LLM CONTENT DETECTION AND INTERPRETATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In LLM-assisted writing, who develops a document's content can differ from who produces its final wording. Existing content detectors, however, can follow expression cues even when the content source remains unchanged. We study both content-source detection under changes in expression and the automatic discovery of recurring differences between human and LLM content choices across topics. To support both tasks, we decompose documents into self-contained content points and restate them in a common, plain register. Using these points as detector inputs substantially improves robustness to changes in paragraph structure and voice on HART and RAID. Across RoBERTa, RACE, and DETree, content-point inputs raise mean AUROC over 36 HART expression-change settings from 0.733 to 0.956, while avoiding the prediction reversals observed with raw-text inputs. For interpretation, we introduce *content repertoire occupancy* (CRO), which discovers and names recurring content choices without a predefined domain taxonomy. CRO learns a shared, overlapping repertoire of content choices without content-source labels, then compares how often human and LLM content points express each choice within matched topics. Across four writing domains, CRO identifies tens of content choices per domain, with both their population differences and descriptions supported on held-out data, revealing distinctions not made explicit in existing methods' topic words and steering descriptions. Together, these methods support a more systematic study of human–LLM content differences, combining more robust source detection across forms of expression with automatic discovery and naming of recurring content choices across topics.

1 INTRODUCTION

Writing and language production involve choices about both what to convey and how to convey it. These dimensions are often described as *content*, including the specific claims, examples, events, and details in a document, and *expression*, including its wording, syntax, tone, and register. The distinction between what to convey and how to convey it has a long history, reaching back to classical rhetoric (Aristotle, 2006) and appearing in linguistic theory and models of language production (Hjelmslev, 1961; Levelt, 1989); studies of text comprehension further show that readers can retain meaning after forgetting the exact wording (Sachs, 1967). We call whoever develops the underlying claims and details the *content source*, and whoever produces the final wording the *expression source*. This distinction becomes especially consequential in LLM-assisted writing: a human's ideas can be polished by an LLM, and machine-generated content can be rewritten in a human register. Identifying the expression source therefore does not necessarily identify the content source.

Existing detectors can struggle with this distinction. Many AI-generated text detectors use statistical or learned cues that can depend on expression (Mitchell et al., 2023; Bao et al., 2023; Hans et al., 2024; Soto et al., 2024), and paraphrasing can substantially change their predictions (Krishna et al., 2023). Recent work addresses the distinct contributions of humans and LLMs to a document's creation and revision, including the distinction between content and expression sources (Zhang et al., 2024; Bao et al., 2025; Li et al., 2026; He et al., 2025). Yet, as we demonstrate, even methods intended to identify the content source can still change or even invert their content source prediction under expression changes. Our experiments show that predictions can cease to track the content source under changes as simple as paragraph splitting and merging, which preserve the text up to whitespace, as well as under re-voicing instructed to preserve substantive content. This behavior exposes a central challenge for content-source detection: how to reduce expression cues while preserving the substantive choices that carry evidence about the source label.

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.

In this work, we study two related problems concerning human- and LLM-sourced content: *content-source detection*, which identifies the content source under content-preserving changes in expression, and *content-choice difference interpretation*, which discovers and names recurring differences in human and LLM content choices across diverse topics. To support both tasks, we introduce a content extractor, ϕ, that decomposes each document into self-contained *content points* and restates each point in a common, plain register. These operations are complementary: decomposition seeks to retain individual claims, examples, events, and details, while plain restatement reduces variation in their expression. Together, they aim to reduce expression cues while retaining substantive detail. For detection, content points markedly outperform abstract summaries in our experiments, supporting the value of retaining localized content choices. A binary content-source objective alone does not remove sensitivity to expression, motivating normalization of the detector input (Appendix C.3). Across RACE, RoBERTa, and DETree, ϕ substantially improves robustness to expression changes.

For *interpretation*, we seek to discover and name recurring differences in the content choices of human and LLM writing across topics, without relying on a domain-specific taxonomy. Existing work approaches this question from two directions. Broad studies primarily identify linguistic signatures, topical preferences, and diversity patterns in human and LLM writing (Zanotto & Aroyehun, 2025; Kwon & Zou, 2026; Zhang et al., 2025), rather than a shared repertoire of content choices that recur across topics and differ in their inclusion rates between humans and LLMs. Richer content analyses reveal substantive differences, but through domain-specific questions and feature spaces, such as character portrayal and narrative construction in fiction or research taste in scientific ideation (Huang et al., 2024; Russell et al., 2026; Chen et al., 2026). This leaves open how to discover nuanced content differences across writing domains within a shared analytical framework. We seek a domain-general method for discovering and naming nuanced human–LLM content-choice differences across topics. This task poses two challenges. First, topic variation can overwhelm the differences of interest: for example, differences between writing about AI and writing about cars can exceed human–LLM differences within either subject. A general method must therefore reduce the influence of shared subject matter and distinguish recurring content-choice differences from those confined to individual topics. Our comparisons show how topic vocabulary can dominate the patterns recovered by existing interpretation methods. Second, as our study reveals, human and LLM content choices overlap substantially: many meaningful choices occur in both populations and do not individually determine the source. Methods designed to simultaneously separate the populations and explain their differences tend to reduce this overlap to a few broad dimensions, obscuring the nuanced choices we seek to name (Kuznetsov et al., 2025).

To meet these challenges, we introduce *content repertoire occupancy* (CRO), which first learns a shared repertoire of content choices and then measures how differently human- and LLM-sourced writing occupies it (Figure 1). To reduce shared topic variation, CRO subtracts from each content-point embedding the mean embedding of all points within its matched topic or prompt, then pools the resulting residuals across topics. Our extractor ϕ provides a common unit of analysis across domains and multiple points within each topic, making this centering possible even when only one original human document is available. CRO fits a sparse autoencoder to the centered embeddings without access to content-source labels. The learned repertoire is overlapping: a point may express several choices, and either population may occupy any choice. Separating repertoire learning from population comparison allows CRO to retain choices shared by both populations whose differences lie in their inclusion rates. Only after the repertoire and its membership rules are fixed does CRO introduce source labels to measure occupancy, defined as the proportion of each population's content points that include a choice. It compares occupancy within matched groups addressing the same topic or writing task and weights these groups equally, reporting which population occupies each choice more often and by how much, expressed as an odds ratio. CRO names statistically supported choices using examples drawn from diverse topics, while withholding source labels and occupancy statistics from the naming model. A recurring content choice can admit descriptions at different levels of specificity. CRO therefore evaluates the population difference and its description separately on held-out matched groups: the occupancy difference must replicate in the same direction, and the description must pass member recognition against nonmembers from the same group, with accuracy exceeding a measured random-control floor. Together, these checks establish replicated population differences with descriptions supported by held-out recognition, while the examples and response coverage inform how broadly each description can be read.

CRO reveals recurring differences in content choices between human and LLM writing across domains. For example, in HART's news corpus, which covers diverse topics, CRO reports human-to-LLM odds ratios of $4.5\times$ and $4.1\times$ for "quantitative ratios or metrics" and "specific dates or timeframes," respectively. It reports an LLM-to-human odds ratio of $7.0\times$ for "emphasizing importance." Applied identically to creative writing without prior domain knowledge, CRO finds that LLM writing favors "transformation and renewal," with an LLM-to-human odds ratio of $5.0\times$, while being less likely to include "use or presence of firearms," for which human writing leads with a human-to-LLM odds ratio of $2.7\times$. CRO is not limited to domain-specific repertoires: it can learn a single, interpretable repertoire spanning heterogeneous forms of writing. When content points from HART's news, arXiv, creative-writing, and essay domains are pooled for repertoire learning, supported content choices include human-favored "quantitative specifics" and LLM-favored "emphasis on collective progress and shared values" and "emphasizing significance," with population contrasts estimated separately within each domain.

Many linguists and philosophers who initially approached thought, expression, or grammatical structure as distinct objects of analysis later emphasized how closely they are bound together: words and grammar actively participate in shaping what is conveyed (Cicero, 2001; Wittgenstein, 2009; Pires De Oliveira, 2001; Langacker, 2016). These intellectual trajectories highlight the difficulty of separating content from expression. Given one piece of writing, it is difficult to draw a precise boundary between what constitutes content and what constitutes expression. We make a related observation in our experiments: even rewrites instructed to preserve content sometimes introduce new claims or commitments. With access to both the original texts and their rewrites, we provide a supplementary analysis (Appendix C.7) examining content shifts under re-expression alongside the associated detector responses. In the detection setting we study, however, the detector must assess the text as presented, without access to its earlier versions or the changes through which it was produced. We use ϕ to represent the substantive points that text conveys, providing a shared basis for detecting content-source differences across forms of expression and interpreting recurring content choices across topics. *Despite the broader challenge of precisely separating content from expression, our results show that this representation substantially improves the robustness of content-source detection under changes in expression.* Our contributions are:

- **More robust content-source detection.** We introduce a plain content-point representation that substantially improves detection under changes in expression. Across RoBERTa, RACE, and DE-Tree on HART and RAID, it improves content-source discrimination and reduces false positives.
- **Automatic discovery and interpretation of content differences.** We reveal substantial overlap between human and LLM content choices and develop CRO to discover nuanced differences in their inclusion rates across topics. CRO reveals recurring population differences not made explicit in the compared methods' topic words and steering descriptions, with both the differences and their descriptions supported on held-out data.
- **Representation benefits and applicability across domains.** Through extensive experiments and ablations, we show how retaining substantive detail, plain restatement, and point decomposition contribute to detection robustness. We also demonstrate that the same CRO procedure identifies tens of supported content choices across different domains without the need to predefine any domain-specific categories.

2 Related Work

2.1 Content and expression in writing

The distinction between what is conveyed and how it is expressed appears across research on language and writing. Rhetoric considers both what to say and how to say it (Aristotle, 2006); models of language production separate message planning from linguistic formulation (Levelt, 1989; REITER & DALE, 1997). Psychological research on text comprehension shows that readers can retain a passage's meaning after forgetting its exact wording (Sachs, 1967). In natural language processing, paraphrasing and text style transfer give this distinction a practical role, seeking to preserve specified information while changing its presentation (Krishna et al., 2020; Jin et al., 2022).

Content and expression nevertheless interact. Alternative expressions can convey different perspectives on a situation (Langacker, 2008; Trott et al., 2020), and grammatical constructions themselves

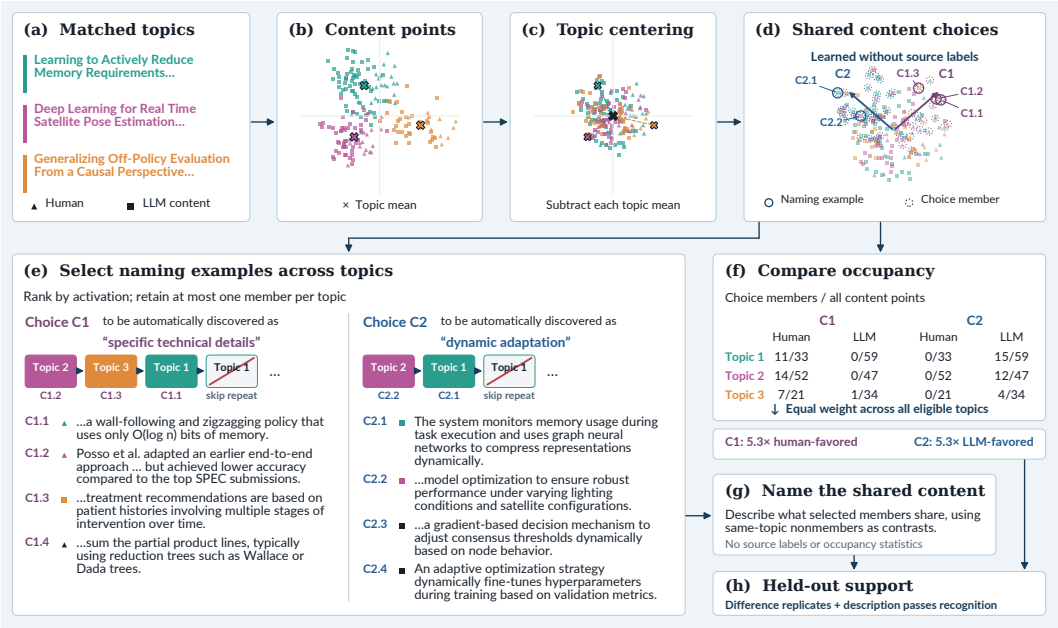

Figure 1: **CRO makes recurring human–LLM content differences explicit and testable.** Real arXiv examples share a 2D embedding projection; memberships are fixed before projection. (a–d) Topic centering reduces shared subject variation before source-blind learning of overlapping choices. (e, g) Naming uses members across topics and same-topic nonmembers, without source labels or occupancy statistics. (f) Population comparisons fix memberships and weight all eligible topics equally, beyond the three shown. (h) Held-out tests separately assess population replication and member recognition from descriptions.

contribute meaning (Goldberg, 1995). These accounts suggest that some aspects of expression cannot be removed without changing what is conveyed, making complete separation a questionable objective. For our purposes, we focus on the substantive points conveyed by a document: we instruct an LLM to retain its claims, events, examples, and details while restating them in plain language, with the aim of reducing expression cues.

2.2 AI-GENERATED TEXT DETECTION AND CONTENT SOURCE

In collaborative writing, content and final wording can have different sources (Zhang et al., 2024; Bao et al., 2025). Generated-text detectors score observed text through probability curvature (Mitchell et al., 2023) or comparisons between language models (Hans et al., 2024). Several approaches directly emphasize how text is expressed: Fast-DetectGPT motivates its conditional-probability score through differences in word choice given a context (Bao et al., 2023), while Soto et al. (2024) explicitly learn writing-style representations for detection. Paraphrase vulnerabilities (Krishna et al., 2023; Sadasivan et al., 2023) motivate defenses through retrieval (Krishna et al., 2023), adversarial training (Hu et al., 2023), and reconstruction of earlier wording (Rivera Soto et al., 2025). Beyond robustness, collaborative writing raises the question of which contribution to identify. HART separates content and expression in a two-dimensional detector, using content to improve robustness (Bao et al., 2025); RACE models creator–editor roles (Li et al., 2026), and DETree represents collaborative production processes hierarchically (He et al., 2025). However, specifying content source as the target does not determine which cues a detector learns: expression may correlate with that source during training, allowing predictions to depend on this association. Changes in expression can then disrupt predictions even when substantive content is retained. We examine this dependence through expression changes during both training and evaluation, testing whether detailed content points in a common register improve robustness.

2.3 DISCOVERING AND INTERPRETING DIFFERENCES BETWEEN HUMAN AND LLM WRITING

Characterizing differences in human and LLM writing. Content-source detection also raises an interpretive question: which recurring content choices differ between populations? Existing studies

characterize linguistic variation (Zanotto & Aroyehun, 2025), topical preferences (Kwon & Zou, 2026), and response diversity (Zhang et al., 2025). Assisted-writing studies report reduced collective diversity in some settings (Padmakumar & He, 2024; Doshi & Hauser, 2024). More focused analyses compare character portrayal (Huang et al., 2024); StoryScope automatically induces narrative features concerning agency, plot, and temporal structure (Russell et al., 2026). Research-taste analysis compares the opportunities and methods selected by humans and LLMs (Chen et al., 2026), while analyses of LLM justifications identify semantic tropes recurring across prompts (Wright et al., 2024). These setting-specific studies motivate a broader question: how can we discover and describe recurring content-choice differences across domains without domain-specific taxonomies?

Methods for discovering and explaining differences. Comparative topic models characterize shared themes and variation across collections or document attributes (Chen et al., 2015; Roberts et al., 2019). SAE-TM learns overlapping sparse features from text embeddings and interprets them through word distributions and merged topics (Girrbach & Akata, 2026). Kuznetsov et al. (2025) select internal language-model features useful for human–machine discrimination and investigate their effects through steering, encompassing wording, discourse, and subject matter. Automated interpretation methods test whether descriptions predict feature behavior on unseen examples (Bills et al., 2023; Paulo et al., 2025), including attempts to falsify explanations using close negative examples (Ma et al., 2025). Our objective brings these concerns together around recurring content choices beyond shared subject matter. Our experiments reveal choices shared by both populations but included at different rates, so interpretation requires identifying the shared choices and comparing their inclusion rates. CRO learns a shared repertoire without source labels, compares inclusion within matched contexts, and separately validates the population differences and their descriptions on held-out groups.

3 METHOD

3.1 CONTENT-POINT EXTRACTION

Both tasks share one representation. Let X be a document. The extractor maps it to an ordered list of content points,

$$\phi(X) = (p_1, \ldots, p_m), \tag{1}$$

where each p_i is a self-contained sentence expressing a substantive claim, event, example, relation, qualification, or detail conveyed by X. We implement ϕ with a prompted language model that applies identical instructions independently to each document, without source labels or rewrite histories, and returns numbered points in document order.

The extraction prompt specifies two operations. *Decomposition* asks the model to enumerate every distinct substantive point, retaining examples, qualifications, and supporting details that a main-idea summary could omit. *Plain restatement* asks the model to express each point in plain, neutral language without reusing the author's distinctive phrasing, tone, imagery, or stylistic quirks. The complete extraction prompt appears in Appendix B.4.

Use in content-source detection. We apply ϕ before the detector's input processing, representing each document by the text of its ordered content points. Training, validation, and test documents undergo the same transformation, with their content-source labels unchanged. We compare each detector fitted and evaluated on original texts with its counterpart fitted and evaluated on content points. Within each comparison, the architecture, learning objective, and scoring rule remain fixed; text-dependent features and structures are recomputed from the corresponding representation. Expression changes may affect training, test, or both; extraction follows each change.

3.2 CONTENT REPERTOIRE OCCUPANCY

Analysis setup and output. CRO compares human and LLM content choices across matched groups of documents. Each group $g \in \mathcal{G}$ contains documents addressing the same topic or writing task, with content-source labels $y \in \{\mathrm{H}, \mathrm{M}\}$; expression source may vary independently. We encode each content point p_{gdi} from document d as a unit-normalized sentence embedding h_{gdi}. For each supported choice, CRO returns a membership rule, a name and description, and a human–LLM occupancy contrast. Groups are partitioned before fitting, with held-out groups reserved for statistical and semantic evaluation.

Learning the shared repertoire. To reduce variation associated with the shared subject, we subtract each group's mean embedding from its points. With N_g points in group g,

$$\bar{h}_g = \frac{1}{N_g} \sum_{d,i} h_{gdi}, \qquad \widetilde{h}_{gdi} = h_{gdi} - \bar{h}_g. \tag{2}$$

All documents and expression conditions in the group contribute to this mean, without using source labels. The residuals describe each point relative to the content shared within its group.

We pool the residuals from discovery groups and fit a sparse autoencoder (SAE). Its activations and reconstruction are $a(\widetilde{h}) = \mathrm{ReLU}(W\widetilde{h} + b)$ and $\widehat{h} = Da(\widetilde{h})$. For the set $\mathcal{T}$ of discovery points, the objective is

$$\mathcal{L}_{\mathrm{rep}} = \frac{1}{|\mathcal{T}|} \sum_{\widetilde{h} \in \mathcal{T}} \left[\|\widetilde{h} - Da(\widetilde{h})\|_2^2 + \lambda \|a(\widetilde{h})\|_1 \right], \qquad \|D_{\cdot j}\|_2 = 1. \tag{3}$$

The sparse code permits overlapping choices, while unit-norm decoder columns constrain the representation's scale. Neither reconstruction nor sparsity uses source labels, allowing the repertoire to retain recurring choices shared by both populations. Capacity and sparsity are assessed without source labels, using criteria for sparse overlap and semantic coherence (Appendix D.2.7).

Measuring differential occupancy. For each choice j, we fix the threshold τ_j at the median positive discovery activation; choices with insufficient support are inactive. A point is a member when $a_j(\widetilde{h}_{gdi}) \geq \tau_j$. Source labels enter only after the repertoire and membership rules are fixed. Occupancy is the proportion of a population's content points assigned to a choice. We compare human and LLM membership odds within matched groups and average the corrected log-odds ratios δ_{gj} equally over groups $\mathcal{G}_j$ with sufficient membership support:

$$\Delta_j = \frac{1}{|\mathcal{G}_j|} \sum_{g \in \mathcal{G}_j} \delta_{gj}, \qquad R_j = |\mathcal{G}_j|. \tag{4}$$

Here R_j records recurrence across groups. Occupancy accounts for each population's available points; equal weighting prevents larger groups from dominating. Positive Δ_j indicates greater human occupancy and negative Δ_j greater LLM occupancy. We report $\exp(|\Delta_j|)$ as the membership odds ratio in that direction, not an inclusion-probability ratio. For both domain-specific and pooled repertoires, contrasts are estimated within each domain. Count definitions, correction, and support rules appear in Appendix D.2.3.

Naming choices across topics. CRO names choices whose discovery differences pass statistical tests and checks of effect consistency across discovery subsets. A naming model receives high-activation members spanning matched groups and nonmembers from the same groups, excluding held-out groups. It returns a short name and one-sentence description of the shared content or communicative move. Spanning groups encourages descriptions that apply across subjects; source labels, generator identities, and occupancy statistics are withheld. Selection rules and the prompt appear in Appendices D.2.4 and D.2.8.

Supporting the difference and its description. We test occupancy contrasts by permuting source labels among documents within matched groups, keeping each document's points together and controlling the false discovery rate across active choices. Discovery differences must also retain their direction and sufficient magnitude across group subsets. Replication requires a significant held-out contrast in the same direction under multiple-testing correction, using the fixed repertoire and membership rules; each held-out group is centered on its own mean. Separately, a judge uses only the name and description to identify the member in unseen member/nonmember pairs from the same held-out group, testing recognition beyond shared subject matter. Group-matched random memberships undergo the same naming and judging procedure to establish a control floor. Descriptions must pass the recognition criteria, including statistical support above chance and accuracy above that floor. Content choices passing both replication and recognition are supported. Full acceptance criteria appear in Appendix D.

