376, 0.801]	+1.98	-6.58	-4.60	Inconclusive
JailbreakBench	Safety	200	200/100	85	0.623 [0.540, 0.693]	+3.49	-1.71	+1.77	Positive
JailbreakClassification*	Safety	258	259/500	134	0.811 [0.747, 0.846]	+5.51	+1.38	+6.89	Positive
JailbreakHub	Safety	200	200/200	155	0.500 [0.463, 0.575]	+0.54	-0.21	+0.32	Inconclusive
PromptInjection	Safety	171	172/203	87	0.711 [0.641, 0.779]	+1.04	+2.68	+3.72	Positive
AGNews	Semantic	500	500/500	442	0.687 [0.659, 0.724]	+3.17	+2.05	+5.23	Positive
AmazonReviews	Semantic	500	500/500	464	0.566 [0.537, 0.598]	+0.63	+1.06	+1.69	Positive
DBpedia	Semantic	500	500/500	396	0.792 [0.768, 0.819]	+9.36	-1.29	+8.06	Positive
IMDB	Semantic	500	500/500	454	0.507 [0.477, 0.543]	+0.62	-0.24	+0.39	Inconclusive



* Shown for context; excluded from the 10-case summary.

Figure 4: **Mean squared-radius differences and AUROC across the cached cohort.** The headline summary excludes JailbreakClassification, shown as a hollow diamond and marked with an asterisk because it overlaps development. Circles and squares distinguish safety and semantic tasks; marker size reflects matched sample count. Error bars are stratified-bootstrap 95% intervals. WildGuard is the sole sign disagreement and has the smallest sample. Table 8 includes the overlapping row.

B.5 Civil Comments validation and test

Civil validation was inconclusive; the separately evaluated official test ranks OOD below ID (Table 9). The validation candidate pool initially targeted 5,000 OOD examples, but only 1,633 met toxicity ≥ 0.80. Before model loading, we allowed every eligible OOD example into the candidate pool while keeping thresholds, quotas, seed, features, orientation and decision criteria fixed. After observing validation, we doubled the official-test evaluation quota based on class-count feasibility, before extracting test features; all other choices remained fixed. We report both stages without pooling. The global-label two-stage adjusted $p = 0.0024$ does not condition on length; a within-bin permutation check gives $p = 0.004$. Both A/D/G-only controls retain the negative direction.

Table 9: **Two-stage prospective Civil Comments result.** Validation uses 250 ID references and 500 evaluation examples per class; official test uses 500 and 1,000. Only ADFG-9 rows carry bootstrap intervals because they define the prospectively specified primary criterion.

Stage	Features	AUROC [95% CI]	Δ_c	Δ_d	$\Delta\mathbb{E}[q]$	Verdict
Validation	ADFG-9	0.478 [0.449, 0.506]	+0.137	-1.863	-1.726	Inconclusive
Official test	ADFG-9	0.467 [0.448, 0.487]	+0.183	-0.529	-0.346	Confirmed
Official test	ADG-7	0.471	+0.094	-0.301	-0.207	Robustness
Official test	ADG-5	0.456	+0.051	-0.514	-0.463	Robustness

B.6 Independent TweetEval irony check

On held-out irony tweets, the fixed distance score reverses even though the same features support a within-source labeled readout. TweetEval's binary irony task (Barbieri et al., 2020; Van Hee et al., 2018) defines non-irony as ID and irony as OOD. Source revision `b3a375baf0f409c77e6bc7aa35102b7b3534f8be` supplies 2,862 official train, 955 validation and 784 test rows. Before seeing target scores, we fixed the nine features, Qwen3-0.6B checkpoint, positive radius direction, one ID-only alternative and all decision thresholds. After the six-token filter and within-source duplicate checks, we drew 250 train ID references and disjoint 300/class train examples for a labeled logistic diagnostic. We did not use the validation split. The test contains 472 ID and 310 OOD examples; all five length-matching seeds retain 310/class. A checksum-verified overlap audit completed after scoring found zero matches against 6,149 texts from the other datasets evaluated here; the sampled rows did not change. Eight feature and token-length parity checks passed before target extraction.

The ID-only alternative is mean distance to the ten nearest references in the ID-standardized nine-dimensional space; larger scores denote OOD for both scores. Raw radius, 10-neighbor and linear-readout AUROCs are 0.412, 0.397 and 0.610; matched values are 0.435 $[0.396, 0.474]$, 0.415 $[0.379, 0.459]$ and 0.614 $[0.579, 0.660]$. Matched length-only AUROC is 0.500. The neighbor-minus-radius difference is -0.020 $[-0.039, 0.005]$. All intervals use 1,000 within-class paired test-row draws, rematching each draw with references and readout fixed. The prespecified ID-only success criterion required neighbor AUROC at least 0.65 and a gain of at least 0.10, among other conditions; neither target was reached. No alternate score or source was tried in response.

Applying Equation 2 after this result gives matched differences of $+0.273$ $[0.156, 0.412]$ for center displacement, -1.748 $[-2.458, -0.993]$ for spread and -1.475 $[-2.149, -0.725]$ for mean squared radius, with the same fixed-reference row bootstrap. This post hoc geometry check does not identify a linguistic cause or change the failed ID-only criterion. Intervals here are pointwise and do not account for sequential task choice, repeated authors or reference uncertainty. The irony subset has no stated license and is subject to Twitter terms; no tweet text is redistributed.

