AI-Text Detection Is at Least Two-Dimensional: A Pooling Artifact, a Factor Test, and Why the Field's Disputes Are Conditional

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

A recurring claim in machine-text detection is that detectors, however different their construction, all read a single underlying axis: post-training style, alignment, or statistical typicality. If true, detectors are interchangeable and their disagreements are noise. We show first that the standard way this claim would be measured is confounded. On a pooled human-and-AI corpus every competent detector correlates through the class label, forcing the first principal component of the detector-score matrix to 83% of variance regardless of what the detectors measure; conditioning on class is necessary and drops it to 58–64%. A participation ratio above one does not by itself refute one axis (a single factor plus per-detector noise also produces it), so we apply Horn's parallel analysis: on a heterogeneous suite of nine detectors spanning six model backbones, including a trained detector, two components exceed the independence null in nine of ten class-by-seed cells, and the effective rank is 3.0 (95% CI $[2.95, 3.06]$). Detection is therefore at least two-dimensional, not one axis. We then show that two live disputes are comparisons made under different unstated conditions rather than disagreements about the phenomenon. Whether detectability tracks alignment or typicality resolves, across six seeds, toward typicality. Whether detectors protect or penalize non-native English writers is conditional on text length: a native-calibrated threshold flags near zero long non-native essays but up to 9.4% (95% CI $[8.0, 11.2]$) of short ones for DeBERTa, while a deployed commercial detector flags 51.6% of short non-native exam essays. Underlying the study is a methodological contribution: four controls for detector-axis experiments, each demonstrated against a specific false conclusion it prevents.

1 Introduction

Automatic detectors of machine-generated text are used in high-stakes settings, notably educational assessment [10], where a positive verdict carries real consequences for a person. A recent thread argues that these detectors, despite ranging from likelihood-curvature methods [12] to trained classifiers [9] to contrastive scorers [7], all read *one* underlying signal. The proposed identity of that signal varies: a post-training style direction ablatable in activation space [2], a pretrained typicality axis [16], or a base-versus-aligned likelihood ratio [18]. What the accounts share is a definite article: *the* axis of AI detection. If they are right, detectors are interchangeable, and a false accusation from one is a false accusation from all.

This paper asks whether the definite article is earned, and finds that the question is harder to measure than it looks. The natural operationalization, build a matrix of detector scores over

*Corresponding author: deep@lab.cloud

documents and measure its effective dimensionality, is confounded in a way that biases the answer toward "one axis." Once the confound is removed the picture is more nuanced than either "one" or "many," and, more usefully, the field's substantive disputes turn out to be conditional rather than contradictory. Along the way we encountered several ways a detector-axis study produces a confident but false conclusion that passes ordinary validation, and we distill them into a small protocol.

Contributions.

1. **Detection is at least two-dimensional, and the usual measurement hides it** (Section 3). On a two-class corpus the first principal component of the detector-score matrix is forced high (83%) by the class label alone; conditioning on class is necessary. On a heterogeneous suite of nine detectors across six backbones, Horn's parallel analysis finds two factors above the independence null robustly across seeds, and the effective rank is 3.0 (CI $[2.95, 3.06]$), ruling out the one-axis-plus-noise account rather than merely the pooled evidence for it.

2. **Detectability tracks typicality more than recipe** (Section 4). Using a post-training pipeline that varies recipe while holding base weights fixed, and replicating across six seeds, base-model typicality is the stronger scalar predictor of detectability in both prompting regimes. This concerns which covariate best predicts detectability, a separate question from how many directions detectors span.

3. **The non-native-writer fairness contradiction is conditional on length** (Section 5). Two published results with opposite conclusions hold under different text lengths: pretrained axes under-flag long non-native essays and, for DeBERTa specifically, over-flag short ones (9.4%, CI $[8.0, 11.2]$), while a deployed detector flags 51.6% of short non-native exam essays. Which detector flips is encoder-specific, so detectors do not share one bias.

4. **A control protocol for detector-axis studies** (Section 6). Four controls, within-class rank, log-space aggregation, non-circular detector selection, and multi-seed replication, each demonstrated against the specific false conclusion it prevents.

2 Related work

Machine-text detection. Zero-shot detectors score text under a language model: likelihood curvature [12], faster curvature estimates [4], and contrastive base-versus-chat scoring [7]. Trained detectors add adversarial robustness [9]. These methods differ in construction but are reported to transfer poorly across generators [7], which motivates the question of whether they nonetheless share an axis.