4 Experimental setup

Corpora and source labels. We evaluate on HART (Bao et al., 2025), covering news, arXiv introductions, creative writing, and student essays, and a corpus constructed from RAID's public training

release (Dugan et al., 2024), covering eight registers (hereafter RAID). Human-content labels cover original and polished human text (t_0, t_1); LLM-content labels cover humanized and original LLM text (t_2, t_3). These labels remain fixed under expression changes intended to preserve substantive content. All versions within a matched group stay in the same data partition; some writing prompts recur across partitions. Dataset construction and partitions are detailed in Appendix B.1.

Detectors and inputs. We compare separately fitted raw-text and content-point detectors using RoBERTa, DETree (He et al., 2025), and RACE (Li et al., 2026). HART's likelihood-based methods, which infer content source from how a document's wording relates to its simplification, are examined separately in Appendix C.4. Appendix B.4 details input processing and implementations.

Expression changes. We evaluate 36 settings that vary paragraph layout or writing voice during training, testing, or both, alongside an unmodified reference. Paragraph splitting and merging preserve words and their order; GPT-4o and Llama voice rewrites request content preservation. Joint settings retain or reverse the assignment of transformations to content sources, using the same fitted detector and threshold. Appendix B.2 defines the transformations and assignments, with the complete setting grid in Table 7.

Measures and reporting. We report AUROC, with TPR/FPR measured at a threshold calibrated to 5% human-content validation FPR. Under a distribution shift, human-content scores can move across this fixed threshold, changing the test FPR even when source discrimination remains strong. Expression leakage measures score separation between production conditions within each content-source class. Metrics weight domains equally; the primary HART comparisons average three seeds. Appendix B.3 details metrics, calibration, and aggregation.

CRO evaluation. CRO uses separate discovery, naming, and held-out group partitions from HART's development material, with domain-specific and pooled repertoires and population comparisons within each domain (Appendix D.2). Baselines are SAE-TM's native words and topics (Girrbach & Akata, 2026) and SAE-ATD's steering descriptions (Kuznetsov et al., 2025), with random-feature controls for SAE-ATD (Appendices D.3 and D.4).

5 Results

5.1 Content points sustain source discrimination when expression changes

Near-perfect detection can hide expression dependence. Consider training with human-content documents split into short paragraphs and LLM-content documents merged into one paragraph. When this assignment is retained at test time, raw-text RoBERTa and DETree both reach AUROC 1.000. Reversing only the assignment at test time reduces their AUROCs to 0.002 and 0.019 (Figure 2): the same fitted models rank content sources almost entirely backward, although the source words and their order are unchanged, and no threshold can repair the reversed ranking. Their content-point counterparts hold steady (RoBERTa 0.971 to 0.966; DETree 0.965 to 0.965). Near-perfect performance under a persistent expression–source association therefore need not indicate reliable content-source discrimination.

The improvement extends from ranking to decisions. Across all 36 settings, content points raise mean AUROC from 0.712 to 0.960 for RoBERTa and from 0.797 to 0.955 for DETree. Raw-text predictions invert in seven and six settings, respectively; neither content-point counterpart inverts, with lowest setting-level AUROCs (averaged over domains and seeds) of 0.935 and 0.937. At the validation-calibrated operating point, RoBERTa's mean TPR/FPR changes from 0.750/0.485 to 0.867/0.071, and DETree's from 0.815/0.392 to 0.843/0.048, so fewer false positives come with higher recall. Mean expression leakage also falls, from 0.134 to 0.066 for RoBERTa and from 0.144 to 0.072 for DETree. The cost on unmodified text is modest: AUROC changes from 0.994 to 0.970 for RoBERTa and from 0.992 to 0.962 for DETree. Ranking does not guarantee threshold transfer: when only human-content training and validation documents receive the GPT-4o storyteller rewrite, content-point RoBERTa keeps AUROC 0.948 but reaches FPR 0.289 on unmodified test documents.

The improvement extends to RACE (Appendix C.1); across all three detectors, mean AUROC rises from 0.733 to 0.956 over the 36 HART expression-change settings.

The ablations explain the representation's benefits. On unmodified HART, content points outperform central-claim summaries across all three learned detectors, supporting the value of retaining

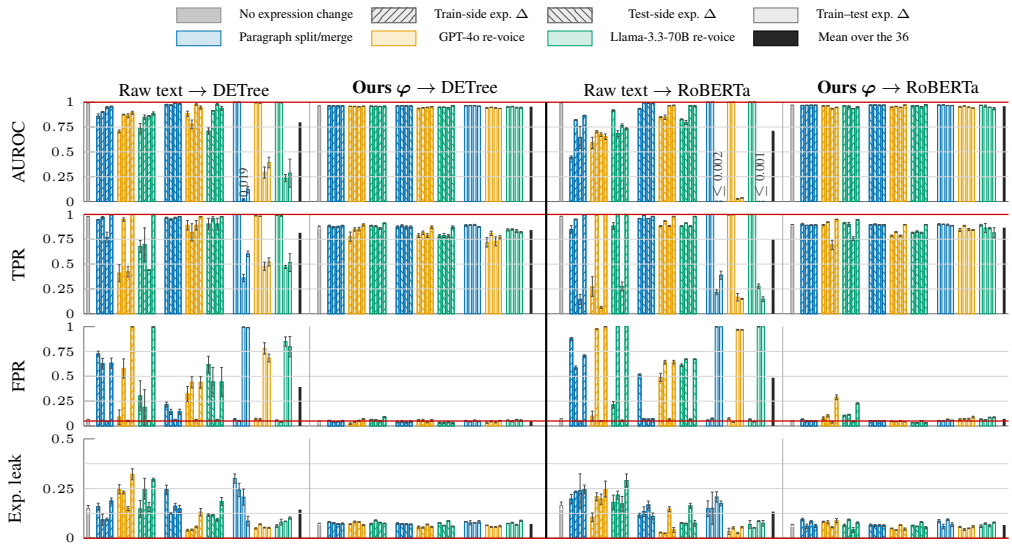

Figure 2: **Content points substantially improve robustness across expression changes.** Each column shows the unmodified reference, 12 training-only, 12 test-only, and 12 joint settings, followed by their mean. Colors identify intervention families and patterns identify scope. Bars average four HART domains equally and then three seeds; whiskers show one sample standard deviation, except on the 36-setting mean. TPR/FPR use fixed thresholds targeting 5% validation FPR. Red lines mark AUROC/TPR 1, FPR 0.05, and leakage 0.

substantive detail. In RoBERTa tests, detailed plain rewriting nearly matches content points' robustness to voice changes, while point extraction adds robustness to paragraph structure: under the human-split/LLM-merged test assignment, FPR rises by only 0.002 from the unmodified reference, compared with 0.086 after plain rewriting; the points also supply CRO's units of analysis (Appendix C.2). Baseline source scores can follow rewrite conditions despite whitespace normalization (Appendix C.6). Changing the training objective does not substitute for the representation: under binary supervision and the tested adversarial approaches, even with adversaries selected using test performance, raw-text detectors still fail under reversals that content points withstand (Appendix C.3).

Transfer and sensitivity to substance. Trained on three HART domains and tested on the fourth, content-point RoBERTa reaches mean AUROC 0.875 across expression changes, versus 0.657 on raw text; calibration with known human-content target documents reduces its mean FPR from 0.138 to 0.052 without retraining (Appendix C.5). With human documents and the threshold unchanged, removing evaluations from LLM-content points reduces a fixed RoBERTa detector's TPR from 0.896 to 0.760, whereas rewording instructed to preserve their claims yields 0.917 (Appendix C.7). The contrast supports a contribution from evaluative commitments, although the edits do not perfectly isolate them from wording.

5.2 CRO DISCOVERS SPECIFIC, RECURRING CONTENT-CHOICE DIFFERENCES

CRO reveals specific, recurring differences in what human and LLM writing conveys, spanning quantitative details, technical conditions, evaluations, and narrative developments. Across the four domains, CRO identifies 32–54 content choices per repertoire with both replicated population differences and descriptions supported by held-out recognition (Appendix Table 33). Learned separately from the detectors, these findings characterize populations rather than explain detector decisions.

Discovering recurring choices across topics. With shared topic variation reduced before repertoire learning, CRO recovers content choices that recur across different subjects. In news, "quantitative ratios or metrics" brings together debt-to-equity ratios, demographic proportions, and test positivity rates: a specific informational choice shared across otherwise different stories, which human writing occupies more often within matched topics. The centering ablation supports this design: removing centering shifts the returned vocabulary toward subjects and entities (Appendix D.5). The same procedure recovers technical and narrative choices without predefined domain categories (Table 1); Figure 1 illustrates the arXiv findings on specific technical details and dynamic adaptation.

Domain	Human-favored	LLM-favored
News	$4.5\times$ quantitative ratios or metrics $4.1\times$ specific dates or timeframes	$7.0\times$ emphasizing importance $5.3\times$ environmental sustainability
arXiv	$5.3\times$ specific technical details $4.9\times$ formal mathematical definitions or assumptions	$5.3\times$ dynamic adaptation $3.5\times$ collaboration and interdisciplinarity
Writing	$2.7\times$ use or presence of firearms $2.5\times$ physical injury and its effects	$5.0\times$ transformation and renewal $4.4\times$ patterns and symbols

Table 1: **Supported content choices across domains.** Prefixes give the favored population's discovery odds ratio, from corrected within-topic membership odds. Writing: creative writing; essays and pooled repertoire: Appendix D.1. Choices may overlap; names link to source-labeled examples.

SAE-TM: ranked words	**SAE-ATD: returned names**
Feature: three, third, year, quarter, record, million, approximately, previous, total, least, iii, percent	counterarguments emphasis
Feature: sustainable, sustainability, environmental, environmentally, green, conscious, carbon, adopt, renewable, friendly, promote, transportation	multilingual insertion
Merged topic: stock, price, earnings, dividend, trading, volatility, target, volume, capitalization, pricing, market, fiscal	rebuttal emphasis
Merged topic: incident, police, investigation, woman, man, death, court, scene, suspect, victim, case, charge	emphasize argument support argumentative structure argumentative structure argument structure run-on-technical scientific argumentation counterargument emphasis

Table 2: **Ranked words and recurring steering descriptions from the compared methods (news outputs).** SAE-TM: first 12 ranked words for two features and two separately output merged topics. SAE-ATD: all ten detector-selected names under the common essay probe, repetitions retained and underscores shown as spaces. The panels are independent; links lead to examples (Appendix E.4).

Learning shared choices before comparing populations. CRO's repertoire retains choices that occur in both populations while distinguishing their occupancy patterns: human-favored ratios and dates, for example, alongside LLM-favored importance claims and sustainability. Learning without source labels supports this differentiated analysis. In the news ablation, adding source supervision strengthens population associations but produces more repeated broad descriptions of factual detail and significance (Appendix D.5). A repertoire learned jointly from all four domains also yields shared descriptions with domain-specific associations: "emphasis on collective progress and shared values" is supported and LLM-favored separately in news, arXiv, creative writing, and essays.

Interpreting the recovered differences. Table 2 contrasts the interpretations supplied by existing methods. SAE-TM recovers related memberships but characterizes them through ranked feature words and merged topics. SAE-ATD achieves strong news source discrimination, yet its selected features repeatedly receive broad argumentation-related steering names under the common essay probe. CRO makes specific recurring content choices explicit and tests whether their descriptions identify unseen members, separately from testing the population difference (Appendices D.3–D.6). Relearning CRO across five seeds yields substantial correspondence among active directions and partial membership overlap; for features with replicated population differences, discovery-matched counterparts generally retain the favored population (Appendix D.7.3).

6 Conclusion

Our results show that detailed content choices provide a productive basis for detecting and interpreting human–LLM differences. Restating those choices in a common register substantially improves content-source discrimination under expression changes, including shifts that defeat the tested training alternatives. Complementing detection, CRO discovers recurring differences in how often shared content choices are expressed across topics, without predefined domain-specific categories. These findings show that substantial overlap between human and LLM content can coexist with specific, supported population differences. Although content and expression remain closely intertwined, this difficulty does not preclude more robust detection or systematic interpretation. Together, our methods make progress on both: identifying content source across forms of expression and making recurring differences in what human and LLM writing conveys explicit and testable.

AI USE STATEMENT

The authors developed the conceptual framework, designed the methodology and experiments, and directed the analysis. AI agents assisted with implementing author-specified methods and experimental procedures. The authors reviewed the code and empirical results and used the resulting evidence to refine their hypotheses. AI-assisted discussions helped visualize and present empirical comparisons; the authors assessed those suggestions and determined the interpretations and conclusions.

As part of the experimental methodology, LLMs generated texts and rewrites, enabling comparisons with human writing on matched topics under controlled rewriting conditions. They also extracted content points, proposed names and descriptions for learned content choices, and performed recognition-based evaluations under author-designed procedures. These uses are documented in the methods and appendices. Dataset-related uses involved generation, rewriting, extraction, and implementation of preprocessing, rather than a separate LLM-based cleaning task.

The authors conducted the primary computer-science literature review, with AI assistance supporting a supplementary survey of linguistic and philosophical perspectives on content and expression. AI also assisted with drafting, organizing, and editing portions of the manuscript; the authors substantially revised the text and checked its claims against the experimental evidence. AI was not used to originate mathematical claims or develop proofs; this work contains no proofs, and no translation was needed. The authors reviewed all code, results, and writing and take responsibility for the final work.

ETHICS STATEMENT

This work improves the detection of LLM-sourced content under changes in expression. At the same time, our analysis shows that human and LLM content choices overlap substantially: meaningful choices occur in both populations and differ in how often they are included. Detection outputs alone are therefore not a sound basis for judging that a particular piece of writing is machine-generated, since wrongly attributing human writing to an LLM can cause real harm. Any deployment of our method, or a derivative of it, should proceed with caution and clear communication to both developers and users.

CRO's findings could in principle inform attempts to make LLM writing appear human. They are population-level differences in inclusion rates, expressed as odds ratios, not rules that any single document follows; and our content representation is designed to remain informative when expression changes.

All corpora are built from publicly released datasets (Bao et al., 2025; Dugan et al., 2024); we use them in accordance with their licenses.

REPRODUCIBILITY STATEMENT

Appendix B specifies corpora, partitions, intervention design, detector training, and evaluation, and Appendix D specifies CRO's fitting, statistical tests, naming, and held-out validation; all prompts are included. The primary HART comparisons for RoBERTa, RACE, and DETree are averaged over three seeds and report the sample standard deviation. The detection and original CRO experiments used NVIDIA L40S GPUs with 4 vCPUs and 32 GB of RAM. The additional five-seed CRO refitting study (Appendix D.7.3) used an NVIDIA RTX 4090 with 24 GB of GPU memory, regenerating Harrier embeddings from the saved content points and fixing those embeddings across fitting seeds.

LLM API outputs cannot be regenerated exactly, so we will release the generated texts themselves: our four-condition corpus constructed from RAID, all expression-changed texts for HART and RAID, and all extracted content points, in accordance with the original datasets' licenses. We will also release our code. Together, these allow our detection and CRO results to be re-analyzed without repeating any generation.

REFERENCES

Aristotle. *On rhetoric: a theory of civic discourse*. Oxford University Press, New York, 2006. ISBN 9780195305098.

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. *ArXiv*, abs/2310.05130, 2023.

Guangsheng Bao, Lihua Rong, Yanbin Zhao, Qiji Zhou, and Yue Zhang. Decoupling content and expression: Two-dimensional detection of ai-generated text. *ArXiv*, abs/2503.00258, 2025.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`, 2023.

Changyou Chen, Wray L. Buntine, Nan Ding, Lexing Xie, and Lan Du. Differential topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:230–242, 2015.

Ziyu Chen, Yilun Zhao, and Arman Cohan. Measuring the gap between human and llm research ideas. *ArXiv*, abs/2607.01233, 2026.

Cicero. *Cicero on the ideal orator*. Oxford University Press, New York, 2001. ISBN 9780195091984.

Anil R. Doshi and Oliver P. Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290. URL `https://www.science.org/doi/abs/10.1126/sciadv.adn5290`.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.674. URL `https://aclanthology.org/2024.acl-long.674/`.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL `http://jmlr.org/papers/v17/15-239.html`.

Leander Girrbach and Zeynep Akata. Sparse autoencoders are topic models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Adele Goldberg. *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, Chicago, 1995. ISBN 9780226300863.

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. In *International Conference on Machine Learning*, 2024.

Yongxin He, Shan Zhang, Yixuan Cao, Lei Ma, and Ping Luo. Detree: Detecting human-ai collaborative texts via tree-structured hierarchical representation learning. *ArXiv*, abs/2510.17489, 2025.

Louis Hjelmslev. *Prolegomena to a theory of language*. University of Wisconsin Press, Madison, Wis, 1961. ISBN 9780299024703.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *ArXiv*, abs/2307.03838, 2023.

Xi Yu Huang, Krishnapriya Vishnubhotla, and Frank Rudzicz. The gpt-writingprompts dataset: A comparative analysis of character portrayal in short stories, 2024. URL https://arxiv.org/abs/2406.16767.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205, March 2022. doi: 10.1162/coli_a_00426. URL https://aclanthology.org/2022.cl-1.6/.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. Reformulating unsupervised style transfer as paraphrase generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 737–762, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.55. URL https://aclanthology.org/2020.emnlp-main.55/.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *ArXiv*, abs/2303.13408, 2023.

Kristian Kuznetsov, Laida Kushnareva, Anton Razzhigaev, Polina Druzhinina, Anastasia Voznyuk, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. Feature-level insights into artificial text detection with sparse autoencoders. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25727–25748, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1321. URL https://aclanthology.org/2025.findings-acl.1321/.

Yongchan Kwon and James Zou. What llms think when you don't tell them what to think about? *ArXiv*, abs/2602.01689, 2026.

Ronald Langacker. *Cognitive Grammar: A Basic Introduction*. Oxford University Press, 2008. ISBN 9780195331967. doi: 10.1093/acprof:oso/9780195331967.001.0001. URL https://doi.org/10.1093/acprof:oso/9780195331967.001.0001.

Ronald W. Langacker. Working toward a synthesis. *Cognitive Linguistics*, 27:465–477, 2016.

Willem J. M. Levelt. *Speaking: From Intention to Articulation*. The MIT Press, 1989. ISBN 9780262278225. doi: 10.7551/mitpress/6393.001.0001. URL https://doi.org/10.7551/mitpress/6393.001.0001.

Yang Li, Qiang Sheng, Zhengjia Wang, Yehan Yang, Danding Wang, and Juan Cao. Beyond the final actor: Modeling the dual roles of creator and editor for fine-grained LLM-generated text detection. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5188–5203, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.235. URL https://aclanthology.org/2026.acl-long.235/.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, 2018.

George Ma, Samuel Pfrommer, and Somayeh Sojoudi. Revising and falsifying sparse autoencoder feature explanations. *Advances in Neural Information Processing Systems 38*, 2025.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, 2023.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 642–669, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/02dec8877fb7c6aa9a79f81661baca7c-Paper-Conference.pdf.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48393–48421. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/paulo25a.html.

Roberta Pires De Oliveira. Language and ideology: An interview with george lakoff. In *Language and Ideology*, pp. 23–47. John Benjamins Publishing Company, February 2001. ISBN 9789027299543. doi: 10.1075/cilt.204.02pir. URL http://dx.doi.org/10.1075/cilt.204.02pir.

EHUD REITER and ROBERT DALE. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997. doi: 10.1017/S1351324997001502.

Rafael Alberto Rivera Soto, Barry Y. Chen, and Nicholas Andrews. Mitigating paraphrase attacks on machine-text detection via paraphrase inversion. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4421–4433, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.227. URL https://aclanthology.org/2025.findings-acl.227/.

Margaret E. Roberts, Brandon M Stewart, and Dustin Tingley. stm: An r package for structural topic models. *Journal of Statistical Software*, 2019.

Jenna Russell, Rishanth Rajendhran, Chau Minh Pham, Mohit Iyyer, and John Wieting. Storyscope: Investigating idiosyncrasies in ai fiction. *ArXiv*, abs/2604.03136, 2026.

Jacqueline Sachs. Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, 2:437–442, 1967.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *ArXiv*, abs/2303.11156, 2023.

Rafael A. Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, and Nicholas Andrews. Few-shot detection of machine-generated text using style representations. *ArXiv*, abs/2401.06712, 2024.

Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. (Re)construing meaning in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5170–5184, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.462. URL https://aclanthology.org/2020.acl-main.462/.

Ludwig Wittgenstein. *Philosophical investigations*. Wiley-Blackwell, Malden, MA, 2009. ISBN 9781405159289.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. LLM tropes: Revealing fine-grained values and opinions in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 17085–17112, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.995. URL https://aclanthology.org/2024.findings-emnlp.995/.

Sergio E. Zanotto and Segun Aroyehun. Linguistic and embedding-based profiling of texts generated by humans and large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22841–22858, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1163. URL https://aclanthology.org/2025.emnlp-main.1163/.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 409–436, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.29. URL https://aclanthology.org/2024.findings-naacl.29/.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating language models for humanlike diversity, 2025. URL https://arxiv.org/abs/2504.05228.

A Overview of additional findings

The additional experiments show how content points reduce expression sensitivity while retaining detailed distinctions useful for source detection and the study of recurring human–LLM content choices. They connect these benefits to the representation's design and provide evidence that substantive commitments contribute to detection.

Changing the representation sustains discrimination under severe expression shifts. Content-point inputs retain strong discrimination under expression assignments that nearly reverse raw-text predictions. The advantage extends across three learned detectors on HART and a separate RAID evaluation, improving mean discrimination and reducing mean false-positive rates. Binary supervision leaves severe reversals intact, as do the tested adversarial alternatives, even when their configurations are selected using test performance. These comparisons show why specifying content source as the target does not determine which evidence a detector learns to use (Appendices C.1 and C.3). The HART comparison further supports pairing extraction with detectors that can use the retained detail: its likelihood features become less expression-sensitive but lose average discrimination (Appendix C.4).

Detail, plain restatement, and point structure provide complementary benefits. Detailed points outperform central-claim summaries, while detailed plain rewriting recovers much of the tested voice robustness. Point extraction adds robustness to paragraph structure: when human-content documents are split and LLM-content documents merged, RoBERTa's false-positive increase from each representation's unmodified reference is 0.002 for content points, versus 0.086 for plain rewriting. The points also supply CRO's units of analysis, motivating a shared representation for detection and interpretation (Appendix C.2).