C Score Orientation and Mean Squared Radius

Proposition C.1 (Centroid–dispersion decomposition)**.** *For a standardized random vector Z, a fixed reference mean μ and $P \succeq 0$, Equation 2 holds for any group with finite second moments. Its empirical form computes evaluation covariance with the group size as the divisor, regardless of the reference covariance estimator.*

Proof. Write $Z - \mu = (Z - \nu_G) + (\nu_G - \mu)$, expand the quadratic and take expectations. The cross term vanishes because $\mathbb{E}_G[Z - \nu_G] = 0$; the centered quadratic has expectation $\mathrm{tr}(PC_G)$. □

Definition C.2 (Fixed-orientation detection and separability)**.** For independent X_{ID} and X_{OOD}, define

$$A(s) = \Pr[s(X_{\mathrm{OOD}}) > s(X_{\mathrm{ID}})] + \tfrac{1}{2}\Pr[s(X_{\mathrm{OOD}}) = s(X_{\mathrm{ID}})]. \tag{4}$$

Fixed-orientation detection evaluates the predefined s using $A(s)$. Allowing either score sign measures class separation as $S(s) = \max\{A(s), A(-s)\}$.

Proposition C.3 (Orientation is part of the detector)**.** *For every scalar score, $A(-s) = 1 - A(s)$ and $S(s) = \frac{1}{2} + |A(s) - \frac{1}{2}|$. Choosing between s and $-s$ with evaluation labels measures separability. It cannot evaluate the predefined detector.*

Proof. Negation exchanges greater-than and less-than events while preserving ties. Their probabilities sum to one, with half of each tie assigned to either AUROC. □

D Exact Trajectory Features and Selection Provenance

For each adjacent layer pair, the probe averages tokenwise cosine over positions $1, \ldots, T-1$, excluding the final input position:

$$c_l(x) = \frac{1}{T-1} \sum_{t=1}^{T-1} \cos\left(h_l^{(t)}, h_{l+1}^{(t)}\right). \tag{5}$$

Let $a_l^{(t)}$ and $m_l^{(t)}$ denote the attention and MLP outputs at layer l and position t. The trajectory radius uses:

$$\texttt{hs_cos_mean} = \text{mean}_l\, c_l, \qquad \texttt{hs_cos_min} = \min_l c_l, \tag{6}$$

$$\texttt{hs_churn_late} = \text{mean}_{l \in \text{last third}}(1 - c_l), \tag{7}$$

$$\texttt{hs_churn_ratio} = \frac{\text{mean}_{l \in \text{last third}}(1 - c_l)}{\text{mean}_{l \in \text{first third}}(1 - c_l) + 10^{-10}}. \tag{8}$$

For representation spread, define

$$q_l = \frac{1}{d} \sum_{j=1}^{d} \text{Std}_{t=1,\ldots,T-1}\left(h_{l,j}^{(t)}\right). \tag{9}$$

The tokenwise standard deviation uses sample normalization (denominator $T-2$ for $T-1$ positions). Then `repr_spread_slope` is the least-squares slope of q_l against layer index, and `repr_spread_std` is its population standard deviation across layers.

For components, define layerwise averages

$$A_l = \text{mean}_{t<T}\, \|a_l^{(t)}\|_2, \qquad M_l = \text{mean}_{t<T}\, \|m_l^{(t)}\|_2, \tag{10}$$

$$R_l = \frac{A_l}{A_l + M_l + 10^{-10}}, \qquad U_l = \text{mean}_{t<T} \cos(a_l^{(t)}, m_l^{(t)}). \tag{11}$$

The retained features are `comp_align_mean`$= \text{mean}_l\, U_l$ and `comp_attn_ratio_shift`, the last-third mean of R_l minus its first-third mean. Finally, `hs_cos_at_p17` is $c_{\lfloor L/6 \rfloor}$.

These definitions are used for every reported trajectory-radius result. Category A is a token-averaged inter-layer cosine, and churn is defined as $1 - c_l$.

D.1 Exact exploratory selection procedure

Candidate pool and order. The search used the following 21 coordinates; the displayed order is also the deterministic tie-breaking order:

- A: `hs_cos_mean`, `hs_cos_std`, `hs_cos_min`, `hs_cos_max`, `hs_cos_range`, `hs_churn_early`, `hs_churn_late`, `hs_churn_ratio`;
- D: `repr_spread_slope`, `repr_spread_std`, `repr_spread_ratio`;
- F: `comp_attn_ratio_mean`, `comp_attn_ratio_std`, `comp_align_mean`, `comp_align_std`, `comp_attn_ratio_shift`;
- G: `hs_cos_at_p17`, `hs_cos_at_p33`, `hs_cos_at_p50`, `hs_cos_at_p67`, `hs_cos_at_p83`.

For completeness, these names exhaust the following deterministic operators. A contains mean, population standard deviation, minimum, maximum, and range of $(c_l)_l$, followed by first-third and last-third means of $1 - c_l$ and their ratio. D contains the least-squares slope and population standard deviation of $(q_l)_l$, and $q_{\text{final}}/(q_{\text{first}} + 10^{-10})$. F contains the mean and population standard deviation of $(R_l)_l$, the mean and population standard deviation of $(U_l)_l$, and the last-third minus first-third mean of R_l. G samples c_l at clipped indices $\lfloor \rho L \rfloor$ for $\rho \in \{1/6, 1/3, 1/2, 2/3, 5/6\}$. Every third has size $\max(1, \lfloor L/3 \rfloor)$.