One-axis accounts. Three recent works argue that a single detectable direction, tied either to post-training or to pretraining, underlies detection. Anand et al. [2] isolate an aligned-minus-base residual direction, ablatable at decode time across many models. Smirnov [16] argue instead for a pretrained typicality direction, explicitly not an AI-content direction, and report that fine-tuning merely rescales pretrained geometry. Wu et al. [18] detect from the base-versus-aligned likelihood ratio. The accounts make different, sometimes opposite, predictions, which our recipe and length experiments adjudicate.

Directions in representation space. The premise that a behaviour corresponds to a single linear direction has a large supporting literature, from linear probes [1] to difference-in-means steering [3] and linear representations of truth [11, 5]. It is also contested: refusal has been argued to be a cone rather than a direction [17], and the superficial-alignment hypothesis has been called an over-simplification [15]. Our rank measurement speaks to this debate for the detection setting specifically. On methodology, control tasks [8] are the standard guard against a probe reading structure that is not there; our controls are in that spirit.

Post-training and fairness. That instruction tuning can be a narrow, mostly stylistic shift is argued from minimal alignment data [19]; the preference-optimization methods our recipe pipeline varies are established separately [14, 13]. Detectors are documented to penalize non-native English writers [10], a finding with direct stakes that our length analysis revisits and reconciles with the opposite typicality prediction.

3 Measuring detection dimensionality

Setup. We assemble eight detectors spanning the two families the one-axis accounts concern: five likelihood-family scores under a base language model (mean log-probability, log-rank, entropy, maximum log-probability, and a curvature estimate) and three encoder [CLS] typicality axes (RoBERTa, DeBERTa, ELECTRA), each fit as a difference-of-class-centroids direction on a held-out split following Smirnov [16]. We score 520 human and 520 machine answers from the HC3 corpus [6], forming an $N \times 8$ score matrix. We z-score columns and compute the participation ratio $\mathrm{PR} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ of the correlation matrix (eigenvalues λ_i), an effective-rank statistic bounded between 1 (one axis) and 8 (eight independent axes) that needs no eigenvalue threshold.

The pooled matrix is confounded. Pooling human and AI documents, the participation ratio is 1.42 with the first principal component explaining 83.2% of variance, which reads as near-one-dimensional. This number is an artifact. On a two-class corpus every competent detector tracks the class label, so all eight correlate through the label whatever else they measure, and the first component is forced high by construction rather than by a shared mechanism. Any dimensionality measurement on pooled text therefore cannot distinguish "one shared axis" from "eight detectors that each separate the classes." The correct test conditions on class.

Within class. Removing the label axis, the within-class participation ratio is 2.39, averaging 2.24 for human text and 2.55 for AI text, with the first component explaining 58–64% of variance rather than 83%. A 1000× document bootstrap gives per-class 95% intervals of $[2.13, 2.35]$ for human text and $[2.41, 2.67]$ for AI text (we do not splice these into a single interval for the cross-class mean of 2.39). As a null, independently shuffling each detector's column destroys cross-detector correlation and yields $\mathrm{PR} = 7.93$, confirming the estimator reaches the independent ceiling when correlation is absent. Table 1 summarizes.

A participation ratio above one is not enough; a factor test is. A single latent axis plus independent per-detector measurement noise also produces $\mathrm{PR} > 1$ and a first component well below 100%; on synthetic rank-one-plus-noise data our own estimator returns $\mathrm{PR} = 3.35$ while the true factor count is one. The participation ratio therefore cannot, on its own, distinguish one axis from

Table 1: Effective dimensionality of the detector-score matrix. Top block: the initial eight-detector suite; pooling is a label artifact, the within-class value is the measurement of interest, the null is a per-column shuffle. Bottom block: the heterogeneous nine-detector, six-backbone suite, where Horn's parallel analysis finds two factors above the independence null (median over five seeds, per class).

matrix	participation ratio	PC1 variance	bootstrap 95% CI	PA factors
pooled (human + AI)	1.42	83.2%	n/a	n/a
within class, human	2.24	64%	[2.13, 2.35]	n/a
within class, AI	2.55	58%	[2.41, 2.67]	n/a
shuffle null	7.93	n/a	n/a	n/a
heterogeneous, within class	**3.00**		[2.95, 3.06]	**2 / 2**