Re-voicing can make source scores track rewrite conditions. A direct diagnostic separates source discrimination under a common rewrite from rewrite discrimination on original-human documents. Source-associated training strengthens the latter while weakening the former for raw RoBERTa. Common-layout controls and consistently normalized training show that paragraph formatting alone does not account for the vulnerability; content points retain substantially stronger source discrimination (Appendix C.6).

Detection generalizes while remaining sensitive to substantive content. Content-point RoBERTa retains useful discrimination in unseen domains, including under expression changes. Domain-dependent score distributions can shift the operating threshold even on unmodified documents; calibration with 100 known human-content target documents brings mean FPR across domains and expression changes from 0.138 to 0.052 without retraining (Appendix C.5). Separately, targeted removal of evaluative commitments reduces detection evidence, while rewording instructed to preserve them retains strong discrimination. This provides evidence of substantive sensitivity alongside reduced expression dependence, although the edits do not perfectly separate commitments from wording (Appendix C.7).

CRO makes differences within shared content choices explicit and testable. CRO recovers recurring differences in quantitative detail, technical conditions, evaluations, and narrative development across topics. Its comparisons distinguish recovering related content from explaining the population difference: SAE-TM learns features with overlapping memberships, while SAE-ATD achieves strong source discrimination but returns recurring steering descriptions for both selected features and our random controls. CRO supplies specific content-choice descriptions with held-out recognition and separately replicated population differences (Appendices D.1–D.4).

The learning and naming comparisons connect these findings to the method's design. Centering reduces topic dominance; adding source supervision can produce repeated broad descriptions despite strong population associations. Naming from examples spread across topics can yield broader descriptions than concentrating on high-activation examples, even when recognition accuracy is similar (Appendices D.5 and D.6).

The reliability analyses first hold memberships fixed: the favored population largely persists across original and polished/humanized texts and under an alternative occupancy estimator. Relearning the repertoire across five seeds then shows substantial correspondence among active directions and partial membership overlap. For features with replicated population differences, discovery-matched counterparts generally retain the favored population (Appendix D.7).

Reading guide. Appendix B gives shared conventions; Appendices C–D, experiments and prompts; Appendix E, CRO and baseline examples; Appendix F, full source/rewrite pairs with selected content points and commentary.

Condition	Production history	Content-source label
t_0	Original human text	Human
t_1	LLM-polished or rephrased human text	Human
t_2	Humanized LLM text	LLM
t_3	Original LLM text	LLM

Table 3: **The four production conditions share a binary content-source task.** HART's humanization includes LLM rewriting, commercial tools, and human editing; RAID uses LLM rewriting only. The t_2 label therefore does not imply that a human wrote the final wording.

B Experimental setup and implementation

This section specifies the corpora and partitions, intervention design, reporting conventions, and detector implementations underlying the comparisons in Section 5.1 and Appendix C. The central comparison changes the detector's input from raw text to content points while retaining its training recipe and content-source scoring rule. We distinguish the datasets' production conditions from the additional interventions used to test robustness. CRO draws on the same HART development material but uses separate partitions and evaluation procedures, described in Appendix D.

B.1 Datasets, production conditions, and partitions

Production conditions and source labels. Each matched group contains four documents associated with the same human source and its title or writing prompt (Table 3). All four remain together when groups are partitioned. The binary task assigns human content to t_0, t_1 and LLM content to t_2, t_3. These labels record production history: humanized LLM text remains in the LLM-content class even when a person edits it. They do not certify that polishing or humanization preserves every substantive claim. Appendix C.7 examines realized changes separately. The four production conditions are part of the corpus construction, before our additional expression interventions.

HART. We use HART's four English domains (Bao et al., 2025): CC-News articles, arXiv computer-science introductions, WritingPrompts stories, and ASAP student essays. The benchmark supplies original human texts, LLM generations conditioned on their titles or prompts, and rewritten versions. Its six original-generation models are GPT-3.5 Turbo (0125), GPT-4o (2024-11-20), Claude 3.5 Sonnet (2024-10-22), Gemini 1.5 Pro (002), Qwen 2.5 72B Instruct, and Llama 3.3 70B Instruct. Rephrasing uses polishing or restructuring; humanization uses LLM diversification or imitation, commercial humanization tools, and, in the test partition, human editing. We retain the released documents and HART's development/test boundary.

We divide HART's development groups within each domain into training and validation by a fixed group shuffle, leaving the released test groups untouched (Table 4). Group separation does not imply universal prompt separation: the same five essay assignments occur in training, validation, and testing, and some creative-writing prompts are shared between training and validation. Creative-writing development and test prompts are disjoint.

A four-condition corpus constructed from RAID. We construct a separate corpus from the attack-free portion of RAID's public training release (Dugan et al., 2024). It covers abstracts, books, news, poetry, recipes, Reddit, reviews, and Wikipedia. This construction uses neither RAID's official test set nor the RAID subset distributed with HART. We select human sources with both native GPT-4 and ChatGPT generations. Both native classes are restricted to 120–1,200 words and deduplicated by normalized text; LLM generations must use sampling without a repetition penalty and differ from the paired human source. A fixed deterministic assignment selects one native generation as t_3 for each source; RAID records the GPT-4/ChatGPT labels but not their exact API snapshots.

We add t_1 by polishing or restructuring the paired human text, and t_2 by diversifying or imitating human expression in the selected t_3, using HART's corresponding instructions. Imitation uses the paired t_0 as its reference. The rewriting models are GPT-3.5 Turbo (0125), GPT-4o (2024-11-20), Claude Sonnet 4.5, and Gemini 2.5 Flash; commercial tools and human editing are omitted. We retain complete four-condition groups after generation. Length processing is asymmetric: t_1, t_2 are truncated at sentence boundaries relative to the human source, with a target cap of 1.1 times its

Detection corpus	Domains	Train groups	Validation groups	Test groups	Docs/group
HART	4	1,752	248	2,000	4
RAID	8	978	136	160	4

Table 4: **Canonical detection populations, counted in matched groups.** HART has 500 development groups per domain, split into 438 training and 62 validation groups, and 500 test groups. RAID has 1,274 groups in total, with 121–123/17/20 train/validation/test groups per register. Each group contains two human-content and two LLM-content documents. These are the starting populations; setting-specific availability can reduce the number scored.

word count, while native t_3 is not length-matched to t_0. Some short rewrites remain, so the native 120–1,200-word filter is not a bound on every final document.

Detection partitions and available populations. Table 4 gives the canonical counts, before additional expression interventions. RAID's partitions are fixed, group-disjoint, and stratified by register. Validation supports checkpoint selection where the detector uses it and calibration of every detector's operating threshold; the test partition is separate from both. The exact membership and construction settings are retained in the reproducibility records.

Incomplete rewrites or extractions cause small reductions in some intervention settings, and detector preprocessing can further restrict the scored population. HART's creative-writing test population can fall from 500 to 495 groups; the RAID test transformations retain all 20 groups per register. Metrics use the available scored documents under the conventions in Appendix B.3.

CRO's corpus boundary. CRO uses the 500 HART development groups per domain, repartitioned for repertoire discovery, additional naming examples, and held-out evaluation. It does not use the detection test partition or inherit the detection train/validation division. Its realized group and point counts, sentence encoder, and information flow are specified with the CRO procedure in Appendix D.

B.2 Expression-intervention design

The design uses three transformation families, each with four training-only, four test-only, and four joint settings: 36 settings in total, plus an unmodified reference. These separate changes during training, changes at test time, and reversals of learned expression–source associations. Every transformation acts on the document in its existing production condition, so intervening on t_1 rewrites the already-polished human text. Content-source labels remain fixed.

Three transformation families. Each family pairs transformations A and B (Table 5). Paragraph edits preserve words and their order; both voice families request preservation of the original facts and details.

Family	Transformation A	Transformation B	Execution
Paragraph	Sentence-per-paragraph split	Merge paragraphs	Deterministic whitespace edits
GPT-4o re-voice	Casual rewrite	Storyteller rewrite	GPT-4o (2024-11-20)
Llama re-voice	Casual rewrite	Storyteller rewrite	Llama 3.3 70B Instruct

Table 5: **Transformation pairs used throughout the 36-setting evaluation.** The two voice families use identical prompts, an empty system message, temperature 0.7, top-p 1, and a maximum of 2,500 output tokens. Their executor differs; paragraph transformations use no language model.

Assignments to content-source classes. Table 6 defines four source assignments. Both production conditions within a source class receive the indicated transformation independently. A dash retains the existing dataset text, including any polishing or humanization.

Training, testing, and checkpoint reuse. Table 7 gives all train/test combinations. Training-only settings apply I, R, M, or H to training and validation; test-only settings apply them only at test time. Joint settings retain the assignment (I/I or R/R) or reverse it (I/R or R/I). Training interventions always apply to validation for both checkpoint selection and calibration.

Assignment	Human content (t_0, t_1)	LLM content (t_2, t_3)
I	A	B
R	B	A
M	–	B
H	B	–

Table 6: **Assignments specify which transformation each source class receives.** I and R exchange A and B between the two classes. M changes only LLM-content documents; H changes only human-content documents. The dash means no additional intervention.

All standard test-only settings reuse the unmodified-training checkpoint and its threshold. Each I- or R-trained checkpoint is evaluated on unmodified, I, and R test inputs with its threshold unchanged. Thus matched/reversed comparisons change only the test assignment. Checkpoint reuse is within a detector, input representation, and seed; the condition-specific adversarial fits in Appendix C.3 are identified locally.

Reading the full setting grid. The figure order is training-only, test-only, then joint, with paragraph, GPT-4o, and Llama families within each block. Family colors and scope patterns follow this order; Table 7 maps every bar to its assignments. Uniform-renderer controls and other focused diagnostics in Appendix C.7 are separate from the 36-setting evaluation.

	Intervention position in the figure			Assignment	
Scope	Paragraph	GPT-4o	Llama	Train/validation	Test
Training only	1	5	9	I	–
	2	6	10	R	–
	3	7	11	M	–
	4	8	12	H	–
Test only	13	17	21	–	I
	14	18	22	–	R
	15	19	23	–	M
	16	20	24	–	H
Joint	25	29	33	I	I
	26	30	34	R	R
	27	31	35	I	R
	28	32	36	R	I

Table 7: **Complete key to the 36 intervention settings.** Positions count left to right after the canonical reference and exclude the final mean bar. Read each family's numbered column from top to bottom; the three columns together identify all 36 positions. Assignments follow Table 6; a dash leaves the partition canonical. In the joint block, I/I and R/R are matched, while I/R and R/I reverse the association. The same key applies to raw text and **Ours** ϕ.

Transformation execution. Paragraph splitting places each detected sentence in its own paragraph; merging removes paragraph breaks. Sentence segmentation retains the original substrings and accounts for common abbreviations and initials. Both operations preserve the characters and their order after whitespace is collapsed. Their effect depends on the source layout: merging can leave an already single-paragraph document unchanged.

GPT-4o uses the OpenAI API; Llama uses OpenRouter without a fixed underlying inference provider. The storyteller prompt additionally forbids additions and specifies a document-dependent length band. These instructions constrain the requested rewrite, but do not establish that its realized content is unchanged; Appendix C.7 examines additions and weakened commitments. The alternative storyteller formulation analyzed there is separate from the three families in Table 5.

Re-voice prompts and generation. The following user messages are shared by both voice executors; the system message is empty. We substitute the document for $\{\texttt{text}\}$. For a source of w whitespace-delimited words, $\{\texttt{lo}\}$ is $\max(40, 0.8w)$ rounded to the nearest ten, and $\{\texttt{hi}\}$ is $1.2w$ rounded to the nearest ten, raised to $\{\texttt{lo}\}$ if necessary. Only the storyteller message includes this requested range. Generation settings are given in Table 5.

Replies receive a separate length check, with bounds $\max(20, \lfloor 0.4w \rfloor)$ and $\max(80, \lfloor 2.5w \rfloor)$ words. A reply failing this check receives one retry; the final reply is retained on a best-effort basis, so neither the requested band nor the check guarantees a final length bound. Fixed cleanup removes wrapping code fences or quotation marks and matching introductory assistant lines, and normalizes curly apostrophes.

Casual rewrite (A).

```
Rewrite the text below in a casual, conversational register, as if
↪ explaining it to a friend in everyday spoken language. Follow these
↪ rules:
- Use relaxed, informal wording and a natural conversational flow.
- Output only the rewritten text, with no preamble.
- Preserve every fact, claim, example, number, name, and detail of the
↪ original; do not drop, simplify, shorten, or summarize anything ---
↪ every point of the original must still appear.

TEXT:
{text}
```

Storyteller rewrite (B).

```
Rewrite the text below in the voice of a talkative storyteller speaking
↪ out loud to an audience --- warm, plain-spoken, full of everyday idiom
↪ . Follow these rules:
- Every sentence must sound SPOKEN, not written: contractions throughout,
↪  second person, discourse markers ("look", "now", "see", "here's the
↪ thing"), the occasional sentence fragment or rhetorical question, and
↪ the loose rhythm of somebody telling you about it in person.
- Do not use formal, literary, or academic wording; no ornate imagery and
↪  no elevated vocabulary. Replace bookish words with the plainest
↪ spoken equivalent.
- Do this by REPLACING the original's phrasing, never by adding to it. -
↪ Add nothing of your own: no new examples, comparisons, asides,
↪ opinions, greetings, sign-offs or commentary. Every sentence must
↪ correspond to something already in the original.
- Write roughly {lo} to {hi} words.
- Output only the rewritten text, with no preamble.
- Preserve every fact, claim, example, number, name, and detail of the
↪ original; do not drop, shorten, or summarize anything --- every point
↪ of the original must still appear.

TEXT:
{text}
```

B.3 Metrics, calibration, and aggregation

AUROC measures content-source ranking; TPR and FPR measure decisions at a fixed threshold; expression leakage measures score separation within a content-source class. Let $s(x)$ denote the detector's score, with larger values indicating LLM content. For four-way RoBERTa and RACE it sums the probabilities of t_2 and t_3; the other detector scores are specified in Appendix B.4. Unless stated otherwise, these conventions apply to both raw-text and content-point inputs.

Content-source ranking. AUROC compares all scored LLM-content documents (t_2, t_3) with human-content documents (t_0, t_1) within a domain. It is the probability that an LLM-content score exceeds a human-content score, with tied scores contributing one half. Scores are not reversed after observing a failure: an AUROC below 0.5 records a reversed ranking. AUROC is independent of the decision threshold.

A validation-targeted operating point. We set a threshold using human-content validation scores and carry it unchanged to the associated test evaluations. TPR is the fraction of LLM-content documents flagged; FPR is the fraction of human-content documents flagged. A distribution shift can

move human-content scores across the fixed threshold, changing test FPR even when ranking remains strong. The 5% validation target therefore does not impose a 5% test FPR, and TPR and FPR must be read together.

Each fitted checkpoint c has its own threshold,

$$\tau_c = Q_{0.95}\big(\{s_c(x) : x \in V_{c,\mathrm{H}}\}\big), \tag{5}$$

where $V_{c,\mathrm{H}}$ contains that checkpoint's available human-content validation documents and $Q_{0.95}$ is the linearly interpolated empirical 95th percentile. We pool validation humans across the training domains, including both t_0 and t_1. For a manipulated training condition, validation follows the same assignment; for content-point input, calibration uses the corresponding extracted text. Thresholds are specific to the detector, representation, training condition, and seed.

We flag a document as LLM content only when $s_c(x) > \tau_c$; equality is not flagged. TPR and FPR are computed within each domain at this same threshold, excluding missing scores from their denominators. Test labels do not determine the threshold. Pooling validation scores for calibration does not pool test scores for AUROC or average separate domain thresholds.

Table 8 records the validation populations used for calibration. Finite samples and distribution shifts both affect achieved rates, including within individual domains. The separate target-domain and per-domain calibration comparisons in Appendix C.5 change thresholds while leaving scores and ranking unchanged.

Calibration population	Documents	Availability
HART, all four training domains	496	Canonical validation
HART, three training domains	372	Held-out-domain fitting
RAID, all eight registers	272	Canonical validation
RAID, GPT-4o voice training under I, R, or H	270	One Reddit validation group unavailable

Table 8: **Human-content validation populations for threshold calibration.** Counts include t_0, t_1 and are distinct from test-set sizes. Other RAID training assignments retain 272 documents; incomplete inputs can reduce available populations.

Expression leakage. For a fixed setting and seed, let $A_d(a, b)$ denote the AUROC of the same source scores within domain d, using production condition b as the positive class and a as the negative class. We define

$$L_d = \frac{1}{2}\left(\left|A_d(t_0, t_1) - \tfrac{1}{2}\right| + \left|A_d(t_2, t_3) - \tfrac{1}{2}\right|\right), \qquad L = \frac{1}{D}\sum_{d=1}^{D} L_d. \tag{6}$$

Absolute departures are taken within each domain and pair before averaging domains or seeds. The statistic ranges from 0 to 0.5 and is threshold-free. It measures score separation between production conditions within a content-source class, rather than all expression information in the input or hidden representation. Rewriting can change substantive content, and finite samples can yield nonzero separation even without a population difference. Low leakage alone therefore does not establish robustness: an imposed cue shared by both conditions within a source class can drive source predictions without separating either pair. Appendix C.3 illustrates this distinction.

Aggregation, extrema, and uncertainty. For each setting and seed, we compute metrics within domains and average domains equally: four on HART and eight on RAID. We then average these domain-macro values over seeds; detector scores are never pooled across seeds. A domain-specific panel retains that domain's metric before averaging seeds. Held-out-domain results retain their named target domain, and any average over held-out domains weights them equally.

The primary HART comparisons for RoBERTa, RACE, and DETree average three seeds. Their whiskers show one sample standard deviation across seeds, rather than a confidence interval. RACE and DETree content-point fits include implementation-version differences, so their whiskers reflect that variation as well as initialization. The plain-rewrite comparison and the training-renderer references also average three seeds. Tables containing differently replicated fits retain row-level seed counts; calibration results instead average draws of calibration documents, as specified locally. Exact seed and implementation identifiers are retained in the reproducibility records.

Detector	Representation and training	Source score and selection
RoBERTa	RoBERTa-base and classifier; four-condition cross-entropy. Train the final encoder layer and head.	$P(t_2) + P(t_3)$; select by validation four-class macro-F1.
RACE	RoBERTa-base with an RST graph; supervised contrastive and cross-entropy losses. Train the final encoder layer, graph components, and head.	$P(t_2) + P(t_3)$; select by validation four-class macro-F1.
DETree	Released RoBERTa-large encoder, adapted with two LoRA contrastive stages. Build the process-class hierarchy and retrieval database from training documents.	Content-source neighbor score; layer and retrieval settings fixed without validation selection.
HART variants	Fast-DetectGPT or Binoculars criterion on the input and its simplification; fit an RBF support-vector regressor to binary labels.	Regressor prediction; fixed fitting settings without validation selection.

Table 9: **Detector recipes shared by raw-text and content-point comparisons.** Each representation receives its own fit. Validation calibrates the operating threshold for every detector, including those without validation-based model selection. All implementations use a maximum input length of 512 tokens with their respective tokenizers.

The mean over expression changes gives equal weight to the 36 setting-level means and excludes the canonical reference; its bar has no whisker. These designed settings are neither independent replications nor a deployment-frequency distribution. Unless stated otherwise, headline minima, maxima, and inversion counts are over setting-level domain-macro values after averaging seeds, not over individual domains or seeds. The smallest AUROC and largest FPR need not occur in the same setting. Full-grid reference lines mark AUROC and TPR at 1, FPR at the 0.05 target, and leakage at 0; the FPR line is a calibration target, not a constraint on the observed test rates.

B.4 Content-point input and detector implementations

Constructing the content-point input. We implement the extractor in Section 3.1 with GPT-4o (`gpt-4o-2024-11-20`), temperature 0.2, top-p 1, and a maximum of 1,024 output tokens. Each document is processed independently, without its source label or rewrite history. We strip leading list numbers and surrounding whitespace, discard empty lines, and join the points in extraction order with a single line break, without bullets or a trailing line break. Extraction follows any expression intervention and applies to training, validation, and test documents. Source labels and group assignments remain unchanged.

Each detector is trained separately on raw text and on **Ours** ϕ with the same recipe. Text-dependent structures are recomputed: RACE parses the content-point text itself, and DETree adapts its encoder and rebuilds its class hierarchy and retrieval database from the corresponding training inputs. HART's likelihood variants also generate a new simplification of each content-point input. The extraction messages at the end of this subsection use $\{\texttt{text}\}$ for the available document.

Training targets and source scores. Table 9 summarizes the evaluated implementations. Every detector is evaluated on binary content source, but the training targets differ. Standard RoBERTa and RACE predict the four production conditions and use

$$s(x) = P(t_2 \mid x) + P(t_3 \mid x) \tag{7}$$

as the LLM-content score, where x is the input representation being evaluated. DETree instead learns process classes and aggregates retrieved neighbors by content source; HART fits directly to binary labels. Binary and adversarial training variants are specified beside their comparisons in Appendix C.3.

RoBERTa and RACE. Our RoBERTa classifier uses the final hidden state of the initial token and a two-layer MLP head, with cross-entropy training. RACE (Li et al., 2026) instead represents discourse units and relations as nodes of a rhetorical structure theory (RST) graph, applies two relational graph-convolution layers, and classifies the root representation. We parse intervened text and content-point inputs with isanlp_rst 3.2.2, using the DMRST `rstdt` checkpoint. Canonical raw documents retain their supplied trees. The shared data loader requires a nonempty tree even for RoBERTa, although that model does not read it.

Both start from RoBERTa-base and update only the final encoder layer and the components above it. We use AdamW with learning rate 2.9×10^{-5}, weight decay 0.01, batch size 16 with stratified sampling, a linear learning-rate schedule with 10% warmup, and gradient clipping at 1. Training lasts at most 20 epochs, with patience 5 and checkpoint selection by validation four-class macro-F1. RACE adds supervised contrastive loss with temperature 0.07 to cross-entropy with weight 1. Its four labels are the corpus conditions in Table 3; both the label mapping and the binary score in Equation 7 follow this evaluation, rather than assuming the released detector's original classification task is identical.