Objective. For candidate subset F, task t, and its positive Mahalanobis radius $r_{t,F}$, development labels first determine

$$\sigma_{t,F} = \begin{cases} +1, & \text{AUC}_{\text{raw}}(r_{t,F}) \geq \frac{1}{2}, \\ -1, & \text{AUC}_{\text{raw}}(r_{t,F}) < \frac{1}{2}. \end{cases} \tag{12}$$

The search objective is

$$J(F) = \frac{1}{4} \sum_{t=1}^{4} \frac{1}{5} \sum_{b \in \{42,123,456,789,1337\}} \text{AUC}_{\text{LM},b}(\sigma_{t,F} r_{t,F}), \tag{13}$$

where $\text{AUC}_{\text{LM},b}$ is ten-bin length-matched AUROC for matching seed b. Thus J measures development-set class separation after choosing the more favorable score sign. It is not an ID-only detector estimate.

Elimination and stopping. Starting from all 21 coordinates, at each iteration we evaluate all one-coordinate deletions and retain the subset with largest J. Exact ties choose the first feature in the current candidate order because replacement occurs only for a strict improvement. We continue to three coordinates, including steps where J declines, then select the global maximum along the nested path. The maximum is $J = 0.720565$ at nine coordinates. Table 10 gives every deletion; the selected nine are exactly those defined above. After this search, the feature set is frozen and the favorable development signs are discarded. Every ID-only result uses the rule that larger radius receives a higher anomaly score.

Table 10: **Complete greedy backward-elimination trace.** "Deleted" is the coordinate removed to reach that row; J is Equation 13. The search continued through three features and selected the global path maximum.

Features	Deleted	J
21	—	0.693
20	comp_align_std	0.699
19	hs_cos_std	0.704
18	comp_attn_ratio_std	0.708
17	hs_cos_at_p83	0.710
16	hs_cos_at_p67	0.712
15	comp_attn_ratio_mean	0.714
14	hs_churn_early	0.715
13	repr_spread_ratio	0.716
12	hs_cos_max	0.716
11	hs_cos_range	0.718
10	hs_cos_at_p33	0.718
9	hs_cos_at_p50	0.721 ← selected
8	repr_spread_std	0.719
7	hs_cos_at_p17	0.711
6	repr_spread_slope	0.712
5	hs_cos_mean	0.699
4	hs_churn_late	0.702
3	hs_churn_ratio	0.687

Both $\sigma_{t,F}$ and $J(F)$ use the same development labels, so the selected feature set is exploratory. The nine features are not independently hypothesized tests and require evaluation on the frozen confirmation data.

D.2 CATEGORY-F ABLATION WITH FIXED SCORE DIRECTION

Table 11: **Fixed-direction Category-F ablation.** All rows use larger Mahalanobis radius as OOD. ADG-7 removes the two selected component features; ADG-5 is a separately identified five-feature A/D/G subset.

Subset	Toxic	Jailbreak	Hate	Spam	Mean
ADFG-9	0.420	0.850	0.395	0.848	0.628
ADG-7	0.437	0.627	0.405	0.849	0.579
ADG-5	0.449	0.774	0.387	0.850	0.615

Removing F from the selected representation lowers mean AUROC from 0.628 to 0.579, primarily because Jailbreak falls from 0.850 to 0.627; Spam is unchanged and the two reversed tasks remain reversed. The compact ADG-5 subset reaches 0.615. These fixed-direction AUROCs are lower than the selection objective in Table 10 because J chooses a favorable sign per task. Category F is empirically useful for this jailbreak shift; the observed gain does not establish universal transfer or length independence.

Table 12: **Category-F gain under progressively narrower length bins.** Values average fixed-direction AUROC across five matching seeds and four tasks; Δ_J is the ADFG-9 minus ADG-7 gain on Jailbreak.

B	ADFG-9	ADG-7	Δ	Δ_J
5	0.626	0.578	+0.048	+0.223
8	0.633	0.592	+0.042	+0.199
10	0.628	0.579	+0.049	+0.223
15	0.627	0.583	+0.044	+0.206
20	0.633	0.594	+0.040	+0.181

The overall F gain remains 0.040–0.049 and the Jailbreak gain remains 0.181–0.223 as B increases from 5 to 20. The effect therefore survives progressively narrower length matching, although residual within-bin length effects cannot be ruled out.

Category-F length control. In a synthetic truncation sweep at $T \in \{32, 64, 128, 256, 512\}$ with source text held fixed, the Category-F attention-ratio mean and standard deviation have $|r| = 0.63$ and 0.71 with length. They remain range-bounded, but they are not length-independent. Length matching controls the evaluated length distribution; it does not turn these features into invariant quantities.