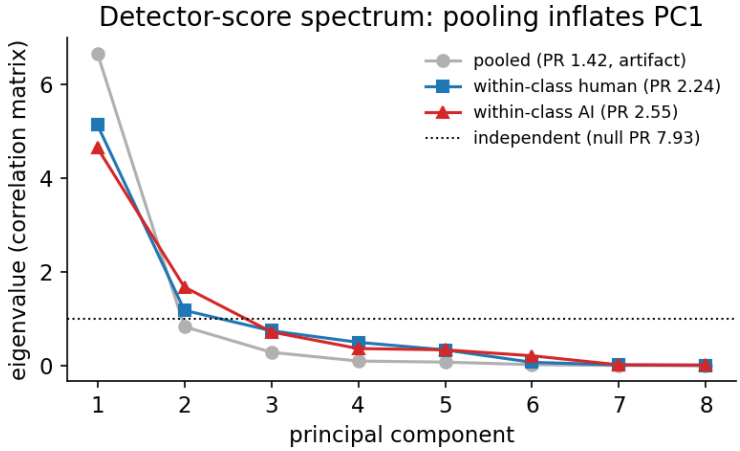

Figure 1: Eigenvalue spectrum of the detector-score correlation matrix. Pooling human and AI text concentrates variance in the first component (grey); conditioning on class spreads it. The dotted line marks the independent-null level ($\lambda = 1$).

several. The correct test is how many components exceed the independence null. We use Horn's parallel analysis: a component is a real factor only if its eigenvalue exceeds the 95th percentile of eigenvalues from column-permuted matrices at that component rank, which the synthetic control passes (one factor for rank-one-plus-noise, three for rank-three).

Two factors survive on a heterogeneous suite. To ensure the result is not an artifact of a suite of near-redundant detectors, we expand to nine detectors spanning six model backbones: three likelihood scores under one base model, two under a second, unrelated base model, the three encoder axes, and a separately trained adversarial detector [9]. Across five seeds, parallel analysis finds two factors above the null in nine of ten class-by-seed cells (median two for each class), and the within-class participation ratio is 3.0 (95% CI [2.95, 3.06], five-seed bootstrap), higher than the 2.4 of the smaller suite rather than lower. A more heterogeneous suite raises the effective rank, the opposite of what a suite-composition artifact would predict. We therefore conclude that detection is at least two-dimensional: not one axis, but a small number of genuine factors rather

than eight independent ones. This is corroborated in deployment (Section 5), where the same essay is confidently AI to one encoder and confidently human to another.

Pre-registration. We pre-registered a decision rule on first-component variance ($\geq 70\%$ confirms one-dimensionality, $< 50\%$ falsifies). We report it for completeness and note that neither the pooled value (83%) nor the within-class value (58–64%) crosses the falsification threshold, so the one-dimensionality question is not settled by the registered variance cut. It is settled instead by parallel analysis, a factor test the registration required us to compute but did not assign a threshold. The exact participation-ratio value depends on the detector suite, which is why the load-bearing claim is the parallel-analysis factor count (robust at two across two suites of different size and composition) rather than the ratio itself.

4 Recipe: what drives detectability?

Setup. To separate post-training recipe from base-model typicality we use a pipeline of four checkpoints derived from one base model by successive post-training stages (base, supervised fine-tuning, preference optimization, and reinforcement learning), holding base weights fixed; capability is intended to be approximately held as well, though we did not verify capability invariance with matched task accuracy. For each checkpoint we generate continuations of HC3 questions, score detectability with an encoder [CLS] axis that is not a function of base-model likelihood, and measure base-model log-probability as the typicality regressor. We compute two partial correlations per document: detectability versus recipe stage controlling for typicality, and detectability versus typicality controlling for recipe. The two regressors are near-independent per document ($r = 0.15$).

Relation to dimensionality. This asks which scalar covariate best predicts a checkpoint's detectability, a distinct question from Section 3's count of directions. Typicality can be the strongest single predictor of *how detectable* a text is while detectors still disagree *per document* about which texts are AI; the two are compatible, and only the latter bears on interchangeability.

Non-circularity. An earlier version used base-model log-probability as both the detector and the typicality regressor, which returned a partial correlation of 1.0 to machine precision, an identity rather than a measurement. The reported experiment uses an encoder detector distinct from the regressor, with the detector-regressor correlation held below a guard of 0.99 (observed maximum 0.61).