DETree. We continue training DETree's released RoBERTa-large encoder (He et al., 2025), using its two-stage tree-guided supervised contrastive procedure. Training classes retain production condition, generator, and rewriter identities: the canonical training corpora have 41 process classes on HART and 15 on RAID. A first LoRA stage uses a flat class tree; embeddings of the training classes then define a hierarchical affinity tree for the second stage. Each stage runs for six epochs with batch size 8, learning rate decaying from 10^{-5} to 10^{-6}, 50 warmup steps, contrastive temperature 0.07, and LoRA rank 64 and scale 128. The model uses max pooling over token states. We retain its built-in training perturbations: with probability 0.5, a document receives one randomly chosen surface edit, such as spelling, number, character, or whitespace changes. These are part of DETree's recipe, separately from our experimental interventions.

The adapted training documents form the retrieval database; validation and test documents are queries and never database members. Retrieval uses normalized layer-19 embeddings and the ten nearest training neighbors by cosine similarity. We sum $0.05 \cos(q, h_i)$ separately over human-content neighbors to obtain H and over LLM-content neighbors to obtain M, then score $s(x) = e^M/(e^H + e^M)$. Layer, neighbor count, and score scale are fixed. This readout groups t_0, t_1 together and t_2, t_3 together, adapting the released detector's original-human-versus-other distinction to content source. Encoder adaptation and database construction are repeated for each training input representation; the score rule remains the same.

HART likelihood-based variants. Following HART (Bao et al., 2025), each variant computes a criterion on the input, T, and on its simplification, C, and fits a support-vector regressor to the pair (C, T). Fast-DetectGPT (Bao et al., 2023) uses Falcon-7B-Instruct as the scoring model and Falcon-7B as the reference; Binoculars (Hans et al., 2024) uses Falcon-7B as observer and Falcon-7B-Instruct as performer. Both criteria use half-precision arithmetic. The RBF regressor receives unscaled features, replaces missing criterion values by zero, and uses $C_{\mathrm{SVR}} = 1$, $\epsilon = 0.1$, and `gamma=scale`. The two criterion values remain separate features, not their literal difference.

For every available input, we generate one fixed GPT-4o (2024-11-20) simplification using HART's simplification instruction, temperature 1, top-p 1, and a 4,096-token output cap. Both criteria share this draw. On content points, the simplification is generated from the point text itself. These are newly generated simplifications, including for canonical documents, rather than the benchmark's released draws. We fit each regressor on the training partition and calibrate it on held-out validation using the shared threshold convention in Appendix B.3. This replaces HART's published all-development fit and in-sample maximum-F1 threshold. Appendix C.4 examines expression dependence in these features and the effect of content extraction.

Content-extraction prompt. *System message.*

```
You are a careful analyst. You extract the substantive content of a text
↪ and restate it plainly, ignoring the author's writing style.
```

Task message.

```
List the substantive content of the text below, in the order it appears.
↪ Follow these rules:
- Write each point as one self-contained sentence in plain, neutral
↪ language.
- Capture what the text conveys (what it claims, argues, or proposes),
↪ not how it is expressed. Do not reuse the author's distinctive
↪ phrasing, tone, imagery, or stylistic quirks.
- Include every distinct substantive point; do not merge restatements
↪ together and do not drop detail as "filler".
- Be exhaustive rather than concise: roughly one line per distinct point,
↪  not a summary.
- Output only a numbered list, one point per line, with no preamble or
↪ commentary.

TEXT:
{text}
```

C DETECTION: ADDITIONAL RESULTS AND ABLATIONS

Content-point inputs substantially strengthen content-source detection under expression changes that cause severe failures in raw-text detectors. Across three learned detectors on HART and a separate RAID evaluation, they improve mean discrimination and reduce mean false-positive rates; none exhibits the setting-level, domain-averaged ranking reversals observed on raw text. The representation ablations show how preserving substantive detail and reducing variation in expression contribute together to this gain. The training comparisons show that severe reversals persist under binary supervision and the tested adversarial objectives, even when adversarial configurations are selected using test performance. The likelihood-feature comparison further motivates using detectors that can learn from the retained content. We then demonstrate retained robustness in held-out domains and investigate sensitivity to substantive commitments. Shared experimental conventions are specified in Appendix B.

C.1 ROBUSTNESS ACROSS DETECTORS AND DATASETS

The robustness advantage extends across RACE, RoBERTa, and DETree on HART and RAID (Figures 3–5). Each detector is trained and evaluated within its corpus; RAID provides a separate replication. The comparison uses the shared 36-setting design and reporting conventions (Appendix B).

All three learned detectors retain high discrimination on HART. For RACE, content points raise mean AUROC from 0.690 to 0.954 and change mean TPR/FPR from 0.755/0.482 to 0.854/0.073. Mean expression leakage falls from 0.199 to 0.093. Raw-text RACE inverts its ranking in ten of the 36 settings, compared with seven for RoBERTa and six for DETree; none of the content-point counterparts inverts. Across the three detectors, the smallest setting-level domain-average AUROC on content points is 0.925. Thus the improvement spans all three architectures and prevents the severe ranking reversals observed on raw text. RACE's unmodified-reference AUROC changes from 0.992 to 0.966. A remaining threshold failure occurs when the GPT-4o storyteller rewrite is applied only to human-content training and validation documents: content-point RACE then has test FPR 0.317 on unmodified text.

RAID replicates the ranking and false-positive gains. On RAID, mean AUROC rises from 0.714 to 0.930 for RACE, from 0.716 to 0.937 for RoBERTa, and from 0.880 to 0.948 for DETree. Raw-text RACE and RoBERTa invert in eight and six settings. Raw-text DETree is less disrupted here than on HART, but still inverts once, reaching AUROC 0.320. Content points remove these

domain-average inversions for all three detectors; their minimum AUROCs are 0.864, 0.881, and 0.913, respectively.

At the fixed validation-calibrated thresholds, mean TPR/FPR changes from 0.752/0.424 to 0.751/0.070 for RACE, from 0.733/0.413 to 0.809/0.073 for RoBERTa, and from 0.860/0.311 to 0.835/0.050 for DETree. The large false-positive reductions accompany nearly unchanged recall for RACE, higher recall for RoBERTa, and lower recall for DETree. Mean expression leakage also decreases for all three. These gains under expression changes come at a larger unmodified-text cost than on HART: reference AUROC decreases by 0.041–0.050, and RACE's reference TPR falls from 0.991 to 0.747.

The gain also has domain-level variation: across detectors, content-point mean AUROC ranges from 0.866 to 0.892 for recipes and from 0.885 to 0.906 for Reddit, two weaker RAID registers. Each register has 40 human-content test documents, so a single changed decision moves FPR by 0.025. These small samples limit domain-level precision and raise the finite-sample floor of expression leakage; its absolute levels should therefore not be compared directly with HART's. Corpus sizes are given in Table 4.

The ranking advantage holds within each HART domain. Figures 6 and 7 show DETree separately in the four HART domains. Content-point mean AUROC is 0.967 in news, 0.975 in arXiv, 0.942 in creative writing, and 0.937 in essays, compared with raw-text values of 0.797, 0.799, 0.821, and 0.772. The raw-text detector reaches near-complete reversals in every domain, with minimum AUROCs between 0.010 and 0.026. The content-point detector's corresponding minima range from 0.911 to 0.954. The ranking gain therefore holds within these domains as well as in their average.

Strong ranking does not ensure a well-calibrated operating point in every domain. For content-point DETree, mean FPR is 0.029 in news and 0.036 in arXiv, versus 0.057 in creative writing and 0.071 in essays; the largest domain/setting FPR is 0.132 in creative writing. All use the same pooled validation threshold. Other detectors have larger threshold failures: under the human-content-only GPT-4o storyteller training intervention, content-point RACE and RoBERTa reach arXiv FPRs of 0.455 and 0.462 despite AUROCs of 0.941 and 0.956. These results motivate the separate calibration analysis in Appendix C.5.

C.2 Content representation ablations

The representation ablations clarify how retaining detail, restating it plainly, and extracting separate points contribute to the approach. A one-sentence summary tests the value of information beyond the central claim. A detailed plain rewrite tests how much voice robustness is retained without point extraction, while paragraph interventions reveal the additional structural benefit of content points. The summary experiment uses three detectors on unmodified HART; the plain-rewrite experiment uses RoBERTa with three seeds and test-time expression changes.

A central-claim summary loses useful detection information. We replace the content-point list with a summary requested as one self-contained sentence stating the document's central claim. RACE, RoBERTa, and DETree are each retrained on the summary input and compared with their content-point counterparts on the same corpus partitions. No additional expression intervention is applied. Figure 8 shows the resulting gap: AUROC rises from 0.825 to 0.970 for RoBERTa, from 0.822 to 0.966 for RACE, and from 0.751 to 0.963 for DETree when the input retains detailed content points. Recall rises substantially as well; for DETree, TPR changes from 0.216 to 0.886. The summary detector's lower FPR, 0.037 versus 0.048, accompanies its much lower recall. Similarly, the summaries' low expression leakage does not compensate for lost discrimination.

The gap varies by domain. A thesis sentence retains more discrimination in essays than in news or creative writing; RoBERTa's AUROC gap between content points and summaries is 0.031 in essays and 0.213 in creative writing. Detailed content therefore improves discrimination even before further expression changes are introduced. Coverage and length change together in this comparison; the next experiment retains detail to examine plain restatement and point structure.

Detailed plain restatement supports robustness to voice changes. We next retain the instruction to preserve every point, detail, and example, but ask for a plain rewrite of the whole document

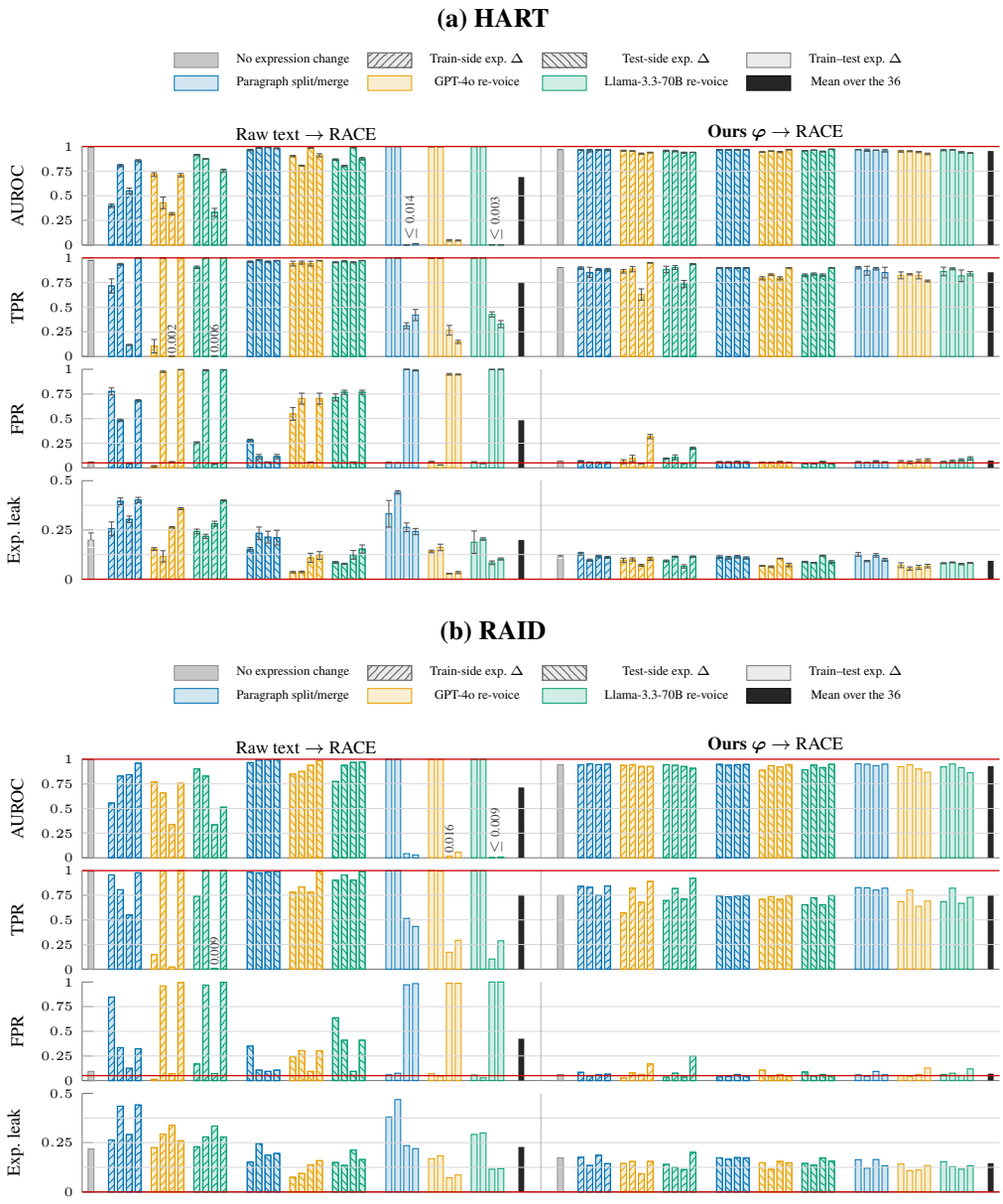

Figure 3: **RACE across HART and RAID.** Raw text and **Ours** ϕ use separately fitted detectors. Each half shows the reference, 12 training-only, 12 test-only, and 12 joint settings, and their 36-setting mean. Colors identify intervention families; patterns identify scope. Table 7 gives the assignment order. Metrics weight domains equally. HART bars average three seeds, with one sample standard deviation as whiskers except on the setting mean. TPR/FPR use fixed thresholds targeting 5% pooled validation FPR. Red lines mark AUROC/TPR 1, FPR 0.05, and leakage 0.

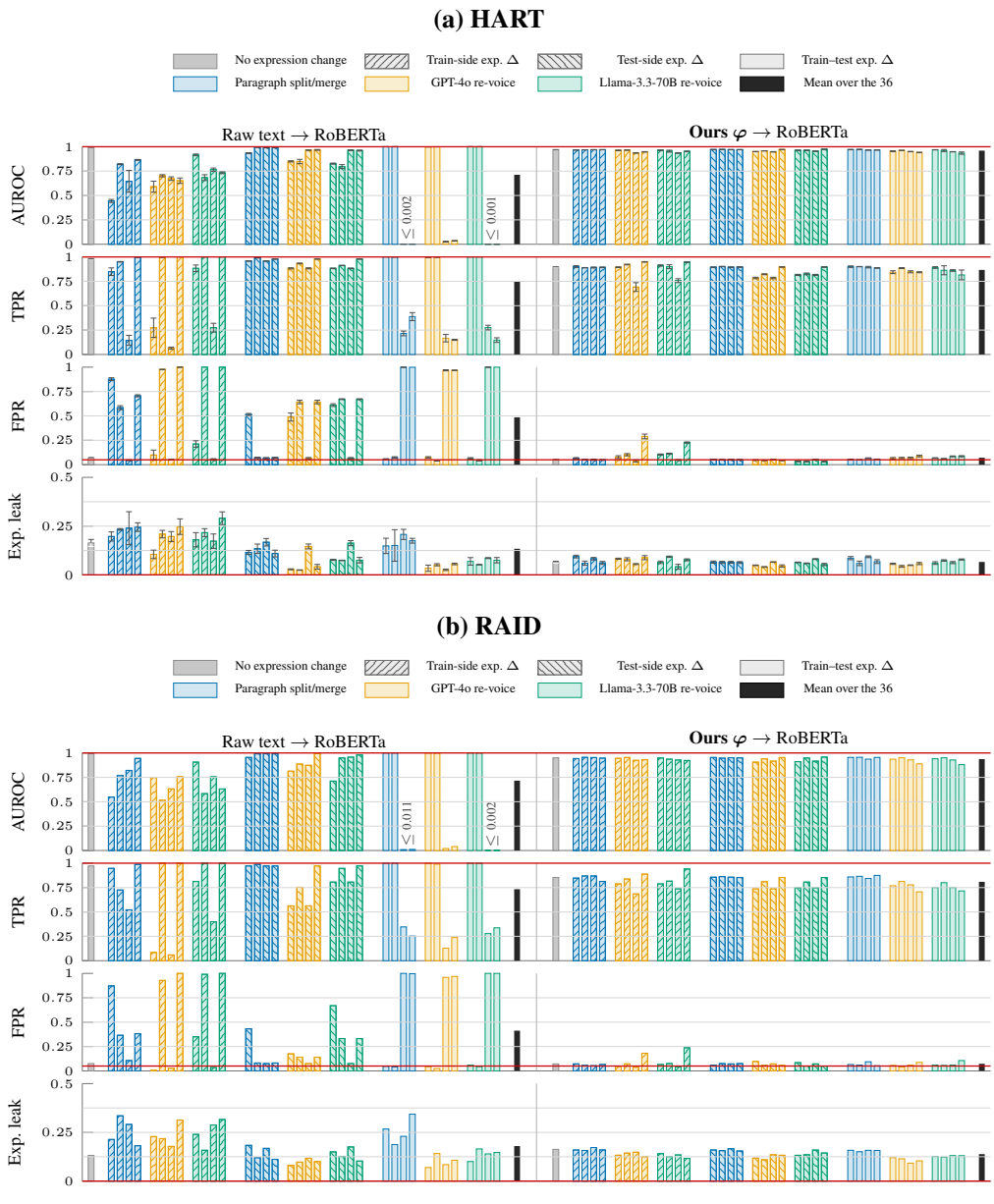

Figure 4: **RoBERTa across HART and RAID.** The corpora use separately trained models. Inputs, setting order, domain averaging, thresholds, and reference lines follow Figure 3. HART shows three-seed means and sample-standard-deviation whiskers, excluding the setting mean.

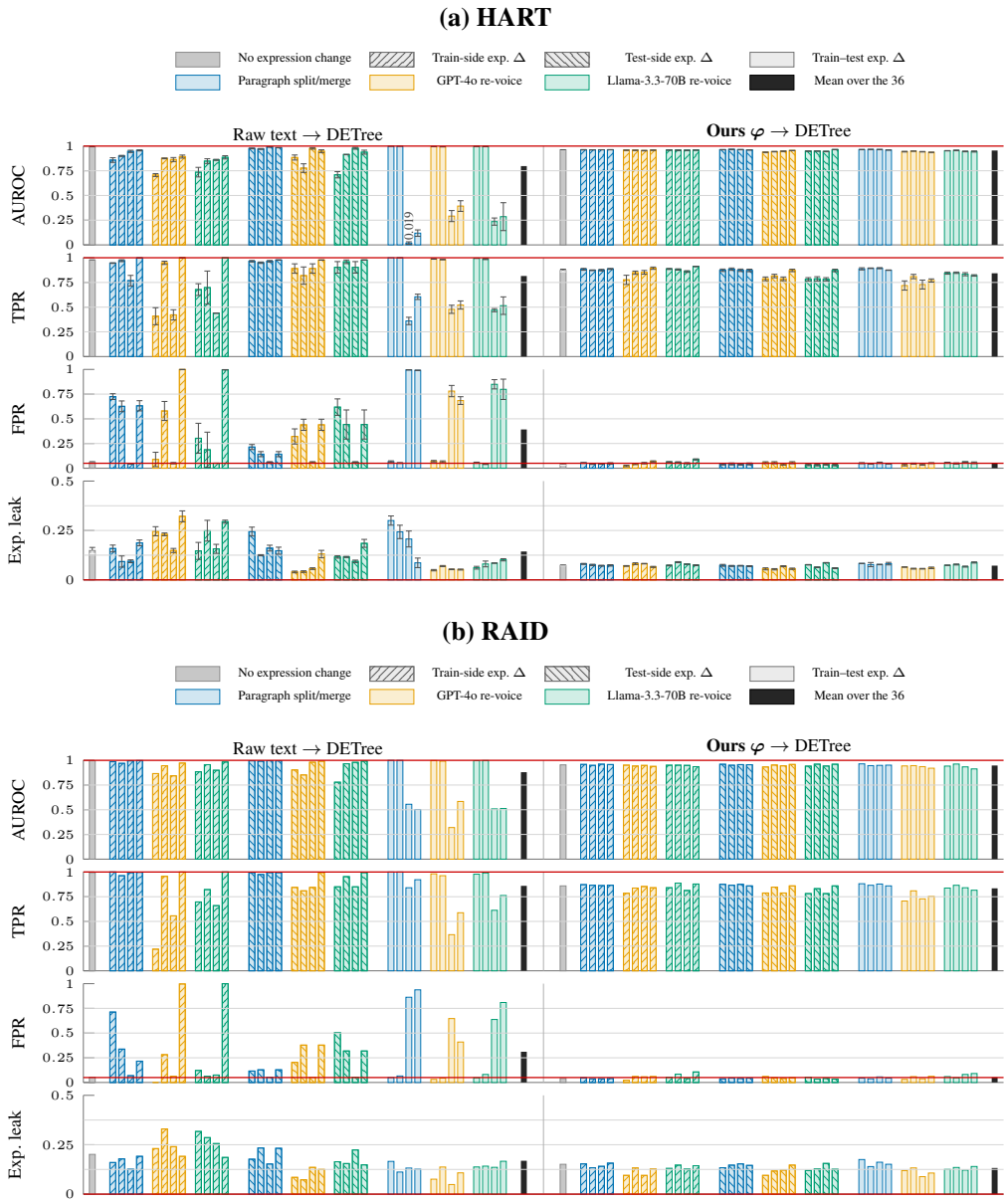

Figure 5: **DETree across HART and RAID.** Display conventions follow Figure 3, including three-seed means and sample-standard-deviation whiskers for HART. Raw-text DETree is less disrupted on RAID than on HART, while content points improve mean AUROC and reduce mean FPR on both corpora.