E INPUT-DOMAIN COMPARISON METHODS

E.1 CED-1C

For target embedding $e(x)$, CED first selects the three nearest ID oracle examples. For each oracle, it ranks ID auxiliary candidates by cosine opposition between $e(x) - e(a)$ and $e(x) - e(o)$, keeps the 100 least similar, and selects the five farthest. Text pairs are concatenated as "target. auxiliary," matching the released implementation. Let $D_c(z)$ be squared Mahalanobis distance to the selected class reference. Our OOD-positive adaptation is

$$s_{\mathrm{CED}}(x) = D_c(x) - 0.7\,\mathbb{E}_o\left[D_c(o)\right] + 0.2\,\mathbb{E}_{o,a}\left[D_c(x \oplus a) - D_c(o \oplus a)\right]. \tag{14}$$

The source implementation uses negative squared distance as ID confidence; this equation is its OOD-oriented negation. We change the source setting to one ID class, replace RoBERTa/OpenAI embeddings with Qwen, and use final-token decoder pooling. The source paper and code are available at https://aclanthology.org/2024.findings-emnlp.874/ and https://github.com/MLAI-Yonsei/CED.

E.2 INPUT-D^{2}H

At each layer, let $\bar{h}_l = T^{-1} \sum_{t=1}^{T} h_l^{(t)}$ be the mean input-token state. The dispersion component is

$$D_l = \frac{1}{T} \sum_{t=1}^{T} \|h_l^{(t)} - \bar{h}_l\|_2. \tag{15}$$

For drift, attention from the final input token is averaged across heads, the top 50% token positions are selected, and their mean representation g_l is formed. Drift is $\mathrm{mean}_l \|g_{l+1} - g_l\|_2$. We standardize the two components with ID-train means and standard deviations and average them equally, as the source gives equal weights but does not specify a portable normalization estimator. The source interprets larger components as faithful-response confidence, so our fixed OOD score is its negative. Applying generated-response equations to input tokens and choosing ID-only normalization make this an adaptation.

E.3 INPUT-TV

For each input, let $y_l = T^{-1} \sum_t h_l^{(t)}$. ID training inputs provide coordinatewise μ_l and σ_l^2. Following the paper's $k = 0$ equation and the released code's diagonal distance mapping, we compute

$$d_l = \left\| \frac{(y_l - \mu_l)^2}{\max(\sigma_l^2, 10^{-10})} \right\|_2, \qquad s_{\mathrm{TV}} = \frac{1}{L-1} \sum_{l=1}^{L-1} |d_{l+1} - d_l|. \tag{16}$$

The square, variance floor and division apply coordinatewise. The norm of squared deviations follows the released mapping; it is not the usual diagonal Mahalanobis distance. Published TV-Score averages hidden states over generated decoding positions from an ID-fine-tuned model and also studies higher-order smoothing. Our score uses input positions, a pretrained model, and $k = 0$ fixed before evaluation.

F DATA AND REPRODUCIBILITY

Reference scaling. Trajectory radius uses `StandardScaler` fitted on ID references only: subtract each coordinate's mean and divide by its population standard deviation, using scale one for constant coordinates. The sample covariance and ridge in Equation 1 apply after this transformation.

Confirmation and matching. Confirmation starts after the first 200 test examples per class, with equal quotas set by the smaller remainder. Filtering can leave unequal counts. The six-token cutoff is an extraction criterion, not a requirement of the layer-third definitions. Main-analysis bins use pooled linear percentiles and half-open intervals, with one added to the final edge. Equal lengths stay together; bins missing either class are skipped. Each seed initializes one generator used across bins. Each backbone uses its own truncated token counts, so matched rows can differ. Appendix H gives the separate comparison protocols.

We divide pooled token counts into $B = \min(10, \max(1, \lfloor \min(n_{\mathrm{ID}}, n_{\mathrm{OOD}})/10 \rfloor))$ equal-frequency bins. Development uses ten bins; confirmation uses nine for ToxicChat, five for Jailbreak, and ten otherwise. Each matched subset combines the retained bins. We compute AUROC on the whole subset, then average across seeds $\{42, 123, 456, 789, 1337\}$. Trajectory-radius development retains 246, 172, 364, and 118 total examples for ToxicChat, Jailbreak, HateSpeech and Spam, respectively.

Table 13: **Checksummed disjoint data partitions.** Feature selection uses 200 development examples per class; the remaining balanced source partition is reserved for confirmation.

Task	Source	ID	OOD	ID train	ID val	ID test	OOD val	OOD test
ToxicChat	lmsys/toxic-chat	non-toxic	toxic	250	50	500	50	300
Jailbreak	jackhhao/jailbreak-classification	benign	jailbreak	200	50	250	50	400
HateSpeech	tweets_hate_speech_detection	normal	hate	250	50	500	50	500
Spam	ucirvine/sms_spam	ham	spam	250	50	500	50	500

All split files carry SHA256 checksums and are verified before extraction. An optimized confirmation extractor agrees with the development implementation to maximum absolute error below 2.4×10^{-8}

on parity examples. The comparison-method run took 708.5 seconds on one RTX 3090; confirmation took 166.1 seconds. The project environment used Python 3.13.7, PyTorch 2.9.1, and Qwen3-0.6B in bfloat16. The implementation saves resumable feature caches, aggregate JSON, and per-sample NPZ scores.