Result. Across six seeds, each re-splitting the data and re-sampling generation, base-model typicality is the stronger predictor of detectability in both prompting regimes. Under a fixed prompt frame the typicality partial is 0.456 ± 0.073 against a recipe partial of 0.161 ± 0.038; under native chat templating it is 0.361 ± 0.043 against 0.193 ± 0.021 (Table 2). A one-sample t-test on the per-seed difference gives $t(5) = -3.01$ (fixed frame) and $t(5) = -5.66$ (native), which we report as directional-confirmatory rather than discovery statistics given the small family of tests. This favours the typicality account over the alignment reading as a directional result, stable to seed and prompt.

A single-seed artifact. A single-seed run of this experiment appeared to show that the winner *flips* with prompt format, recipe winning under a fixed frame and typicality under native templating.

Table 2: Recipe versus typicality, partial correlations with detectability (mean ± SE over six seeds). Typicality is the stronger predictor in both regimes; the seed counts are descriptive, the t-tests inferential.

prompting	recipe partial	typicality partial	seeds favouring typicality
fixed frame	0.161 ± 0.038	$\mathbf{0.456 \pm 0.073}$	5/6 ($t(5) = -3.01$)
native template	0.193 ± 0.021	$\mathbf{0.361 \pm 0.043}$	6/6 ($t(5) = -5.66$)

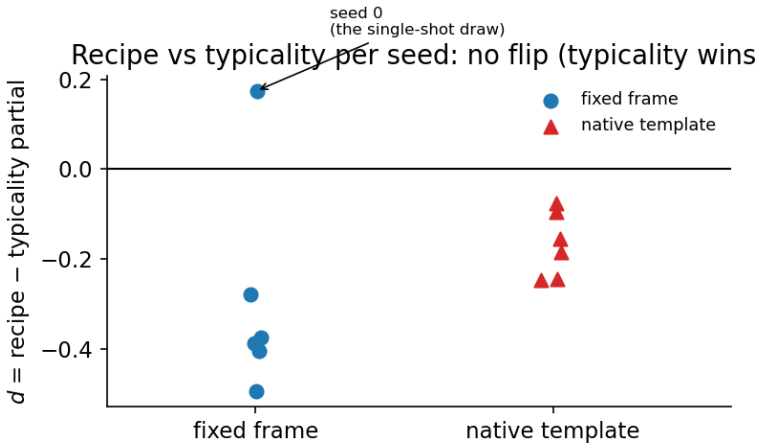

Figure 2: Per-seed difference d between the recipe and typicality partial correlations, for both prompting regimes. All points but one fall below zero (typicality stronger); the single positive point, seed 0 under a fixed frame, is the draw a single-shot run would have reported as a recipe win.

Six-seed replication does not support this: the apparent flip was one seed (a fixed-frame difference of +0.174 at seed 0, against five negative seeds), and the paired flip test is not significant and has the wrong sign ($t(5) = -1.51$). We therefore treat the flip as an unsupported single-seed artifact rather than a finding. The single-seed estimate was sign-unstable because the recipe regressor is constant within each generated arm, so its naive per-document interval understated its variance; the between-seed standard error is the correct inference.

5 Length: the non-native-writer contradiction

Two opposite published results. Liang et al. [10] report that deployed detectors flag non-native English essays as AI at high rates. Smirnov [16] report the opposite for a pretrained typicality axis: non-native essays are atypical under pretraining and project onto the human side, with detection AUROC inverting below chance. The two use corpora of very different length: the deployed-detector study uses essays of median 104 words, the typicality study essays of median 429 words, a fourfold difference with almost no overlap.

The deployed harm. On the 91 non-native exam essays of the deployed-detector study, all human, a commercial detector flags 51.6% (95% CI $[41.5, 61.6]$) as AI at its default threshold; the interval includes the 61% previously reported, so the two are consistent. The false positives are

Table 3: False-positive rate on non-native essays at a threshold calibrated to 5% on native writers, with Wilson 95% intervals (disparity = FPR/0.05). Only DeBERTa's short-length rate excludes the 5% target. The final row is a deployed commercial detector at its default threshold on the separate short-essay exam corpus, not a calibrated column.

detector	full (429 w)	length-matched (94 w)
likelihood log-prob	0.0% [0.0, 0.4]	0.1% [0.0, 0.4]
RoBERTa axis	0.1% [0.0, 0.4]	4.4% [3.4, 5.6]
DeBERTa axis	0.0% [0.0, 0.3]	**9.4%** [8.0, 11.2]
ELECTRA axis	0.1% [0.0, 0.4]	0.2% [0.1, 0.7]
commercial (default threshold, short corpus)	n/a	**51.6%** [41.5, 61.6]

confident, not hedged: the score distribution is bimodal, with 25 of 91 essays scored at mean AI probability 0.89, giving a Brier score of 0.344 against a true label of zero. Within this corpus length does not explain the effect (score-length correlation +0.03).