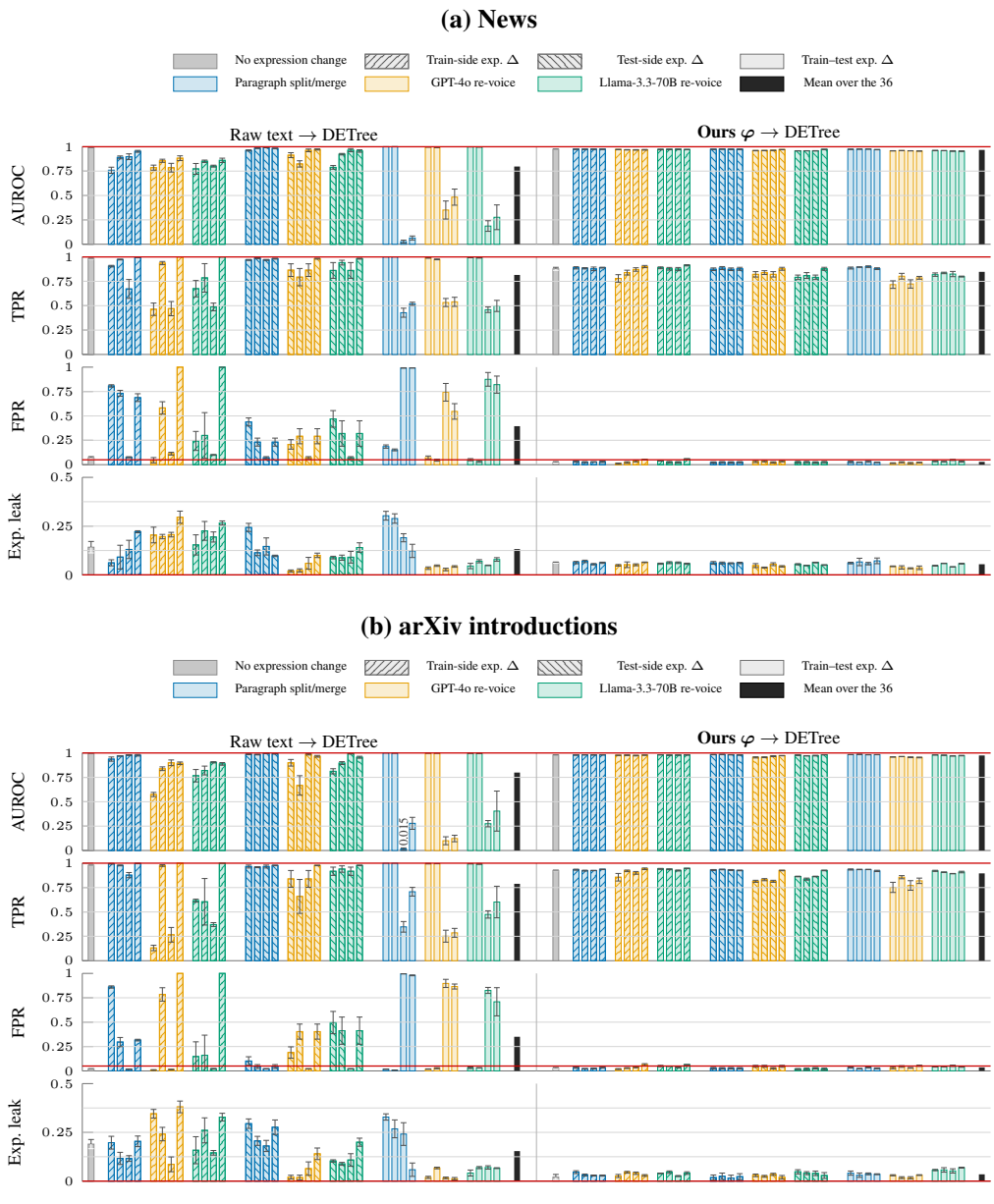

Figure 6: **DETree within HART's news and arXiv domains.** Bars show domain-specific means over three seeds; whiskers show one sample standard deviation except on the 36-setting mean. Both domains use the pooled validation threshold. Settings and display conventions follow Figure 3.

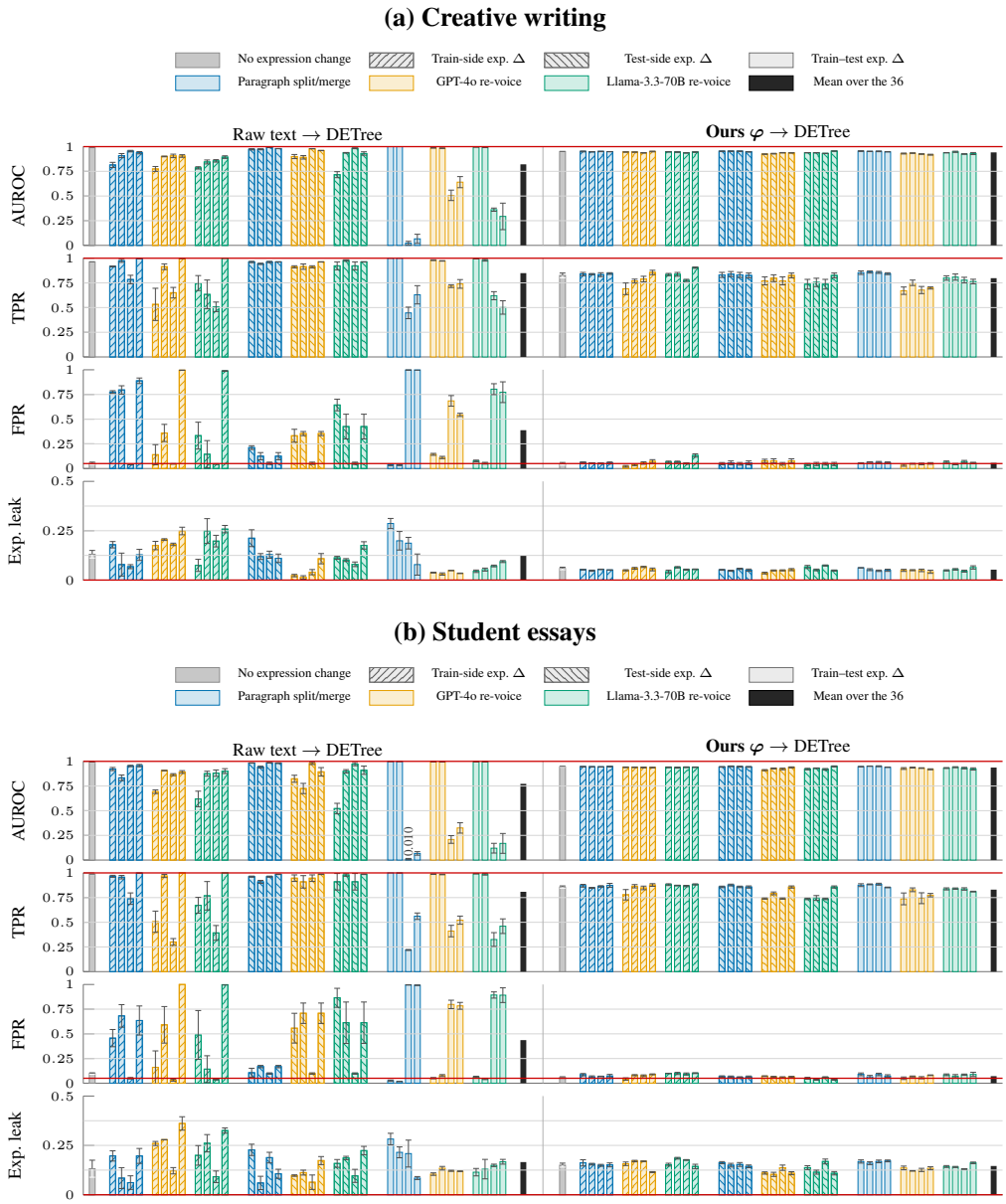

Figure 7: **DETree within HART's creative-writing and essay domains.** Three-seed means, sample-standard-deviation whiskers, and pooled validation thresholds follow Figure 6. Content points reduce expression sensitivity in both domains, while recall and FPR differ from news and arXiv.

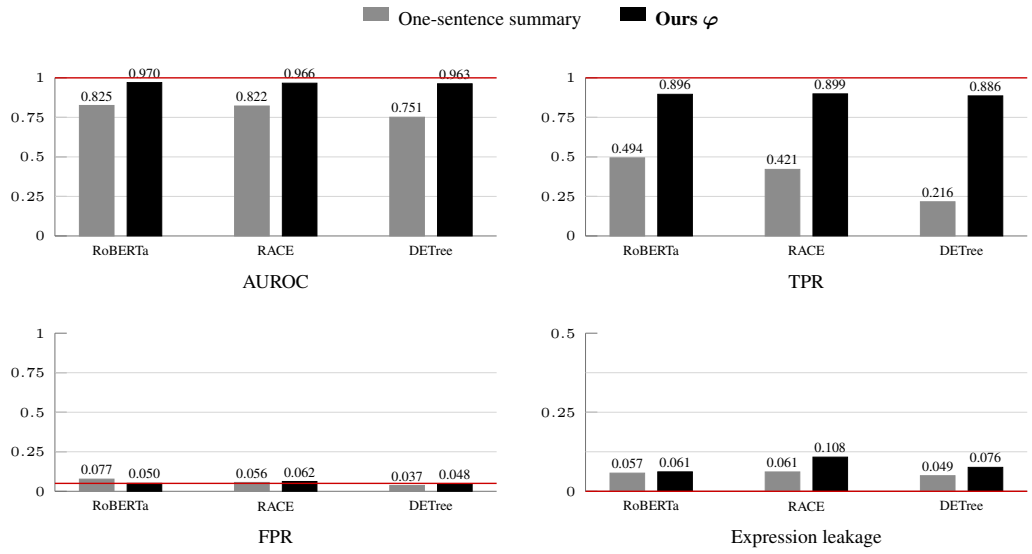

Figure 8: **Detailed points improve discrimination and recall on unmodified HART.** Detectors are retrained on each representation: gray bars use central-claim summaries; black bars use **Ours** ϕ. Metrics weight the four domains equally. TPR/FPR use each fit's threshold targeting 5% pooled validation FPR.

Statistic	Raw text	Plain rewrite	**Ours** ϕ
Unmodified test text			
AUROC	0.994	0.975	0.970
TPR	0.979	0.923	0.897
FPR	0.066	0.061	0.051
Twelve test-only expression settings			
Mean AUROC	0.924	0.965	0.964
Minimum AUROC	0.797	0.947	0.948
Mean TPR	0.935	0.890	0.852
Mean FPR	0.382	0.067	0.045
Maximum FPR	0.672	0.147	0.053
Mean expression leakage	0.096	0.053	0.060

Table 10: **Detailed plain rewriting recovers much of the test-time robustness.** RoBERTa on HART, separately fitted for each representation. Each setting averages four domains equally and then three seeds. The lower block summarizes 12 test-only settings, excluding the reference; extrema may occur in different settings. TPR/FPR use fixed validation-calibrated thresholds.

instead of a list of content points. For each of raw text, plain rewriting, and content points, we fit RoBERTa on its representation of unmodified HART training documents and calibrate a separate threshold on its corresponding validation inputs. The fitted detector and threshold are then fixed across the unmodified test condition and 12 test-only settings: four paragraph assignments and eight voice assignments, as defined in Appendix B.2. Evaluation retains the available test documents within each production condition, with the same domain/condition counts for all three inputs. Each representation uses three seeds.

Table 10 shows that detailed plain rewriting recovers much of the robustness of content points. Mean and minimum AUROC nearly match across the tested settings. Its mean TPR is higher, alongside a higher mean FPR, so the operating points should be read as pairs. Relative to each input's unmodified performance—AUROC 0.9749 for plain rewriting and 0.9696 for content points—the AUROC changes differ by at most 0.006 across all eight voice settings. Plain rewriting also keeps voice-setting FPR between 0.053 and 0.062, close to its unmodified value of 0.061. This comparison supports detailed plain restatement as a major contributor to the observed voice robustness.

Content-point extraction adds robustness to paragraph structure. When human-content documents are split into short paragraphs and LLM-content documents are merged into one paragraph, raw-text RoBERTa reaches TPR/FPR 0.959/0.515. Plain rewriting changes this to 0.906/0.147, while content points yield 0.895/0.053. Relative to each representation's unmodified reference, AUROC changes by -0.059 for raw text, -0.028 for plain rewriting, and $+0.001$ for content points; FPR increases by 0.449, 0.086, and 0.002, respectively. Content-point extraction therefore preserves both ranking and the operating point under this paragraph assignment, substantially reducing the false-positive increase that remains after plain rewriting.

These ablations support the complementary design of ϕ: retaining substantive detail supports discrimination, plain restatement supplies much of the observed voice robustness, and point extraction adds robustness to paragraph structure. The representation thus reduces expression sensitivity while retaining detailed distinctions useful for detection. The same points supply the units for CRO's analysis of recurring content choices, motivating a shared representation for detection and interpretation.

Summary generation and prompts. Summaries are generated independently for training, validation, and test documents using GPT-4o (`gpt-4o-2024-11-20`), temperature 0.2, top-p 1, and a maximum of 256 output tokens. The cleaned reply is retained as a single input unit, without truncating it to its first sentence; test summaries average 28.3 words. In the messages below, {`text`} is replaced by the document. The content-point extraction prompt is in Appendix B.4.

Summary system message.

```
You are a careful analyst. You extract the substantive content of a text
↪ and restate it plainly, ignoring the author's writing style.
```

Summary task message.

```
State the central claim (the thesis) of the text below. Follow these
↪ rules:
- Write it as one self-contained sentence in plain, neutral language.
- Capture what the text conveys (what it claims, argues, or proposes),
↪ not how it is expressed. Do not reuse the author's distinctive
↪ phrasing, tone, imagery, or stylistic quirks.
- Choose the one point that matters most; do not list several points and
↪ do not join separate points into a compound sentence.
- Output only that one sentence, with no preamble or commentary.

TEXT:
{text}
```

Plain-rewrite generation and prompt. This procedure uses GPT-4o (`gpt-4o-2024-11-20`), temperature 0.2, top-p 1, and a maximum of 2,500 output tokens. It applies one rewrite pass to the available text, after any test-time expression intervention, without a target word-count band or a request to decompose the text. There is no system message; the exact task message is below. Output-announcing first lines are removed by a fixed prefix rule. Five persistent rewrite refusals are excluded from the 7,008 training documents, leaving 7,003; this rewrite step excludes no validation or test document.

Plain-rewrite task message.

```
Rewrite the text below in plain, neutral language, as a straightforward
↪ statement of what it says. Follow these rules:
- Keep every point, detail, and example; do not add, remove, or summarize
↪  anything.
- Do not reuse the original's distinctive phrasing, sentence rhythm,
↪ imagery, or stylistic quirks; use ordinary, unremarkable wording
↪ throughout.
- Output only the rewritten text, with no preamble.

TEXT:
{text}
```

	Raw text		**Ours** ϕ	
Statistic	Four-way	Binary	Four-way	Binary
Unmodified test text				
AUROC	0.992	0.992	0.966	0.970
Expression leakage	0.202	0.077	0.108	0.087
All 36 expression-change settings				
Mean AUROC	0.691	0.695	0.953	0.959
Mean TPR	0.758	0.751	0.850	0.875
Mean FPR	0.477	0.507	0.074	0.074
Mean expression leakage	0.201	0.082	0.090	0.075
Minimum AUROC	0.0012	0.0001	0.9229	0.9304
Maximum FPR	1.000	1.000	0.294	0.317
Settings with AUROC < 0.5	10	9	0	0
Joint training/test changes: AUROC ranges				
Matched assignments (6)	0.9978–0.9999	0.9984–1.0000	0.9416–0.9683	0.9543–0.9704
Reversed assignments (6)	0.0012–0.0559	0.0001–0.0288	0.9229–0.9629	0.9354–0.9666

Table 11: **Binary supervision leaves RACE's severe expression sensitivity intact.** Identical-seed comparison; both targets are evaluated for binary content-source detection, with equal-domain averaging. The middle block summarizes the 36 changed settings; the final block separates six matched from six reversed assignments. TPR/FPR use fixed validation-calibrated thresholds. AUROC minima and ranges retain four decimals to distinguish near-complete reversals.

C.3 Changing the training objective

The training comparisons establish why the input representation matters even when the objective explicitly targets content source or suppresses expression information. Binary supervision leaves severe raw-text ranking reversals intact, showing that they are not explained solely by the four-way training target. The tested adversarial objectives also leave severe reversals, even under configuration selection that favors those alternatives. Content-point inputs retain strong discrimination under these same reversals.

Binary content-source supervision. Binary RACE replaces the four-production-condition head with a two-class content-source head, retaining the encoder, rhetorical graph, and nominal training recipe in Appendix B.4. Cross-entropy and supervised contrastive learning both use the binary labels. Within each input representation, the targets use the same split documents and graph inputs; content-point graphs are built from the extracted text.

The evaluation is binary for *both* training targets. Four-way RACE is scored by $P(t_2) + P(t_3)$, whereas binary RACE uses the logit difference $z_1 - z_0$, an increasing function of its LLM-content probability. The saved score is used for both ranking and threshold calibration. Checkpoints maximize validation macro-F1 over the respective four or two classes. Both targets follow the shared validation calibration and checkpoint-reuse rules in Appendix B.

We evaluate both representations and training targets on HART's unmodified reference and all 36 expression-change settings. Table 11 compares the resulting discrimination, operating points, and reversal failures.

The failures persist when the target names only content source. On raw text, binary supervision changes mean AUROC from 0.691 to 0.695, with nine of the 36 settings still below chance, compared with ten under four-way supervision. The joint evaluations expose the failure more directly: binary RACE achieves AUROC 0.998–1.000 across the six matched assignments, but only 0.0001–0.029 when those assignments reverse at test time. This includes paragraph changes that preserve the words and their order. Removing production-condition distinctions from the target therefore does not prevent the fitted detector from relying on expression correlated with content source.

The benefit of content-point input persists under binary supervision. Mean AUROC rises from 0.695 to 0.959, and mean TPR/FPR changes from 0.751/0.507 to 0.875/0.074. No setting-level domain average falls below chance; the minimum AUROC is 0.930. Across individual domains, mean AUROC over the 36 settings ranges from 0.687–0.702 on raw text and 0.947–0.975 on content

Target	Input and training	Mean AUROC	Mean TPR	Mean FPR
Binary	Raw text, no adversary	0.675	0.725	0.513
Binary	Raw text, selected adversary	0.827	0.774	0.280
Binary	**Ours** ϕ, no adversary	0.959	0.873	0.075
Four-way	Raw text, no adversary	0.712	0.741	0.470
Four-way	Raw text, selected adversary	0.809	0.769	0.255
Four-way	**Ours** ϕ, no adversary	0.959	0.864	0.071

Table 12: **Adversarial training improves average performance but leaves a substantial gap.** RoBERTa on HART, identical-seed comparison; metrics average domains equally and then the 36 changed settings. The selected adversary maximizes test AUROC within each training corpus. Content-point rows use no adversary and retain the corresponding target. TPR/FPR use validation-calibrated thresholds.

points. The same calibration weakness remains under binary supervision: rewriting only human-content training and validation documents in the GPT-4o storyteller voice yields content-point FPR 0.317 on unmodified test text.

Direct interventions expose shortcuts that leakage alone can miss. Binary supervision reduces raw-text expression leakage on the unmodified reference from 0.202 to 0.077, yet severe intervention failures remain. Leakage measures score separation *within* each content-source class. A cue shared by both human-content conditions, and a different cue shared by both LLM-content conditions, can predict source with little within-class separation. The unmodified leakage and joint-intervention failures also concern differently trained checkpoints; the direct interventions establish the vulnerability. A content-source label specifies the required answer, but does not determine which evidence the detector learns to use.

Explicit adversarial expression training. We next test whether a detector can learn to suppress expression information while retaining content-source discrimination. The primary comparison uses binary RoBERTa on HART, with adversarial heads attached to its pooled text representation. We consider three kinds of signal: production-condition distinctions within each source class (t_0 versus t_1, and t_2 versus t_3); the procedure used to render a training document; and membership in the labeled training population versus an unlabeled population rendered like the intended test condition. The last approach includes domain-adversarial neural networks (DANN; Ganin et al., 2016) and conditional domain-adversarial networks (CDAN; Long et al., 2018), whose discriminator additionally uses the classifier's predictions. These target populations use validation documents with their labels discarded and remain disjoint from test documents. They provide information about the deployment expression condition that the ordinary supervised detector does not receive.

A comparison selected to favor the adversarial alternatives. For each of the 13 training corpora, we choose the adversarial configuration with the highest mean *test* AUROC across all evaluation settings associated with that corpus, among configurations with complete results for those settings. For unmodified training, selection includes the unmodified test reference and 12 test-only changes; other training corpora use their associated unmodified, matched, or reversed evaluations. The configuration cannot be selected separately for matched and reversed assignments. DANN/CDAN nevertheless have condition-specific fits because their unlabeled target pools depend on the intended test condition. This is a retrospective, test-selected comparison, not a deployable selection protocol or an upper bound on other adversarial methods. Checkpoint selection within each candidate follows its specified validation-F1 or final-epoch rule; the decision threshold is calibrated on that fit's pooled human-content validation scores, not optimized on test labels.

Content points withstand reversals that defeat the selected adversaries. Content points retain high discrimination across the reversed joint assignments that defeat the selected adversaries (Figure 9). Across those six settings, selected-adversary AUROC ranges from approximately 0.0001 to 0.620, compared with 0.939–0.966 for binary RoBERTa on content points. Ranking and threshold behavior can also diverge: after rewriting only human-content training and validation documents in the Llama storyteller voice, the selected adversary reaches AUROC 0.945 on unmodified test text, but TPR is 1.000 and FPR is 0.99975. Its threshold labels almost every document as LLM content.