Source audit and overlap. HateSpeech uses the binary Tweets Hate Speech Detection collection. It does not use Davidson's three-class corpus. Both jailbreak names use the training split of `jackhhao/jailbreak-classification`, whose benign and jailbreak examples have different collection origins. Length control does not remove that source confound. Hashing all 21 cached trajectory coordinates finds 175 exact ID evaluation-row matches and 194 OOD matches between JailbreakClassification and development, plus 17 cohort ID-evaluation matches to the fitted reference. These counts measure feature identity. These caches lack per-row text hashes and pinned dataset revisions, so exact text-level disjointness and the separate crossover resample cannot be certified. Civil Comments has pinned revisions and text hashes. Its official-test OOD pool has four duplicate rows, and the intervals do not account for duplicate clusters.

Archive scope. The code supplement provides cached numeric inputs for the selected analyses and paired estimates for the intervention, readout and endpoint interval checks. It excludes raw text, model weights, feature extraction and inputs for other appendix sensitivity analyses; end-to-end reruns require the external datasets and checkpoints described above.

G Dispersion Intervention and Robustness Checks

G.1 Fixed-centroid dispersion intervention

The intervention fixes all six cases, nine features, five matching seeds, 1,000 bootstrap replicates, RNG seed 20260916, and residual scales $\{0.5, 0.75, 1, 1.25, 1.5, 2\}$. These choices were set after observing the unmodified scores but before computing intervention outcomes. For each matched subset, Equation 3 preserves the OOD centroid and multiplies precision-weighted population dispersion by α^2. The implementation transforms unstandardized coordinates; affine ID standardization makes this equivalent to the standardized expression. Tests verify unchanged ID rows/scores, centroid preservation, the dispersion law and baseline parity with the unmodified scores. The equalized mean-square difference equals the centroid contribution to numerical precision in every case.

Table 14: **Every prespecified intervention outcome.** The first four rows use confirmation data. n_m is retained per class; $\bar{\alpha}_{\mathrm{eq}}$ is the mean subset-specific equalization factor. AUROCs are five-seed means. Intervals are pointwise 95% paired-bootstrap intervals. The final column is the equalized OOD-minus-ID mean squared radius.

Evaluation	n_m	$\bar{\alpha}_{\mathrm{eq}}$	Observed	Equalized [95% CI]	Change [95% CI]	$\Delta_{\mathrm{rad,eq}}$
ToxicChat	62	1.164	0.457	0.572 [0.520, 0.630]	+0.115 [+0.036, +0.208]	+0.734
HateSpeech	251	1.121	0.444	0.545 [0.490, 0.581]	+0.101 [+0.029, +0.143]	+0.382
Jailbreak	18	1.093	0.657	0.685 [0.578, 0.839]	+0.028 [-0.042, +0.138]	+4.064
Spam	93	0.789	0.742	0.636 [0.565, 0.717]	-0.106 [-0.140, -0.046]	+5.089
Civil validation	451	1.098	0.478	0.542 [0.505, 0.567]	+0.064 [+0.026, +0.095]	+0.137
Civil test	872	1.028	0.467	0.488 [0.461, 0.520]	+0.021 [-0.011, +0.054]	+0.183

Each bootstrap replicate resamples evaluation rows within class, rematches lengths using all five seeds, and recomputes centroids, dispersions and equalization factors. Both scores use the same resampled rows. Intervals condition on the fixed reference fit and selected features. They are retrospective and do not control family-wise error. Figure 2's baseline intervals use this paired procedure and seed, rather than the bootstrap draws in Table 20. The code supplement includes the paired AUROC estimates; matched indices and point scores can be recomputed from its feature arrays and matching seeds.

For failure analysis, each OOD example receives its tie-aware rank against the unchanged matched ID scores. Averaging these ranks recovers AUROC exactly. The underlying analysis records rank

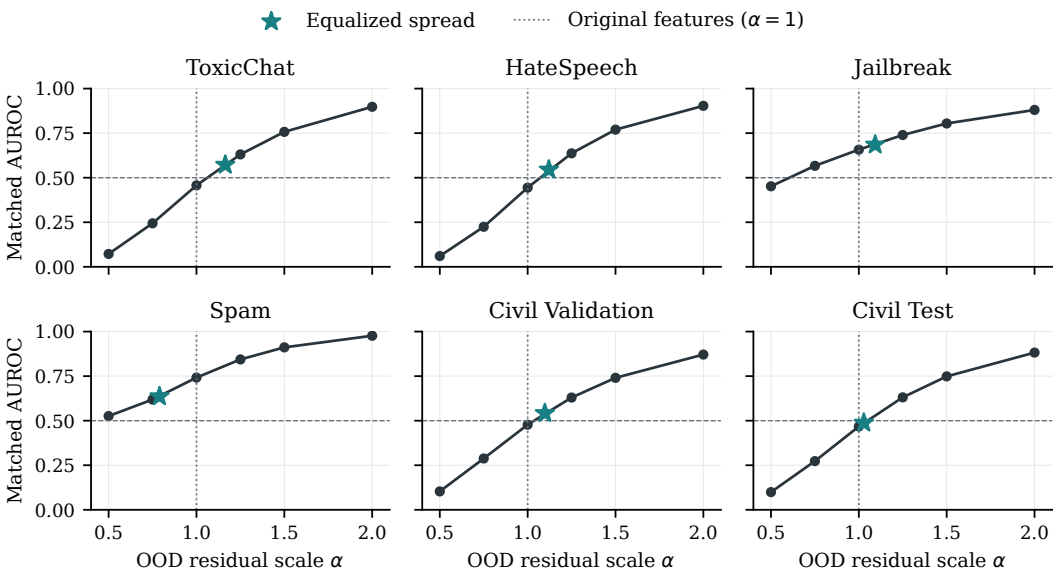


Figure 5: **Full fixed-scale response curves.** No scale is selected to maximize AUROC. The dotted vertical line marks the original features ($\alpha = 1$). Each curve averages five matching seeds; stars show mean equalization scale and mean equalized AUROC, with the factor recomputed per subset. These descriptive curves do not carry simultaneous confidence bands.