Length as a moderator. We calibrate each detector's threshold to a 5% false-positive rate on native reference writers, then measure the false-positive rate on non-native essays at full length and matched to the short deployed length (Table 3, with Wilson intervals). At full length every pretrained axis under-flags non-native writers (near zero). Matched to short length, DeBERTa rises to 9.4% (CI [8.0, 11.2], excluding the 5% target), a genuine over-flag; RoBERTa reaches 4.4% (CI [3.4, 5.6]), statistically indistinguishable from the calibration target rather than a rise; ELECTRA and the likelihood detector stay well under. The two literatures are thus not in direct conflict: read as a moderation by length, the typicality inversion holds for long text and weakens or reverses for short text, encoder by encoder.

We attach two caveats rather than a clean causal claim. First, the full-versus-short contrast here is drawn across two corpora (a long-essay and a short-essay exam set) that differ in domain, register, and first-language population as well as length; the within-corpus truncation control (where only DeBERTa shows a significant length effect and one encoder's inversion does not replicate) supports length as a moderator but does not isolate it. Second, the commercial-detector rate is a different detector at a different operating point (its default threshold) on the short corpus, so it is an illustration of deployed harm, not a row commensurable with the calibrated encoder columns. What is robust and not contingent on these caveats is that the encoders disagree: DeBERTa over-flags short non-native text where ELECTRA does not, so a fairness harm depends jointly on the detector deployed and the length of the text.

6 A control protocol for detector-axis studies

Each result above was first reached in a version that passed ordinary validation (correct array shapes, no missing values, plausible outputs) but was wrong, and was caught only by a targeted check on a quantity whose plausible range is known independently of the pipeline. We offer the following as a checklist to run *before* concluding, stated with the false conclusion each item prevents.

1. **Measure rank within class, not pooled.** A pooled two-class corpus reports the label axis, not the detector structure, and would "confirm" one-dimensionality (PR = 1.42, PC1 83%)

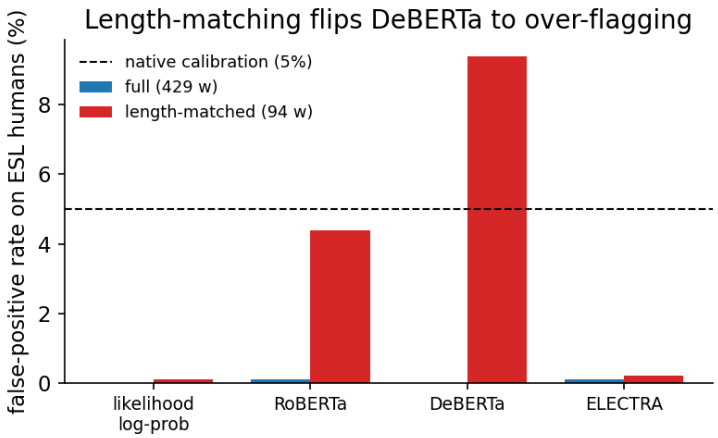

Figure 3: False-positive rate on non-native essays at a threshold calibrated to 5% on native writers. At full length all axes under-flag; matched to short length, only DeBERTa clearly exceeds the native rate, while ELECTRA and the likelihood detector stay low. The bias is encoder-specific.

mechanically.

2. **Aggregate likelihoods in log space.** Correlating per-document perplexity, which is heavy-tailed, let a few degenerate generations move a reported mean between 288 and 54,983 across otherwise comparable runs while the median stayed below 11; log-probability is bounded and reproducible.

3. **Keep the detector distinct from the regressor.** Using base-model likelihood as both makes the adjudicating correlation an identity ($r = 1.0$); assert that their correlation stays below a threshold.

4. **Replicate across seeds where a regressor is arm-constant.** A single-seed run produced a sign-unstable "flip" that six seeds did not support; use the between-seed standard error, since the naive per-document interval understates variance.

We caught these in our own drafts and cannot claim to have surveyed how prevalent they are in published work; establishing that would require a separate audit. We note only that the conditions that produce them, single-seed and single-generator pipelines reported without a within-class or non-circularity check, are common, and offer the checklist accordingly.