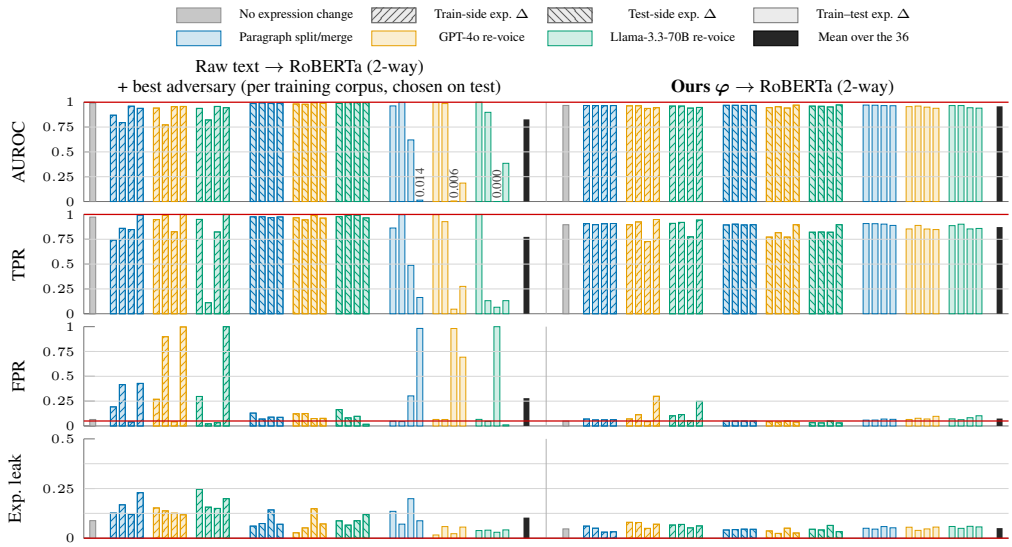

Figure 9: **Binary RoBERTa: adversarial configurations selected using test performance.** Left: raw text with the selected adversary for each training corpus. Right: **Ours** ϕ without an adversary. Both use a binary content-source target. Metrics weight HART's four domains equally; TPR/FPR use fixed validation-calibrated thresholds. Setting order, colors, patterns, and reference lines follow Figure 3; near-zero annotations are rounded.

Content points achieve overall mean TPR/FPR of 0.873/0.075, with setting variation visible in the figure.

The selected binary adversaries raise mean AUROC from 0.675 to 0.827 and improve mean TPR/FPR from 0.725/0.513 to 0.774/0.280 (Table 12). Their strongest average result is on the 12 test-only settings: AUROC reaches 0.990, compared with 0.946 without an adversary, while FPR falls from 0.461 to 0.095. This exceeds the content-point detector's test-only AUROC of 0.961; the content-point detector has the lower FPR, 0.042. Adversarial adaptation therefore provides useful protection in this part of the evaluation.

The four-way replication gives the same qualitative conclusion under the paper's main RoBERTa target. Adversarial selection raises mean AUROC from 0.712 to 0.809 over all 36 settings, but AUROC across the six reversed joint assignments remains 0.009–0.586. Four-way RoBERTa on content points reaches mean AUROC 0.959 over the full 36-setting set. The improvements from adversarial training are substantial, but they do not remove the severe failures under the tested reversals.

Adversarial search details. The binary search contains 49 adversarial configurations, including combined objectives. It varies adversarial weights from 0.1 to 10, gradient reversal versus a loss pushing adversary predictions toward a uniform distribution, weight schedules, validation-macro-F1 versus final-epoch checkpoint selection, and last-layer versus full-encoder fine-tuning. A four-way replication evaluates nine configurations informed by the binary search, excluding the within-source production-condition adversary because those distinctions are part of the four-way target. Thus the two targets do not receive equally broad searches. The no-adversary raw-text controls are fitted with the same experimental trainer. The binary content-point control is also fitted there; the four-way content-point reference is the corresponding main-experiment fit with the identical seed. Shared implementation settings are given in Appendix B.4.

The strength of the expression–source association matters. A separate binary-RoBERTa diagnostic varies the association during training and validation while keeping the test assignment reversed. For each matched group and content-source class, the two production conditions share a rendering assignment: human content uses renderer A and LLM content renderer B with probability $(1+\rho)/2$, and the opposite assignment otherwise. Thus $\rho = 0$ gives an equal mixture in expectation, while $\rho = 1$ makes rendering perfectly identify content source. Assignments use the same underlying draw across ρ values, and the reversed test always assigns B to human content and A to LLM

	Training association strength ρ					
Input and training	0.0	0.2	0.4	0.6	0.8	1.0
Raw text, no adversary	0.969	0.956	0.923	0.880	0.736	0.009
Raw text, renderer adversary ($\lambda = 1$)	0.967	0.971	0.966	0.963	0.913	0.009
Ours ϕ, no adversary	0.963	0.962	0.960	0.954	0.956	0.951

Table 13: **A renderer adversary helps under partial association but fails at perfect association.** Reversed-test AUROC for binary RoBERTa, averaged over four domains and three intervention families. The adversary is a fixed configuration, not the test-selected search winner. Both control rows use no adversary. These training/validation mixtures are separate from the 36-setting comparison.

content. We repeat this construction for paragraph changes and both voice families; content-point inputs are extracted after rendering. The renderer adversary uses gradient reversal with a scheduled maximum weight $\lambda = 1$ and validation-macro-F1 checkpoint selection.

At $\rho = 0.8$, the renderer adversary raises reversed-test AUROC from 0.736 to 0.913 (Table 13). At $\rho = 1$, it reaches only 0.009, while the content-point detector retains AUROC 0.951. This pattern supports a specific difficulty with source-correlated expression. When rendering identifies the source label, suppressing renderer information competes with predicting that label on the training distribution. An adversary separating production conditions within each source class does not directly target a renderer shared by both conditions. For domain alignment, reversing the source–renderer association can preserve the overall renderer mixture, so matching that mixture alone does not determine which association will hold at test time. These observations explain limitations of the tested procedures; they do not establish an impossibility for adversarial learning.

Together, these comparisons establish a substantial advantage for changing the input representation. A content-source target alone leaves expression shortcuts available, and the tested adversaries retain severe reversal failures even when selected using test performance. Content-point extraction applies the same label-blind instructions for detailed content and plain restatement to both populations before fitting. Its strong discrimination under the same reversals persists across binary and four-way supervision. Adversarial adaptation remains useful for other shifts, particularly when given unlabeled examples of the intended expression condition; the reversal comparisons show where normalizing the input provides the stronger protection.

C.4 Expression dependence in HART's likelihood-based detection

HART (Bao et al., 2025) infers content source from likelihood-based measurements of a document and a model-generated simplification. Its two variants use FastDetectGPT (Bao et al., 2023) or Binoculars (Hans et al., 2024) as the criterion. For each input, the evaluated implementation computes the criterion once on the input itself and once on its GPT-4o simplification. An RBF support-vector regressor receives these two values as separate features and is fitted to binary content-source labels. The regressor can use their relationship; its input is the pair of values, not their literal difference.

Inferring content source through the relationship between a document's expression and its rewrite leaves the detection signal sensitive to expression changes. The criterion pair can capture how the original wording relates to the simplified wording, even though the prediction target is content source. Our extraction instead aims to retain substantive details while reducing expression cues. On content-point input, HART measures the restated points and a new simplification of them. It still receives only two likelihood-based values: preserving useful content in the input does not ensure that these features capture it effectively.

We evaluate both variants on HART, fitting separate raw-text and content-point regressors for each training condition on the training partition and evaluating on the test partition. Appendix B.4 specifies the criterion models, fixed regressor settings, and newly generated simplifications. Both variants use the shared held-out validation calibration (Appendix B.3); no test documents enter fitting or threshold selection. Figure 10 retains the unmodified reference and all 36 expression-change settings. Metrics weight the four domains equally.

Expression changes weaken the raw-text signal. Raw-text HART achieves unmodified AUROC 0.857 with FastDetectGPT and 0.861 with Binoculars. Across the 36 expression-change settings,

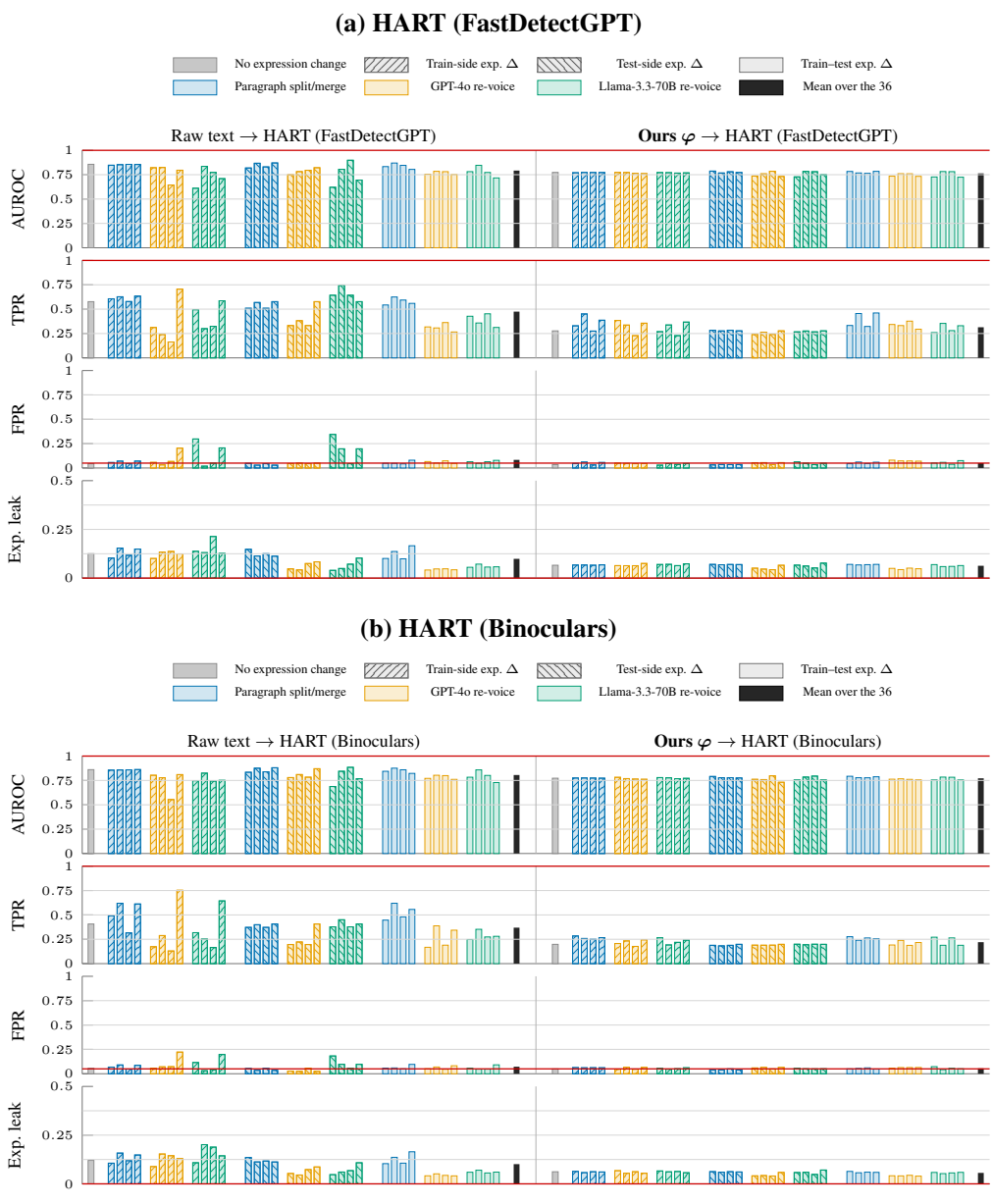

Figure 10: **Content points reduce HART's expression sensitivity and improve its weakest settings.** Average discrimination nevertheless falls, as detailed in the text. Each variant uses separately fitted raw-text and content-point regressors. Metrics weight HART's four domains equally; TPR/FPR use fixed validation-calibrated thresholds. Setting order, colors, patterns, and reference lines follow Figure 3.

AUROC ranges from 0.611 to 0.898 and from 0.554 to 0.884, respectively. Thus the criterion-pair design does not sustain its reference discrimination under changes in expression. These are setting-level averages over domains; neither variant exhibits the below-chance reversals seen in the other raw-text detectors.

Extraction reduces expression dependence; the likelihood features retain limited discrimination. Content points narrow the AUROC ranges to 0.724–0.786 for FastDetectGPT and 0.731–0.798 for Binoculars, improving both variants' weakest settings. Mean expression leakage also falls, from 0.099 to 0.063 and from 0.101 to 0.056, respectively. However, mean AUROC falls from 0.792 to 0.764 and from 0.806 to 0.772; both variants reach only 0.773 on unmodified content points. Extraction therefore reduces expression sensitivity without recovering the strong discrimination that RoBERTa, RACE, and DETree achieve on the same representation (Appendix C.1).

Lower false-positive rates accompany lower recall. At the fixed validation-calibrated thresholds, mean TPR/FPR changes from 0.475/0.083 to 0.314/0.050 for FastDetectGPT, and from 0.369/0.072 to 0.220/0.055 for Binoculars. The lower FPR on content points should therefore be read alongside the loss of recall. This loss is also present without an expression intervention: reference TPR falls from 0.577 to 0.277 and from 0.407 to 0.198, respectively.

HART's domain-dependent score scales account for part of the operating-point limitation. Calibrating a separate threshold from each domain's human-content validation scores can materially improve recall, especially for Binoculars. On its unmodified content-point input, this changes TPR from 0.198 to 0.388 while FPR changes from 0.049 to 0.050. The primary figure retains the pooled convention shared with the other detectors. Appendix C.5 develops the threshold comparison, distinguishing this in-domain calibration from calibration on an unseen domain. Threshold choice leaves AUROC and expression leakage unchanged, so improved calibration cannot remove the ranking loss above.

The reduction in expression sensitivity alongside lower average discrimination is consistent with weakening an expression-dependent signal; changing both criterion inputs does not isolate a unique cause of the loss. The comparison nevertheless makes a practical distinction clear: content extraction alone does not make these likelihood features as effective as detectors fitted to the detailed point text. Together with the representation and commitment ablations (Appendices C.2 and C.7), these results support preserving substantive detail and using detectors that can learn from it. This combination substantially reduces expression dependence while retaining strong content-source discrimination.

C.5 Domain transfer and calibration

Content-point detection retains its robustness advantage when the target domain is excluded from detector training. The learned source distinction remains useful under expression changes, while domain-dependent score distributions can require a different operating threshold even on unmodified documents. We first establish the transferred discrimination, then show how a small sample of known human-content target documents recovers an operating point close to the intended false-positive rate without retraining.

We train RoBERTa on three HART domains and evaluate on the fourth, repeating this leave-one-domain-out procedure for arXiv, essays, news, and creative writing. Raw text and content points use separately fitted detectors with the standard four-condition training target and binary content-source evaluation (Appendix B.4). The held-out domain supplies no documents for fitting or checkpoint selection. Initial calibration uses only human-content validation documents from the three training domains (Table 8). Evaluation uses the unmodified reference and all 36 settings, averaging the four held-out domains equally under the shared reporting conventions.

Expression robustness persists after withholding the domain. With content points, domain-macro AUROC is 0.892 on unmodified text and 0.875 averaged over the 36 changes. Raw-text AUROC falls from 0.930 to 0.657, with nine setting-level domain averages below chance. None of the content-point domain averages inverts; their minimum AUROC is 0.807. This minimum describes the four-domain average, not a bound for each domain. At the individual-domain level, content-point mean AUROC is 0.926 for arXiv, 0.875 for essays, 0.863 for news, and 0.837 for

creative writing, compared with 0.684, 0.659, 0.679, and 0.606 on raw text. The individual-domain plots in Figures 11 and 12 show the retained ranking and the remaining variation across settings.

Threshold mismatch is already present without expression interventions. On unmodified target-domain inputs, content-point AUROC is 0.941 in arXiv and 0.846 in creative writing, yet their FPRs are 0.236 and 0.008. The transferred threshold is therefore too permissive in one domain and too conservative in the other despite useful discrimination in both. Across expression changes, mean FPR is 0.320 in arXiv and 0.166 in essays, versus 0.056 in news and 0.012 in creative writing. Creative-writing mean TPR is only 0.334 despite AUROC 0.837. These results motivate calibrating the operating point to the target domain's score distribution, separately from assessing expression robustness relative to that domain's unmodified reference.

Calibrating with known human content from the target domain. We recalibrate each fixed detector using 100 human-content documents from 50 target-domain groups, averaging results over 400 calibration draws. Human-content documents from sampled groups are excluded from FPR evaluation. Calibration uses no LLM-content labels or scores, and leaves AUROC unchanged. The sampling and evaluation procedure is specified after the results.

Target-domain calibration recovers an appropriate operating point. Across all four domains and the 36 expression changes, target calibration brings content-point mean FPR from 0.138 to 0.052, near the intended 0.05; domain-specific means range from 0.048 to 0.058. These are averages over settings and calibration draws. AUROC remains 0.875: the adjustment uses the retained source discrimination without retraining the detector. Overall mean TPR changes from 0.639 to 0.546 as calibration corrects overly permissive thresholds in some domains.

Creative writing illustrates the complementary benefit of correcting an overly conservative threshold (Figure 12b). Content-point AUROC is 0.846 without interventions and 0.837 averaged across changes. Target calibration raises mean TPR from 0.334 to 0.578 while changing FPR from 0.012 to 0.058. A small human-content sample thus recovers recall at an operating point close to the intended FPR, while preserving the transferred detector and its expression robustness.

Raw text receives the same calibration procedure. Its creative-writing mean TPR/FPR changes from 0.682/0.598 to 0.306/0.154, while AUROC remains 0.606. The threshold can reduce false positives, but it cannot repair expression-induced ranking reversals.

Across domains, raw text reaches mean TPR/FPR 0.381/0.157 after the same target calibration, compared with 0.546/0.052 for content points. The content-point representation therefore supports a substantially better operating point after local calibration as well as stronger ranking across the expression changes.

Target-calibration procedure. For each calibration draw, we sample 50 whole groups without replacement from the target-domain evaluation pool. Their 100 human-content documents are scored by the checkpoint used for the setting being evaluated, using the corresponding extracted inputs for the content-point detector. Detector parameters remain fixed, and no LLM-content labels or scores are used to choose the threshold.

Each draw sets the threshold to the empirical 95th percentile of those 100 scores, using linear interpolation; scores strictly above it are flagged. FPR is measured on human-content documents outside the sampled groups, while TPR uses all LLM-content documents, including those from sampled groups. We average these rates over 400 draws, sharing group selections across settings and input representations within each domain. AUROC and expression leakage retain their full setting-level evaluation populations, so their values are unchanged. This target-domain calibration is evaluated separately from the shared training-domain validation rule.

Domain-specific thresholds when all domains are already represented. The HART comparison in Appendix C.4 raises a related but distinct issue. When fitting includes all four domains, we can replace the pooled threshold based on 496 human-content validation documents with four thresholds, each based on its domain's 124 human-content validation documents. Both choices use the checkpoint's own validation expression condition; neither draws calibration documents from the test partition. Table 14 compares their operating points using the same fitted HART detectors and test scores.

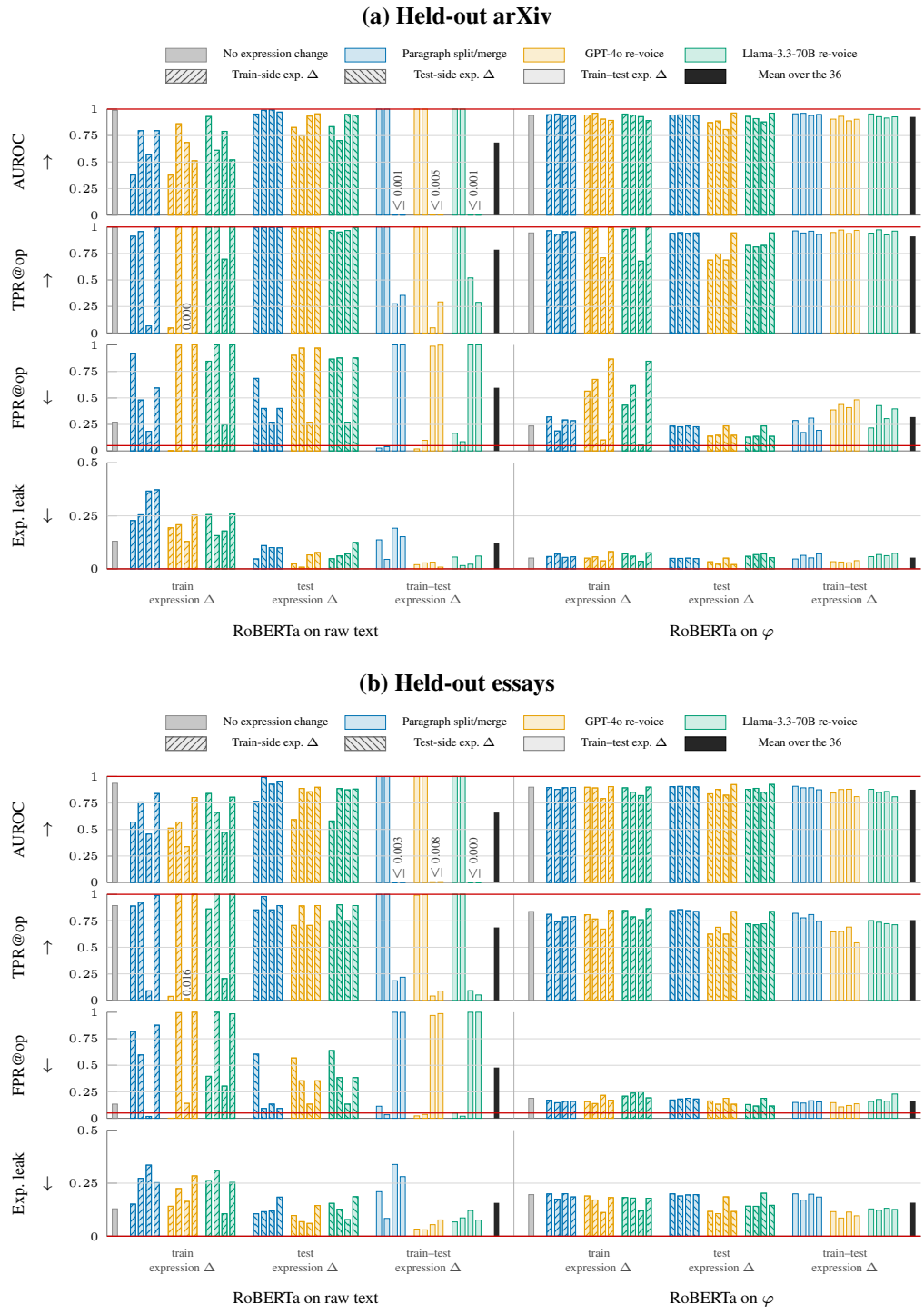

Figure 11: **Content-point ranking remains robust in held-out arXiv and essays.** RoBERTa is trained on the other three domains. Left: Raw text; right: **Ours** ϕ. TPR/FPR use training-domain validation thresholds. Each panel retains the reference, 36 expression-change settings, their mean, and all four metrics; near-zero annotations are rounded.