increases, decreases and unchanged cases for each matching seed; the cached numeric features and fixed matching seeds allow these ranks to be recomputed. Civil test shows why means and rankings must be checked separately: its equalized mean squared-radius difference is positive, its AUROC point estimate is below 0.5, and its paired-change interval includes zero. Spam's negative paired change is also retained.

G.2 Inference sensitivity and typicality control

These sensitivity checks were designed after observing the main score results. We permute labels within fixed pooled length bins (1,999 replicates), preserving each bin's class counts. We also independently bootstrap reference rows and each evaluation class while refitting standardization and precision (1,000 replicates). Both use seed 20260915 and the five main-analysis matching seeds. Permutation within bins assumes that labels are exchangeable only among examples in the same bin. It does not remove residual within-bin length effects. The refit intervals still condition on candidate design and selected features, and assume independent rows without accounting for shared source clusters.

Table 15: **Fixed-score sensitivity checks on cached evaluations.** The first four rows use LOTO features, unlike the confirmation cases in Figure 2. Full radius uses all ID references; CI resamples reference and evaluation rows. p_B is the unadjusted within-bin permutation value. Split radius and density share a half-reference fit; the other half calibrates density.

Evaluation	Full radius	Refit 95% CI	p_B	Split radius	Density
ToxicChat	0.430	[0.377, 0.499]	0.0160	0.437	0.455
Jailbreak	0.777	[0.675, 0.841]	0.0005	0.748	0.666
HateSpeech	0.431	[0.383, 0.486]	0.0055	0.406	0.494
Spam	0.818	[0.747, 0.886]	0.0005	0.831	0.691
CivilValidation	0.478	[0.449, 0.506]	0.1055	0.476	0.480
CivilTest	0.467	[0.447, 0.488]	0.0020	0.468	0.502

The four LOTO bin tests remain significant after Holm correction (maximum $p = 0.016$). The Civil test two-stage adjusted value is 0.004, whereas validation remains inconclusive. ToxicChat's refit upper endpoint remains close to chance. These controls support the reported ranking pattern on the tested data.

For an ID-only control inspired by statistic typicality (Morningstar et al., 2021), we split reference rows with seed 20260915, fit geometry on the first $\lfloor n/2 \rfloor$, and fit a one-dimensional Gaussian kernel density estimate (KDE) with Scott's bandwidth to radii of the remaining ID rows. Negative log density is fixed as the OOD score. This estimates the density of one statistic and does not implement full DoSE. It uses LOTO features for the four development tasks and fixed Civil features; its half-reference baseline therefore differs from full-reference confirmation. No evaluation labels choose its split, bandwidth or direction. Civil test rises from 0.468 to 0.502, but Jailbreak falls from 0.748 to 0.666 and Spam from 0.831 to 0.691. These descriptive point differences do not establish paired significance. All six outcomes and per-row scores are released.

G.3 Precision-matrix sensitivity

Table 16: **Radial metric sensitivity.** Entries are fixed-direction, five-seed length-matched AUROC. Features and ID standardization are unchanged; rows vary only the precision matrix. Euclidean uses identity precision after standardization, and Maha rows add the stated ridge to the ID covariance.

	Development				Confirmation			
Geometry	**Toxic**	**Jail**	**Hate**	**Spam**	**Toxic**	**Jail**	**Hate**	**Spam**
Euclidean	0.422	0.634	0.426	0.784	0.475	0.421	0.445	0.678
Maha 10^{-4}	0.420	0.851	0.395	0.846	0.458	0.660	0.445	0.741
Maha 10^{-3}	0.420	0.850	0.395	0.848	0.457	0.657	0.444	0.742
Maha 10^{-2}	0.423	0.839	0.399	0.850	0.454	0.625	0.444	0.741
Maha 10^{-1}	0.428	0.803	0.409	0.848	0.447	0.567	0.445	0.741
Maha 1	0.432	0.712	0.412	0.828	0.450	0.480	0.442	0.720

ToxicChat and HateSpeech each remain below chance in all 12 split–geometry combinations, and Spam remains above chance in all 12. Development Jailbreak is above chance throughout. On confirmation it reverses under standardized Euclidean distance and a ridge of 1, while intermediate ridges remain above chance. Reversal on ToxicChat and HateSpeech persists beyond the default 10^{-3} ridge. Jailbreak's ranking depends on covariance geometry.

G.4 All candidates and Civil length balance

Table 17 uses all 21 development-designed candidates, without subset elimination or an AUROC-selected transformation. It keeps the four development AUROC estimates on the same side of chance and gives Civil point estimates below chance. This tests dependence on the nine-coordinate subset, not on the design of the candidate pool. The four same-source confirmation caches contain only the selected nine coordinates, so the all-candidate check is unavailable there. Development and bin-sensitivity results were known before the dispersion diagnostic; these checks do not provide independent confirmation.