7 Discussion

The useful reframing is that "the axis of AI detection" is a category error. The pooled measurement that would answer it is confounded toward "one," and once the confound is removed a factor test finds at least two genuine directions, not one. Equally practically important, the field's live disputes, alignment versus typicality and protection versus penalization of non-native writers, are not disagreements about the phenomenon but comparisons made under different conditions (prompt and length), and detectors do not share one bias. The same non-native essay is confidently AI to one encoder and confidently human to another, so a platform that licenses one detector is not

making a neutral implementation choice but selecting a particular, and length-dependent, pattern of false accusation.

8 Limitations

Suite and generality. We report two detector suites: an initial eight-member suite and a heterogeneous nine-member, six-backbone suite that includes a trained detector, and the parallel-analysis factor count is two in both; the participation ratio itself is suite-dependent, so we rely on the factor count. We did not run a leave-one-detector-out analysis or a retrieval-style detector. The generations come from a single instruction-tuned model family and the human text from one question-answering corpus plus one non-native exam corpus, so the categorical reach of the claims is limited and a second generator and corpus are the clearest next step. In particular, the parallel-analysis factor count of two is established on one generator; confirming it holds for AI text from an unrelated generator would make the "at least two-dimensional" claim generator-robust.

Inference. The rank and fairness point estimates carry document-bootstrap or Wilson intervals but, unlike the recipe experiment, a single generation seed, so their intervals are conditional on that seed and do not capture generation variance. The recipe t-tests are low-degree-of-freedom ($df = 5$) and directional; we did not correct for the family of tests and frame them as confirmatory rather than discovery statistics.

Construct validity. The length analysis compares a short-essay and a long-essay corpus that differ in domain and register as well as length; the within-corpus truncation control supports length as a moderator but does not isolate it, and one encoder's inversion did not replicate. The capability-invariance of the recipe checkpoints is assumed, not verified.

Open pre-registered tests. Two pre-registered hypotheses, the simultaneity-of-ablation sweep that discriminates one direction from a cone, and cross-model-family transfer, were not run.

9 Availability

Code and data are available from the authors on request. No model checkpoints are released.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. 2016. URL https://arxiv.org/abs/1610.01644.

[2] Aniket Anand, Janvijay Singh, Zhewei Sun, Dilek Hakkani-Tür, and Nick Feamster. Measuring, Localizing, and Ablating Alignment Signatures in LLMs. 2026. URL https://arxiv.org/abs/2605.30526.

[3] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. 2024. URL https://arxiv.org/abs/2406.11717.

[4] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. 2023. URL https://arxiv.org/abs/2310.05130.

[5] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. 2022. URL https://arxiv.org/abs/2212.03827.

[6] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. 2023. URL https://arxiv.org/abs/2301.07597.

[7] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. 2024. URL https://arxiv.org/abs/2401.12070.

[8] John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. 2019. URL https://arxiv.org/abs/1909.03368.

[9] Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. RADAR: Robust AI-Text Detection via Adversarial Learning. 2023. URL https://arxiv.org/abs/2307.03838.

[10] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. GPT Detectors Are Biased Against Non-Native English Writers. 2023. URL https://arxiv.org/abs/2304.02819.

[11] Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in LLM Representations of True/False Datasets. 2023. URL https://arxiv.org/abs/2310.06824.

[12] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. 2023. URL https://arxiv.org/abs/2301.11305.

[13] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. 2022. URL https://arxiv.org/abs/2203.02155.

[14] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. 2023. URL https://arxiv.org/abs/2305.18290.

[15] Mohit Raghavendra, Vaskar Nath, and Sean Hendryx. Revisiting the Superficial Alignment Hypothesis. 2024. URL https://arxiv.org/abs/2410.03717.

[16] Alexander Smirnov. Amplifying, Not Learning: Fine-Tuned AI Text Detectors Amplify a Pretrained Direction. 2026. URL https://arxiv.org/abs/2605.21653.

[17] Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence. 2025. URL https://arxiv.org/abs/2502.17420.

[18] Junxi Wu, Kailin Huang, Dongjian Hu, Bin Chen, Hao Wu, Shu-Tao Xia, and Changliang Zou. Alignment Imprint: Zero-Shot AI-Generated Text Detection via Provable Preference Discrepancy. 2026. URL https://arxiv.org/abs/2604.16923.

[19] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment. 2023. URL https://arxiv.org/abs/2305.11206.