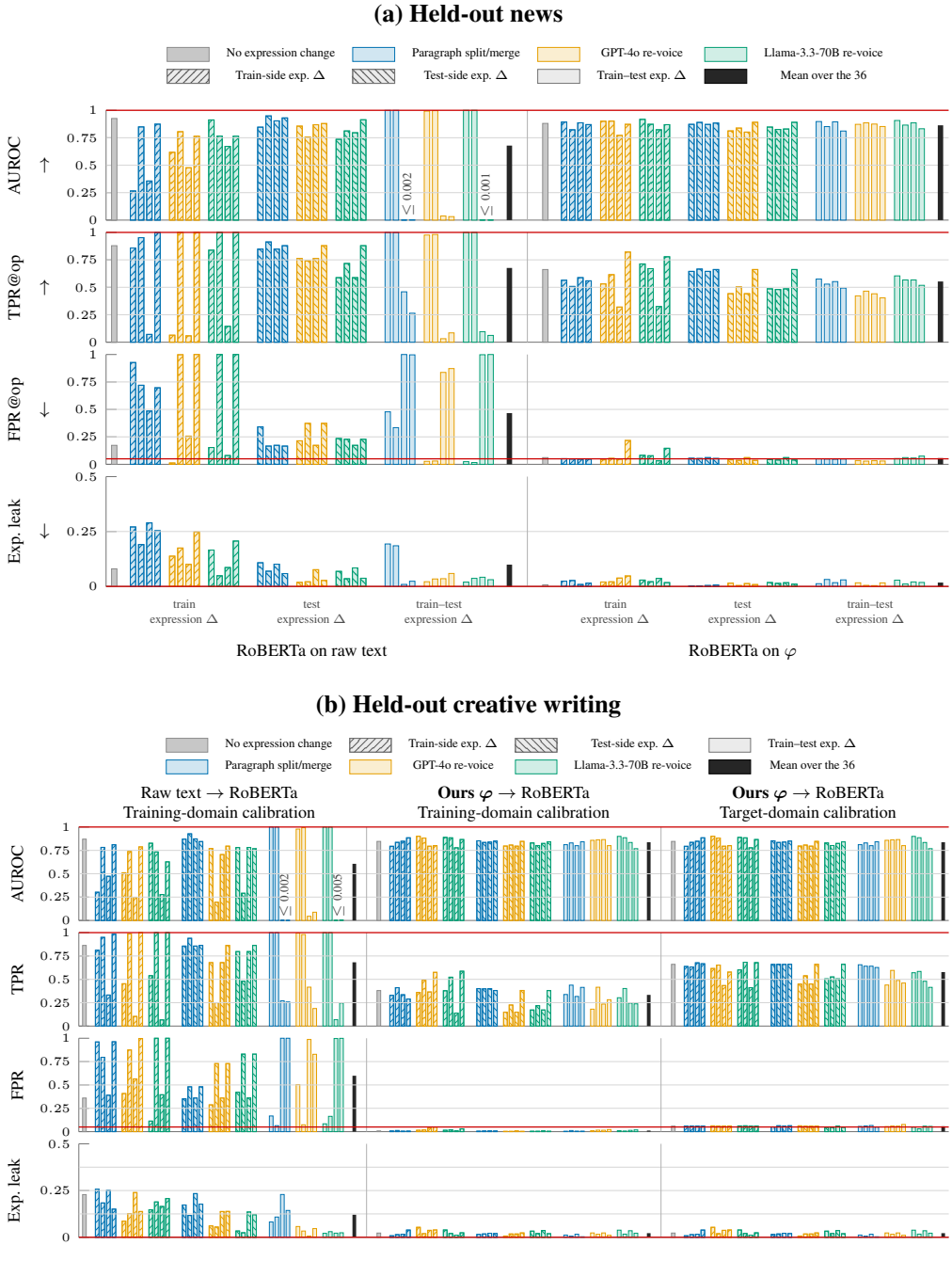

Figure 12: **Target-domain calibration recovers recall in creative writing without changing ranking.** RoBERTa is trained on the other three domains. Top: news with training-domain calibration. Bottom: creative writing with raw and content-point inputs under training-domain calibration, followed by the same ϕ detector with target-domain calibration. Target thresholds use 100 human-content documents from 50 target-domain groups; TPR/FPR average 400 calibration draws. Human-content documents from sampled groups are excluded from FPR evaluation. AUROC and leakage are identical across the two ϕ columns.

		Pooled threshold		Per-domain thresholds	
Detector	Input	TPR	FPR	TPR	FPR
HART (FastDetectGPT)	Raw text	0.475	0.083	0.520	0.118
HART (FastDetectGPT)	**Ours** ϕ	0.314	0.050	0.344	0.061
HART (Binoculars)	Raw text	0.369	0.072	0.488	0.090
HART (Binoculars)	**Ours** ϕ	0.220	0.055	0.371	0.062

Table 14: **In-domain calibration changes HART's recall and false-positive rates together.** Metrics average four domains equally and then the 36 changed settings. Both rules target 5% human-content validation FPR; fitted detectors and test scores are unchanged. All domains are represented during fitting, unlike the held-out-domain experiment.

Input	Training condition	Source AUROC	Rewrite AUROC
Raw text	Unmodified	0.907	0.615
Raw text	Source-associated voice	0.541	0.988
Ours ϕ	Unmodified	0.957	0.511
Ours ϕ	Source-associated voice	0.955	0.522

Table 15: **A source score can distinguish rewrites while losing source discrimination.** RoBERTa, one run (seed 3407), pooling available HART test documents across domains. Source AUROC compares content sources under a common casual rewrite; rewrite AUROC compares two rewrites of original-human documents. Both use the same fitted source score. Rewrite AUROC near 0.5 indicates little directional separation, not identical individual scores.

Per-domain thresholds raise recall for both HART variants, especially Binoculars, but also raise mean FPR across the expression changes. For example, FastDetectGPT on raw text changes from TPR/FPR 0.475/0.083 to 0.520/0.118. The favorable unmodified Binoculars example in Appendix C.4 therefore does not imply that domain-specific calibration improves every operating point. The primary HART results retain the pooled convention, with this comparison showing how domain-specific score scales affect its interpretation.

The held-out-domain experiments establish two complementary benefits: content-point detectors retain useful source discrimination under expression changes in unseen domains, and calibration with known human-content examples adapts their operating point without retraining. Threshold mismatches are already visible on unmodified target inputs; their presence does not erase the transferred ranking advantage. Conversely, calibrating raw-text scores cannot repair the ranking reversals induced by expression changes.

C.6 Re-voicing and learned source-score dependence

Re-voicing can make a detector's source score strongly distinguish rewrite conditions while weakening its discrimination between content sources. We examine this dependence directly, test whether paragraph formatting accounts for the loss, and assess whether consistent whitespace normalization during training prevents it.

Separate source discrimination from rewrite discrimination. We evaluate each fitted RoBERTa's existing LLM-content score, $P(t_2) + P(t_3)$, on two contrasts. For *source discrimination*, all human- and LLM-content test documents receive the same GPT-4o casual rewrite; LLM content is the positive class. For *rewrite discrimination*, we compare casual and main-storyteller versions of original-human (t_0) documents, with storyteller positive. No rewrite classifier is fitted. We compare unmodified training with source-associated training: human-content documents receive the casual rewrite and LLM-content documents the storyteller rewrite, in both training and validation (assignment I). Each training condition and input representation uses a separate fit.

After source-associated training, raw RoBERTa's rewrite AUROC rises from 0.615 to 0.988, while source AUROC under the common rewrite falls from 0.907 to 0.541 (Table 15). Its content-point counterpart retains source AUROC 0.955 and rewrite AUROC 0.522. This diagnostic supports the interpretation that the learned source score can follow the rewrite–source association. The comparison fixes the original-document population, not the commitments realized by each rewrite; it does

Corpus	Detector	Raw text	Whitespace normalized	**Ours** ϕ
All 24 voice settings				
HART	RoBERTa	0.707	0.681	0.955
HART	RACE	0.678	0.683	0.950
HART	DETree	0.791	0.782	0.951
RAID	RoBERTa	0.698	0.687	0.931
Reversed voice assignments (4 joint settings)				
HART	RoBERTa	0.016	0.015	0.944
HART	RACE	0.025	0.026	0.936
HART	DETree	0.302	0.251	0.943
RAID	RoBERTa	0.016	0.015	0.907

Table 16: **Content points retain source discrimination under re-voicing beyond formatting normalization.** Mean source-detection AUROC; each setting weights domains equally, then averages available runs, and each block weights settings equally. The lower block is the four-setting subset in which the voice assignment to human and LLM content reverses between training and testing. HART uses three runs for every input; RAID uses one for raw text and content points and three for normalized text.

not attribute every difference exclusively to expression. This single-run, pooled diagnostic is distinct from the domain-averaged, three-run HART comparisons below.

Control layout while keeping the detector fixed. To test whether paragraph differences account for the re-voicing loss, we give originals and their rewrites the same deterministic layout before scoring with frozen RoBERTa detectors trained on unmodified documents. We test flattened whitespace and several sentence-boundary paragraph policies. With all whitespace flattened and the GPT-4o storyteller rewrite applied to both content classes, raw-text source AUROC falls from 0.981 on originals to 0.798 on rewrites. With content points freshly extracted after flattening, AUROC changes from 0.970 to 0.952. These comparisons average domains equally and then three runs, using 1,995 complete groups for raw text and 1,988 for content points. Thus, source discrimination can deteriorate even when both classes receive the same rewrite and paragraph differences are removed. Wording and any substantive changes remain coupled in this control.

Normalize consistently during training and evaluation. The fixed-detector control leaves open whether learning on normalized inputs would prevent the dependence. We therefore collapse every whitespace run to a single space throughout training, validation, and testing, retaining other characters in order. We refit the three HART detectors and RoBERTa on RAID with their unchanged recipes and thresholds calibrated on their own normalized validation inputs; RACE's discourse trees are recomputed. All 12 paragraph settings become identical to the normalized unmodified reference by construction. The 24 voice settings comprise eight training-only, eight test-only, and eight joint settings (Appendix B.2).

Across the 24 voice settings, normalized-text mean AUROC stays within 0.03 of raw text in all four comparisons, while content points reach 0.931–0.955 (Table 16). Reversing the voice assignment between training and testing still reverses normalized detectors' rankings: mean AUROC is 0.015–0.251, compared with 0.907–0.944 for content points.

Normalization nevertheless helps some operating points. For HART RoBERTa under test-only voice changes, mean TPR/FPR changes from 0.917/0.482 to 0.893/0.307; content points yield 0.830/0.041. Normalized inputs also have higher unmodified-text AUROC than content points in all four comparisons. The advantage here concerns sustained discrimination across re-voicing conditions.

Reduce rewrite dependence while retaining source information. Together with the detailed plain-rewrite ablation (Appendix C.2), these results support plain restatement as a major contributor to voice robustness. That ablation concerns test-time changes and does not isolate restatement's contribution across the full training-and-testing grid. Retained detail and point structure have complementary roles in the shared representation.

Source excerpt	Rewrite excerpt	Commitment comparison
The car which is the driver doesn't drink or text.	The car's doing the driving—it doesn't drink, doesn't text, doesn't get distracted.	The rewrite generalizes two specific behaviors to absence of distraction. That stronger claim is not entailed by the two behaviors alone.
What if someone hacks and goes and controls the car?	What if someone hacks into the car and takes control?	Both raise the possibility of someone hacking the car and controlling it; this wording change does not introduce a different possibility.

Table 17: **A stronger commitment versus different wording.** Excerpts from two HART human essays, rewritten by GPT-4o using the main storyteller and casual instructions, respectively. The final column explains these local contrasts; document-level changes are measured separately below.

Content-point scores also retain systematic rewrite responses, particularly on LLM-content documents, even under a common layout. These controls do not separate altered commitments from expression retained through extraction. The following analysis examines substantive changes and targeted detector responses, complementing the demonstrated robustness advantage.

C.7 Substantive commitments under re-expression

Content-point detection substantially reduces expression sensitivity while retaining evidence about substantive commitments. This distinction matters because commitments are conveyed through expression: a rewrite can preserve a document's topic while weakening its claims or evaluations. We measure these changes independently of the detector, then use targeted edits to test whether evaluative commitments contribute to its predictions. Removing evaluations reduces detection evidence, whereas rewording instructed to preserve them retains strong discrimination. We examine these fixed-detector responses separately from the associations that changed inputs can introduce during training. Here *canonical* means the dataset document before the additional experimental rewrite; it includes the dataset's already-polished human and humanized LLM production conditions. Exact generation and assessment details follow in Appendix C.7.1.

Compare the commitments before and after rewriting. Paragraph changes provide a direct expression control because they preserve the words and their order. Voice rewrites require a further distinction: changing how a passage is expressed can also change what it asserts. With a source and its rewrite, we apply a common criterion in both directions: would accepting one passage commit a reader to everything the other asserts? An addition asserts something the source does not already imply; a strengthening makes an existing claim more specific or more demanding. Removal and weakening go in the opposite direction. Evaluative language matters when it changes an asserted judgment, such as a claim that a procedure is effective or an outcome is important. Tone, emphasis, or a conversational aside does not count merely because the wording differs.

Table 17 illustrates the distinction. Saying that a car does not drink or text does not by itself assert that it cannot be distracted in any way. By contrast, the two questions about hacking raise the same possibility. The comparison concerns what the text asserts or asks the reader to consider, rather than how formal or human-like it sounds.

We retain the original content-source labels throughout evaluation. The analysis concerns changes in the evidence available in the resulting document, not relabeling a rewritten human document as entirely LLM content when a claim is added. Likewise, "machine-content evidence" means information associated with that class in these data, not a kind of claim that humans cannot make. A response to changed evidence can still be an error against the retained provenance label.

A detector-independent measurement. We operationalize this criterion through a fixed procedure that separates locating candidate differences from assessing their direction. GPT-4o first reads the source and rewrite and proposes differences, quoting the relevant spans and stating the difference in commitment. Nonempty spans are checked against their attributed documents after case and whitespace normalization. Reports with no stated difference, or with the model answering that both spans imply each other, are discarded. For candidates with both spans present, Gemini 2.5 Flash then applies the two directional commitment questions to the quoted spans alone, without the

Rewrite	Documents	Additions/document	With an addition
Paragraph split	400	0.000	0.0%
Paragraph merge	400	0.000	0.0%
Storyteller, Llama	800	0.385	27.3%
Storyteller, GPT-4o	800	0.410	30.0%
Casual, GPT-4o	800	0.489	33.9%
Casual, Llama	800	0.565	40.6%
Storyteller alternative, Llama	800	0.786	47.9%
Storyteller alternative, GPT-4o	800	1.225	62.9%

Table 18: **Instructions to preserve content still permit measured additions.** Automated, detector-independent counts of new and strengthened commitments in original human HART test documents. Each sample is balanced across the four domains. Rows are ordered by mean addition count. "Storyteller" denotes the main content-controlled prompt; its alternative is specified locally.

surrounding documents, rewrite identity, content-point representation, or detector scores. Neither stage uses ϕ or a detector prediction. This separation allows us to compare measured commitment changes with detector responses without defining those changes by the responses themselves.

When the source span is empty and the rewrite span is present, the procedure records a new commitment directly. When both are present, a rewrite that implies the source but is not implied by it is classified as a strengthening. The reported addition count is the number of these two kinds of differences. The reverse entailment direction marks weakening; an empty rewrite span marks a missing commitment; failure of entailment in both directions marks a changed commitment. These are measurements from proposed spans, not exhaustive inventories: finding the relevant passage and judging what it implies can both fail. In particular, span verification establishes that the words occur, not that the semantic judgment is correct.

Controls apply to the category being measured. Paragraph splitting and merging preserve the source words and their order, with equality checked after whitespace normalization. On 400 human documents per layout, balanced across HART's four domains, the measurement reports zero added commitments. This provides an unchanged-word reference for the same addition measure applied to voice rewrites below (Table 18). The finder nevertheless proposes other apparent differences on 23 split documents and three merged documents; some are subsequently judged equivalent. The control supports the addition measure in this sample; it does not establish that the procedure never reports a difference on unchanged words or automatically validate weakening or removal counts.

The same layout changes need not yield identical extracted point lists. Extraction can vary in wording, grouping, coverage, or detail, and a later alignment of those lists adds its own comparison errors. Those differences must be distinguished from changes in the source document. The document-level measurement above bypasses extraction; point-alignment diagnostics examine a different combination of effects. The detection control below asks whether the variation that remains after extraction has a substantial aggregate effect, without requiring identical point strings.

Six voice rewrites, including a less constrained storyteller. We examine casual and storyteller rewrites from the main experiments, each generated by GPT-4o or Llama-3.3-70B, and a storyteller alternative from the earlier rewrite study. The main storyteller prompt in Appendix B.2 explicitly prohibits additions and sets a document-specific length band. The alternative retains the instruction to preserve every original point but permits digression and asides, omits the no-addition clause and length band, and uses different voice instructions. It is therefore a comparison between prompt formulations, not an experiment changing only one prohibition. The content-change measurement uses 800 original human (t_0) test documents per voice rewrite, 200 from each HART domain. These samples differ from the larger detection populations below.

Across the six voice rewrites, mean additions range from 0.385 to 1.225 per document, with at least one reported addition in 27.3–62.9% of documents (Table 18). Even the main storyteller instruction yields means of 0.385 for Llama and 0.410 for GPT-4o. The GPT-4o alternative produces the largest mean, 1.225, despite retaining the instruction to preserve the original points. Preserving existing claims and refraining from supplying new ones are separate requirements.

	Raw text			**Ours** ϕ		
Rewrite	RACE	RoBERTa	DETree	RACE	RoBERTa	DETree
Unmodified reference	0.058	0.073	0.061	0.062	0.050	0.048
Paragraph split	0.299	0.520	0.214	0.060	0.051	0.051
Paragraph merge	0.113	0.076	0.142	0.056	0.048	0.053
Storyteller, Llama	0.760	0.671	0.442	0.038	0.029	0.041
Storyteller, GPT-4o	0.720	0.626	0.439	0.050	0.042	0.067
Casual, GPT-4o	0.581	0.469	0.321	0.055	0.047	0.065
Casual, Llama	0.718	0.595	0.618	0.042	0.033	0.042
Storyteller alternative, Llama	0.904	0.752	0.882	0.054	0.045	0.060
Storyteller alternative, GPT-4o	0.822	0.790	0.627	0.081	0.080	0.128

Table 19: **False-positive rates under uniform test-time rewriting.** Thresholds target 5% pooled validation FPR and remain fixed. Rates weight the four domains equally; raw RACE averages eight seeds and raw DETree three. Addition counts in Table 18 use smaller original-human samples; these FPRs use the available full human-content test populations.

Input	Detector	t_0	t_1	t_2	t_3	$t_2 + t_3$
Raw text	RACE	$+0.0019$	$+0.0088$	-0.0206	$+0.0100$	-0.0106
Raw text	RoBERTa	-0.0043	$+0.0365$	-0.0240	-0.0082	-0.0321
Ours ϕ	RACE	-0.0090	$+0.0005$	$+0.0049$	$+0.0036$	$+0.0085$
Ours ϕ	RoBERTa	-0.0177	$+0.0075$	$+0.0049$	$+0.0052$	$+0.0101$

Table 20: **Added commitments have different probability associations on raw text and content points.** Descriptive slopes in class-probability change per measured added commitment, after removing each rewrite's mean probability change. The analysis pools six voice conditions; the final column sums the two LLM-content classes. DETree has no corresponding four-probability head.

Layout sensitivity at a fixed validation-calibrated threshold. We next evaluate the fixed RACE, RoBERTa, and DETree fits from this diagnostic study, trained on unmodified inputs in their respective raw-text or content-point representation. One renderer is applied uniformly at test time across the four production conditions. This eight-renderer diagnostic is separate from the main 36-setting evaluation. Each checkpoint retains the threshold targeting 5% FPR on pooled human-content scores from its unmodified validation partition. There is no refitting or test-time recalibration. FPR is evaluated on the available human-content test documents (t_0, t_1), nominally 4,000 across 2,000 groups, rather than only the 800 t_0 documents used to measure additions.

Despite unchanged source words, splitting versus merging changes raw-text FPR by 0.186 for RACE, 0.445 for RoBERTa, and 0.072 for DETree. The corresponding absolute gaps on content points are only 0.00375, 0.00275, and 0.00225; their FPRs stay between 0.048 and 0.060. Thus all three content-point detectors retain stable aggregate false-positive rates under these layouts despite variation in the extracted lists.

Voice rewrites change expression and content together. Their mean addition counts and FPRs should not be read as a causal dose–response curve. The GPT-4o storyteller alternative has both the largest measured addition count and the highest content-point FPR for each detector, at 0.081 for RACE, 0.080 for RoBERTa, and 0.128 for DETree. That coincidence motivates examining individual documents under a shared instruction; it does not isolate additions from all other changes the instruction produces.

Prediction shifts among documents receiving the same instruction. For RACE and RoBERTa, we inspect the four production-condition probabilities on the original-human (t_0) source–rewrite pairs used to measure additions. For each document and predicted class, we subtract the source probability from the rewrite probability, then subtract that rewrite condition's mean change. We pool these centered probability changes across the six voice conditions and regress them on the document's measured addition count. This describes the direction of the remaining association after removing the average response to each rewrite instruction. Documents receiving that instruction can still differ in realized wording, length, and content changes.