Table 17: **All-candidate sensitivity without subset selection.** Fixed-orientation, five-seed length-matched AUROC. The 21 candidates were designed on the development suite.

Stage	Evaluation	Selected 9	All 21
Development	ToxicChat	0.420	0.411
Development	Jailbreak	0.850	0.770
Development	HateSpeech	0.395	0.431
Development	Spam	0.848	0.844
Validation	Civil Comments	0.478	0.486
Official test	Civil Comments	0.467	0.482

Table 18: **Civil Comments across every recorded bin setting.** No bin count is chosen by performance. n_m reports retained examples per class; AUROC uses the unchanged nine-feature score.

Bins	Validation AUROC	n_m	Test AUROC	n_m
5	0.485	459	0.465	898
8	0.478	447	0.471	901
10	0.478	451	0.467	872
15	0.477	446	0.470	862
20	0.481	435	0.471	859
40	0.483	411	0.470	839

At the canonical ten bins, the five-seed mean length-only AUROC is 0.4977 for validation and 0.5013 for test; the corresponding standardized mean differences (OOD minus ID, pooled within-class standard deviation) are -0.0109 and -0.0068. Test retains 872 of 1,000 examples per class. Small scalar balance diagnostics do not rule out nonlinear or within-bin confounding, but the negative test direction is not isolated to one discretization or a large residual mean-length difference.

G.5 Repeated data draws

Table 19: **Fixed-direction repeated-split evaluation** over 100 Monte Carlo reference/OOD redraws. Score orientation remains fixed in every repeat. The 95% range is the empirical 2.5th–97.5th percentile.

Task	**Mean ± std**	**95% range**	**Fixed prefix**
ToxicChat	0.436 ± 0.029	[0.378, 0.492]	0.420
Jailbreak	0.829 ± 0.033	[0.776, 0.888]	0.850
HateSpeech	0.409 ± 0.025	[0.367, 0.460]	0.395
Spam	0.841 ± 0.041	[0.752, 0.908]	0.848
Task mean	0.629	—	0.628

The redraw means reproduce the direction of every development result: ToxicChat and HateSpeech remain below chance, while Jailbreak and Spam remain positive. Cross-draw standard deviations are 0.025–0.041, much larger than variation from length-matching seeds alone.

G.6 Comparable fixed-score uncertainty

Table 20 applies the same stratified bootstrap to each fixed score. These intervals measure evaluation-sample uncertainty, not just variation across length-matching seeds.

Table 20: **Fixed-orientation AUROC with stratified-bootstrap 95% intervals.** "Radius dev." is exploratory development evaluation; "Radius conf." is the held-out confirmation split from the same source.

Task	CED-1C	Input-D^{2}H	Input-TV	Radius dev.	Radius conf.
ToxicChat	0.520 [0.459, 0.582]	0.504 [0.493, 0.516]	0.411 [0.373, 0.455]	0.420 [0.369, 0.483]	0.457 [0.374, 0.540]
Jailbreak	0.626 [0.567, 0.711]	0.455 [0.423, 0.480]	0.533 [0.455, 0.614]	0.850 [0.774, 0.898]	0.657 [0.488, 0.853]
HateSpeech	0.573 [0.507, 0.623]	0.514 [0.498, 0.526]	0.415 [0.372, 0.463]	0.395 [0.348, 0.450]	0.444 [0.405, 0.492]
Spam	0.828 [0.749, 0.900]	0.530 [0.492, 0.568]	0.805 [0.713, 0.863]	0.848 [0.788, 0.905]	0.742 [0.664, 0.796]

G.7 Bin-count sensitivity and overlap support

Length matching changes the evaluated population. Table 1 reports original and matched AUROC; the checks below vary the bins while preserving score orientation. Matched performance cannot be assumed to hold at all lengths.

Table 21: **Development bin sensitivity.** Cells give five-seed mean AUROC (retained count per class). Counts do not vary across seeds. Features and score direction are fixed.

Task	$B = 5$	$B = 8$	$B = 10$	$B = 15$	$B = 20$	Spread
ToxicChat	0.427 (128)	0.441 (123)	0.420 (123)	0.432 (125)	0.432 (123)	0.021
Jailbreak	0.838 (88)	0.847 (86)	0.850 (86)	0.842 (81)	0.838 (73)	0.012
HateSpeech	0.392 (183)	0.405 (182)	0.395 (182)	0.399 (170)	0.411 (166)	0.018
Spam	0.846 (64)	0.840 (59)	0.848 (59)	0.834 (56)	0.852 (56)	0.018

The maximum movement across $B \in \{5, 8, 10, 15, 20\}$ is 0.021, so bin count does not explain the reversals. At $B = 10$, matching changes the OOD/ID mean-length ratio from 2.29 to 0.96 (ToxicChat), 3.27 to 1.03 (Jailbreak), 1.00 to 0.95 (HateSpeech), and 2.34 to 0.98 (Spam). Effective sample retention is 65%, 43%, 91%, and 30%, respectively. The low Spam retention means its matched result applies to a limited overlap region.

H Endpoint Comparisons and Additional Task Coverage

Paired confirmation comparison. The embedding (endpoint) k-NN score uses final-layer token means and cosine distance to ten ID neighbours. On shared confirmation inputs, it leads on Spam (0.891 versus 0.742 for trajectory radius). On ToxicChat, trajectory radius is higher (0.457 versus 0.302), but both scores remain below chance. Paired intervals support these opposite differences; Jailbreak and HateSpeech remain inconclusive. This retrospective comparison changes both representation and scoring rule.

Table 22: **Fixed scores on shared confirmation inputs and ID references.** Qwen3-0.6B AUROCs average five token-matching seeds; n_m is retained per class. Δ is trajectory minus endpoint AUROC. Intervals use 1,000 paired row-bootstrap draws with rematching. They are retrospective, pointwise and conditional on fixed references.

Task	n_m	Endpoint	Trajectory	Δ	95% interval
ToxicChat	62	0.302	0.457	$+0.154$	$[+0.060, +0.237]$
Jailbreak	18	0.584	0.657	$+0.073$	$[-0.149, +0.295]$
HateSpeech	251	0.486	0.444	-0.042	$[-0.100, +0.014]$
Spam	93	0.891	0.742	-0.149	$[-0.219, -0.094]$

Table 22 uses the four-task confirmation split, not an independent source replication. Endpoint embeddings and trajectory features come from the same forward passes with identical ID-reference members: 236, 200, 248 and 244 for ToxicChat, Jailbreak, HateSpeech and Spam. Endpoint embeddings average final-layer states over all input positions; trajectory summaries omit the final

position, as elsewhere. Both scores treat larger values as OOD. In each of 1,000 paired bootstrap draws, rows are resampled within class and token-length matching is repeated with the five standard seeds. We subtract endpoint AUROC from trajectory AUROC within each draw, then take the 2.5th and 97.5th percentiles. These pointwise intervals condition on the reference fit and feature/source design; matching seeds are not independent replications. No score calibration or combination enters this comparison.

Development comparison. Table 23 compares five fixed scores across six development task contrasts. It shares evaluation rows across extractors, using ten token-length bins and seed 42 reset within each bin. Separately filtered reference pools are truncated to equal sizes, but their members can differ. The available aggregates cannot recover paired intervals for these estimates; Table 22's intervals do not apply to them. AGNews compares World (ID) with Sports (OOD); JailbreakClass. resamples the Jailbreak source, not an independent source.

Table 23: **Development comparison on shared matched evaluation rows.** Descriptive Qwen3-0.6B AUROCs with fixed score directions, ten token-length bins and seed 42. ID-reference members can differ. Bold marks the larger endpoint/trajectory estimate, not a significance test.

Task	Energy	MaxLogit	Perplexity	Embedding k-NN	Trajectory radius
ToxicChat	0.607	0.608	0.591	0.336	**0.410**
Jailbreak	0.775	0.770	0.760	0.684	**0.847**
HateSpeech	0.520	0.530	0.556	**0.530**	0.398
Spam	0.338	0.369	0.298	**0.925**	0.867
AGNews	0.658	0.637	0.539	**0.924**	0.751
JailbreakClass.	0.740	0.737	0.747	0.691	**0.764**

Logit and likelihood scores. Energy averages $-\log \sum_v \exp \ell_t(v)$ and MaxLogit averages $-\max_v \ell_t(v)$ over all input positions t, including the last; $\ell_t(v)$ is the vocabulary logit at temperature one. The Perplexity column stores mean next-token negative log likelihood, $-(T-1)^{-1} \sum_{t=1}^{T-1} \log p(x_{t+1} \mid x_{1:t})$, for input tokens $x_{1:T}$. Exponentiation gives perplexity with unchanged AUROC. All three scores use larger values for OOD with no evaluation-label sign selection.

Additional classification datasets. Table 24 covers six CED datasets unused in feature design. Each detector fits one pooled ID distribution, which can contain several source classes:

- **AGNews:** ID = World (0); OOD = Sports (1).
- **Banking77:** ID = labels 0–49; OOD = labels 50–76.
- **SNIPS:** ID = AddToPlaylist, BookRestaurant, GetWeather, PlayMusic, RateBook (first five alphabetically); OOD = SearchCreativeWork, SearchScreeningEvent.
- **News-Categories:** ID = Politics, Wellness, Entertainment, Travel, Style & Beauty (five most frequent); OOD = all other categories.
- **CLINC150:** ID = archived CLINC-OOS in-scope intents; OOD = out-of-scope class.
- **ROSTD:** ID = ROSTD+ training rows marked ID; OOD = test rows not marked ID.

The first four sample source training splits. All six match whitespace-separated word counts, not tokens; the caches lack row identities needed to recover token lengths.

Table 24: **Endpoint–trajectory comparison.** Qwen3-0.6B uses up to 1,000 ID/500 OOD inputs, reserving half the valid ID inputs as references. n_m is the matched count per class, excluding references. AUROC uses fixed directions, ten word-length bins and seed 42.

Dataset	n_m	k**-NN endpoint**	**Trajectory radius**
AGNews	442	0.904	0.681
ROSTD	275	0.921	0.640
SNIPS	406	0.806	0.509
CLINC150	314	0.668	0.504
News-Categories	449	0.638	0.510
Banking77	444	0.573	0.529

Endpoint k-NN has the larger AUROC on all six datasets. These results concern the tested intent/topic shifts and do not establish that vocabulary statistics alone determine performance. Both scores perform poorly on Banking77, which remains in the table.