The raw-text net slopes toward LLM content are approximately -0.011 for RACE and -0.032 for RoBERTa (Table 20). In particular, raw RoBERTa's probability for polished human writing rises

Rewrite	Added	Weaker or missing	Changed	Total
Paragraph split	0.000	0.010	0.000	0.010
Paragraph merge	0.000	0.000	0.000	0.000
Storyteller, Llama	0.115	1.160	0.315	1.590
Storyteller, GPT-4o	0.180	1.600	1.220	3.000
Casual, Llama	0.260	1.335	0.610	2.205
Casual, GPT-4o	0.290	1.460	1.145	2.895
Storyteller alt., Llama	0.410	1.300	1.100	2.810
Storyteller alt., GPT-4o	0.930	0.880	2.020	3.830

Table 21: **Machine-text rewrites often weaken existing commitments.** Mean automated counts per document on 200 paired HART machine-content test documents per renderer. Added includes new and strengthened commitments; changed means neither quoted span entails the other. Total excludes equivalent spans. The finder, span checks, direction prompt, and model settings are those in Appendix C.7. These counts compare source documents directly, without using extracted points or detector scores.

by 0.0365 per additional measured commitment in this fit. That class counts as human content in the binary score, so its increase accompanies a net movement toward the human-content side. This describes how the fitted detector distributes probability; it does not establish who authored a particular added claim from the score alone.

With content points, the net association reverses direction: approximately $+0.009$ for RACE and $+0.010$ for RoBERTa. Documents gaining more measured commitments tend to receive more probability on the LLM-content classes after the shared instruction's average effect is removed. The effect is small. In a companion analysis of normalized binary-score changes on 4,798 document–rewrite pairs, addition counts explain only 0.18% and 0.24% of the remaining variance for content-point RACE and RoBERTa. Within-rewrite permutation tests support positive associations ($p = 0.0044$ and 0.0008), while a related regression using uncertainty clustered over the six rewrite conditions has a 95% interval that includes zero for content-point RoBERTa. These diagnostics support a limited systematic association, not an explanation of most document-level movements or a causal effect established by holding the instruction fixed.

The layout control establishes reduced sensitivity to unchanged-word transformations; the addition analysis identifies a small association with independently measured content changes. We next examine the complementary case of weakened commitments in machine writing and test their role through direct edits and training comparisons.

Machine-text rewrites can weaken commitments without removing the topic. We apply the same document-level measurement to 200 machine-content test documents per rewrite: 25 from each domain and each LLM-content production condition (t_2, t_3). The same document identities are used for all eight rewrites. This is a new sample, distinct from the original-human addition sample above. Under the main GPT-4o storyteller rewrite, the instrument reports 1.600 weakened or missing commitments per document, comprising 1.565 weakened commitments and 0.035 missing ones (Table 21). Paragraph splitting and merging yield corresponding means of 0.010 and 0.000. These controls concern the document-level weakening measurement; they do not estimate the variability of separately extracted point lists.

For example, an LLM essay calls the Seagoing Cowboys program "an unparalleled opportunity for personal growth, cultural immersion, and meaningful impact on communities in need." Its GPT-4o storyteller rewrite describes "something that'll stretch you, throw you into new cultures, and let you really help people" and calls the program "the real deal." The rewrite retains the endorsement and the opportunities described, but no longer asserts that the opportunity is unparalleled. This is a narrower commitment even though the passage remains enthusiastic.

Removing whole points does not reproduce the training loss. The GPT-4o storyteller rewrite also impairs learning when applied only to machine training and validation documents: RoBERTa's canonical-test AUROC falls from 0.970 to 0.936. We first ask whether disappearance of whole points accounts for this loss. We align the lists extracted from each canonical document and its rewrite, with GPT-4o judging coverage in both directions and allowing one claim to be covered by several points. A canonical point is designated for removal or restoration only if the rewritten list is

Machine-content training/validation input	Seeds	AUROC
Canonical reference	3	0.970
Canonical, identified points removed	1	0.969
Storyteller rewrite	3	0.936
Storyteller, identified points restored	1	0.936

Table 22: **Deleting identified whole points does not reproduce the loss, and restoring them does not repair it.** RoBERTa on **Ours** ϕ, with canonical human training/validation inputs and canonical test inputs. AUROC weights domains equally; the seed count is shown per row. Alignment succeeds for 3,988 of 4,000 machine training/validation documents. The remaining 12 retain their input lists without removal or restoration.

judged not to cover it at all; partly covered points are retained. This diagnostic depends on extraction and alignment, unlike the document-level measurement above.

On the 3,992 machine test documents with complete alignments, approximately 0.33 canonical points per document are judged wholly uncovered, versus 0.04–0.05 in the paragraph-split control. The latter compares extractions from documents whose source words are unchanged. Differences between those lists can reflect extraction as well as alignment, so an uncovered point is not automatically a claim deleted from the source. The intervention asks whether the *identified* whole-point losses explain the detector result.

We use the training and validation documents' own alignments to construct two modified fitting corpora: one removes the identified points from canonical machine lists; the other restores them to the storyteller-derived lists. Human-content documents remain canonical, and evaluation uses canonical test documents. Test-document alignments do not determine these training edits. Removing the identified points leaves AUROC at 0.969; restoring them leaves it at 0.936 (Table 22). The canonical and storyteller references average three seeds. A local repetition of storyteller training without restoration gives 0.935, consistent with the reference comparison. These results do not support the identified whole-point losses as the explanation for the training failure.

Targeted edits test the contribution of evaluative commitments. We next edit canonical machine content points directly, so that the detector receives the intervention without another extraction pass. The first instruction removes evaluation, emphasis, and intensity while preserving other material. Its scope includes judgments of importance and effectiveness, certainty, and what should be done. A point containing only evaluation can become `NONE` and be removed. The second instruction rewords every point while preserving its commitments, explicitly including evaluations and qualifications, at the same vocabulary level. These are interventions defined by their instructions and checked below; neither is assumed to be a perfectly isolated operation.

The checks support different effects, with imperfect separation. We assess 2,000 randomly sampled source–edited point pairs per intervention from the machine test documents. Gemini uses the same commitment-direction assessment as for document spans, now comparing the two whole points; `NONE` is marked missing directly. A separate GPT-4o assessment identifies kinds of differences and which point asserts more, without seeing detector scores or the surrounding document.

Evaluation removal is judged to weaken the source commitment in about 60% of sampled pairs and to remove it in another 3%; the kinds assessment marks an evaluation difference in about 59%. Rewording produces a much smaller weakening rate, about 5%, and Gemini judges about 66% of its pairs equivalent. The same commitment-direction criterion used for document spans (Appendix C.7) therefore distinguishes these edit conditions in the intended direction. The separation remains imperfect: GPT-4o applies the wording-only label to 97.5% while also marking an evaluation difference in 24.8% (Table 23). The kinds prompt permits multiple labels, and the returned categories are not mutually exclusive in practice. We therefore do not interpret the wording-only percentage as the fraction independently certified to have unchanged content. The disagreement limits how cleanly the edits separate wording from commitments.

Point-level judgments also have a substantial control baseline: among 10,339 mutually aligned point pairs from 1,000 machine documents under paragraph splitting, Gemini judges only 42% equivalent, versus 9% of 9,721 storyteller pairs from 999 documents. The storyteller diagnostic selects every fourth document in the sorted alignment records within each machine condition and retains mutual

Assessment outcome	Evaluation removal	Rewording
Equivalent	34.0%	65.6%
Weaker	60.1%	5.2%
Stronger	1.4%	6.5%
Changed	1.6%	22.6%
Missing	3.0%	0.0%
Unresolved	0.0%	0.1%
Evaluation	59.2%	24.8%
Wording only	25.6%	97.5%
Source asserts more	64.4%	2.9%

Table 23: **Automated assessment of the two direct point edits.** Each column uses 2,000 machine-test point pairs. The first six rows are the Gemini commitment-direction categories (missing points are assigned directly). The next two are selected GPT-4o difference-kind labels, which can co-occur; the last is its separate assertion-direction judgment. Percentages retain the saved aggregate precision and may not sum to exactly 100%.

Machine-content test input	AUROC	TPR	Mean log-odds change
Canonical reference	0.970	0.896	$+0.00$
Evaluation removal	0.941	0.760	-1.75
Storyteller rewrite	0.948	0.782	$-1.69^{\dagger}$
Rewording	0.976	0.917	$+0.27$

Table 24: **Evaluation removal reduces detection evidence; the tested rewording preserves discrimination.** A fixed RoBERTa detector on **Ours** ϕ, with unchanged human inputs and threshold. AUROC/TPR use the full test set. Log-odds changes use paired machine inputs: 4,000 for direct edits; †3,992 for the storyteller comparison. Each change uses its own paired canonical reference. Storyteller rewrites precede extraction; direct edits act on extracted points.

one-to-one matches. These are pairs of *extracted* points; the paragraph comparison can include extraction differences despite unchanged source words. This baseline is neither a document-level content-change rate nor an error rate for the earlier addition instrument. It reinforces the need to interpret point categories alongside the direct interventions and their detector effects, rather than treating every non-equivalent pair as a source-level content change.

A fixed detector loses evidence when evaluations are removed. Table 24 applies the two edits to the 4,000 machine-content test documents while leaving the 4,000 human-content test documents canonical. RoBERTa is fitted once on canonical content points and keeps its threshold targeting 5% pooled human-content validation FPR. Its test FPR remains 0.050 because the detector, threshold, and human inputs are unchanged. AUROC and TPR average the four domains equally. To show score movement before thresholding, we also report the mean change in the log-odds of $P(t_2) + P(t_3)$ on paired machine documents, clipping that probability to $[10^{-7}, 1 - 10^{-7}]$ before taking log-odds.

Removing evaluations lowers mean machine log-odds by 1.75 and TPR from 0.896 to 0.760. The storyteller rewrite produces a similar aggregate decrease, approximately 1.69 log-odds and a TPR of 0.782. Rewording instead raises mean log-odds by 0.27, with AUROC/TPR of 0.976/0.917 compared with the canonical 0.970/0.896. The fixed detector therefore retains strong discrimination after the tested rewording while responding markedly to removal of evaluations. Together with the intervention checks, this contrast provides evidence that evaluative commitments contribute to detection. The edits do not perfectly isolate commitments from wording, and similar aggregate decreases do not establish a complete explanation of the storyteller effect.

Training on changed commitments can introduce a different failure. The test-time comparison changes the evidence given to a fixed detector. During training, the same changes can instead determine which features become associated with the source label. To examine this distinction, we compare training with the GPT-4o storyteller rewrite applied to both content classes against training with it applied only to machine content. The corresponding validation inputs follow the training condition; every detector is evaluated on canonical HART test documents. The three-seed RoBERTa AUROC is 0.970 with neither class rewritten, 0.967 with both classes rewritten, and 0.936 with only

Change to training and validation inputs	Seeds	AUROC	TPR	FPR
Neither class changed	3	0.970	0.897	0.051
Both classes: storyteller rewrite	3	0.967	0.910	0.084
Machine only: storyteller rewrite	3	0.936	0.691	0.038
Machine only: evaluation removal	1	0.508	0.249	0.057
Machine only: rewording	1	0.857	0.561	0.046

Table 25: **Changes associated with one class during training can impair canonical-test detection.** RoBERTa on **Ours** ϕ; metrics weight domains equally, with seed counts shown per row. Each model uses its own validation-calibrated threshold; both-class rewriting also rewrites validation inputs. Every row evaluates canonical test documents. Read TPR together with FPR; AUROC compares ranking.

		Ours ϕ: both-class minus machine-only AUROC		
Training renderer	Changes/document	RoBERTa	RACE	DETree
Paragraph merge	0.00	-0.0001 ± 0.0007	$+0.0008 \pm 0.0002$	$+0.0008 \pm 0.0007$
Storyteller, Llama	1.59	$+0.0307 \pm 0.0034$	$+0.0233 \pm 0.0032$	$+0.0005 \pm 0.0023$
Storyteller alt., Llama	2.81	$+0.0367 \pm 0.0052$	$+0.0256 \pm 0.0155$	$+0.0054 \pm 0.0011$
Storyteller, GPT-4o	3.00	$+0.0308 \pm 0.0024$	$+0.0299 \pm 0.0059$	$+0.0054 \pm 0.0011$
Storyteller alt., GPT-4o	3.83	$+0.0581 \pm 0.0033$	$+0.0497 \pm 0.0210$	-0.00003 ± 0.0027

Table 26: **The association between commitment changes and training loss depends on the detector.** Gaps are means $\pm$ sample standard deviations of paired differences across three seeds, after equal-domain AUROC aggregation on canonical test inputs. Changes/document comes from the separate machine-test sample in Table 21. Paragraph fits use merging in both conditions; the extracted-point control uses splitting.

machine content rewritten (Table 25). The content-source distinction therefore largely survives this rewrite when both classes receive it, while the class-specific training condition incurs a larger loss.

We summarize the additional loss associated with rewriting only machine content as the AUROC under both-class rewriting minus the AUROC under machine-only rewriting. This gap is descriptive: changing which classes are rewritten also changes the number and distribution of rewritten documents, so it does not isolate label correlation as the only changed variable. Table 26 compares it across the five available training renderers with independently measured document-level commitment changes. For content-point RoBERTa and RACE, larger measured changes generally accompany larger gaps; the five-renderer rank correlations are 0.90 and 1.00. DETree's gaps are much smaller and differently ordered (rank correlation -0.10). Content and realized wording changes co-vary across these few renderers, so this association does not separate their contributions.

Paragraph merging supplies the direct training-side control. Its both-class-minus-machine-only AUROC gap is 0.329 for raw-text RoBERTa and 0.431 for raw-text RACE, despite unchanged source words and zero reported commitment changes in the corresponding document-level sample. The content-point gaps are -0.0001 and $+0.0008$; DETree's raw/content-point gaps are 0.040 and 0.0008. Thus layout alone supports a harmful training association on raw text, whereas its observed aggregate cost after extraction is very small.

Tolerance of a fixed detector does not prevent learning a wording difference. Finally, we apply the same direct edits from Appendix C.7 only to machine training/validation points, retaining canonical human inputs and canonical test documents. Evaluation removal yields AUROC 0.508, compared with the canonical reference near 0.970. The rewording fit also loses discrimination, reaching 0.857 (Table 25). The former shows that a class-correlated intervention targeting substantive commitments can produce a severe training failure. The latter shows that even the rewording intervention tolerated by the fixed canonical detector can support a harmful learned association. Its imperfect semantic checks preclude attributing that entire loss to wording alone.

Both edits bypass extraction of a rewritten source document. They test what RoBERTa can learn from systematic differences already present in its point-text input, not how successfully extraction removes those differences from a source-level rewrite. The fixed-detector and refitting comparisons also differ in training conditions and replication; they are not a fully matched factorial experiment. Their joint lesson is that robustness of an existing detector and resistance to learning a shortcut

are different properties. The advantage of ϕ is its reduction of expression differences reaching the detector, not an inability of the downstream text classifier to learn wording that survives. This distinction complements the training-scope results in Section 5.1 and the objective comparisons in Appendix C.3.

Together with the representation ablations, these experiments support the combined benefit of reducing expression sensitivity while retaining information about substantive content. Detailed points sustain source discrimination under expression changes; targeted evaluation removal lowers a fixed detector's evidence for LLM content, whereas rewording instructed to preserve commitments retains strong discrimination. This contrast provides evidence that evaluative commitments contribute to detection, although the edits do not perfectly separate commitments from wording. These aggregate responses do not establish recognition of each individual change or explain every residual error. The training comparisons identify a separate vulnerability to class-correlated differences that remain in point text.

C.7.1 Generation and assessment details

The following settings and exact messages specify the measurements and interventions above.

Commitment finding and direction assessment. The finder uses GPT-4o (2024-11-20), temperature 0, top-p 1, and at most 1,600 output tokens, with each input text limited to its first 24,000 characters. Its exact system message and user template are:

```
You are a careful analyst. You compare two versions of the same text and
↪ identify only where they commit to different things about the world,
↪ ignoring how each is worded.
```

```
The PASSAGE below is taken from one version of a text. The OTHER VERSION
↪ is a re-expression of the whole text. Report only the places where the
↪  two do not commit to the same things.

Work claim by claim. A claim is something that could be true or false.
↪ Wording, tone, register, sentence structure, paragraphing, address to
↪ the reader, and turns of phrase that make no claim are not content,
↪ however much of them there is and however differently they read.

For each claim in the PASSAGE, find what the OTHER VERSION says about the
↪  same matter and answer both:
- Would a reader given only the PASSAGE be committed to everything the
↪ OTHER VERSION asserts here?
- Would a reader given only the OTHER VERSION be committed to everything
↪ the PASSAGE asserts here?
If both are yes, the two agree; report nothing. Otherwise report it.

Rules:
- Quote verbatim: the exact span from each version, copied character for
↪ character. Use "" where a version says nothing about the matter.
- State the difference in commitment as a sentence that could be checked
↪ as true or false. If you cannot state it that way, the difference is
↪ one of wording and must not be reported.
- Report each difference once, under the claim it belongs to.
- Output only JSON, with no preamble.

{"differences":[{"passage_span":"...","other_span":"...","
↪ passage_commits_to_other":true,"other_commits_to_passage":true,"
↪ difference":"..."}]}

PASSAGE:
{A}

OTHER VERSION:
{B}
```

For pairs with both spans present, the span evaluator uses Gemini 2.5 Flash, temperature 0, top-p 1, and at most 60 output tokens. Each span is limited to 4,000 characters. Empty-span cases receive the direction described above without an additional model call. The exact system message and user template are:

```
You are a careful analyst of what a statement commits a reader to. You
↪ answer only the two questions asked, with no explanation.
```

```
Here are two versions of the same passage from a text.

A: {a}
B: {b}

Answer both questions about what a reader is committed to:
1. Would a reader who accepted A be committed to everything B asserts?
2. Would a reader who accepted B be committed to everything A asserts?

Output only JSON, with no preamble: {{"a_commits_b": true|false, "
↪ b_commits_a": true|false}}
```

Alternative storyteller generation. The alternative's exact user prompt is:

```
Rewrite the text below in the voice of a talkative storyteller speaking
↪ to an audience --- warm, digressive, full of everyday idiom. Follow
↪ these rules:
- Use spoken idiom, asides, and the loose rhythm of somebody telling you
↪ about it in person.
- Do not use formal, literary, or academic wording; no ornate imagery and
↪  no elevated vocabulary.
- Output only the rewritten text, with no preamble.
- Preserve every fact, claim, example, number, name, and detail of the
↪ original; do not drop, shorten, or summarize anything --- every point
↪ of the original must still appear.

TEXT:
{text}
```

Both alternative renderers use an empty system message, temperature 0.7, top-p 1, and at most 2,500 output tokens, with GPT-4o (2024-11-20) or Llama-3.3-70B-Instruct.

Point alignment. The alignment uses GPT-4o (2024-11-20), temperature 0, top-p 1, and at most 4,000 output tokens. Each list is numbered with its side and point index. Responses must supply a valid coverage category for every point; incomplete responses are retried up to three calls. The exact system message and user template are:

```
You are a careful analyst. You compare two lists of content points
↪ extracted from two versions of one document, and identify which claims
↪  one list asserts that the other does not, ignoring how each is worded
↪ .
```

```
LIST A and LIST B are content points extracted from two versions of the
↪ same document. Each point states one claim.

For each point in LIST A, decide whether a reader given only LIST B would
↪  be committed to what that point asserts. The claim may be stated in
↪ one point of LIST B, split across several, or folded into a longer
↪ point; wording does not matter. Answer "yes" if LIST B commits the
↪ reader to all of it, "partly" if LIST B commits the reader to only
↪ part of it or to a weaker version of it, and "no" if LIST B does not
↪ commit the reader to it at all. List the LIST B points that cover it.
```

```
Then do the same for each point in LIST B against LIST A.

Output only JSON, with no preamble:
{"A":[{"id":"A1","covered":"yes|partly|no","by":["B2"]}],"B":[{"id":"B1
↪ ","covered":"yes|partly|no","by":["A1"]}]}

LIST A:
{A}

LIST B:
{B}
```

Direct point edits. Both direct edits use GPT-4o (2024-11-20), temperature 0, top-p 1, and at most 4,000 output tokens, in one call per document's numbered point list. The response must retain one numbered line per input point before `NONE` lines are removed; malformed responses are retried up to three calls. All 4,000 machine test documents and all 4,000 machine training/validation documents have accepted outputs for both edits. They share this exact system message:

```
You edit numbered lists of statements exactly as instructed.
```

The evaluation-removal user template is:

```
Below is a numbered list of statements. Rewrite each statement so that it
↪  no longer asserts any evaluation, emphasis, or intensity --- how
↪ important, good, bad, large, notable, strong, effective, or certain
↪ something is, or what should be done --- and change nothing else: keep
↪  every other word, fact, name, number, event, and relation as it is.
↪ If a statement asserts only an evaluation, write the underlying fact
↪ without the evaluation if there is one, and write NONE if there is
↪ none. A statement that asserts no evaluation is copied unchanged.

Return exactly {n} lines, numbered 1 to {n}, one per statement, with no
↪ other text.

{list}
```

The rewording user template is:

```
Below is a numbered list of statements. Re-express each statement in
↪ different words while asserting exactly the same thing: keep every
↪ fact, name, number, event, relation, qualification, and every
↪ evaluation, emphasis, or intensity it asserts. Use vocabulary of the
↪ same level as the original, neither simpler nor more elaborate.

Return exactly {n} lines, numbered 1 to {n}, one per statement, with no
↪ other text.

{list}
```

Here `{n}` is the input point count and `{list}` is the numbered list. The resulting retained points are concatenated in order for the detector, using the usual content-point input format.

Point-edit assessment. The Gemini assessment uses exactly the messages in Appendix C.7.1, with the two whole points replacing document spans, at temperature 0, top-p 1, and at most 60 output tokens. The separate GPT-4o assessment uses temperature 0, top-p 1, at most 80 output tokens, and these messages:

```
You are a careful analyst. You compare two statements that restate the
↪ same point and say what kind of difference, if any, there is in what
↪ they assert.
```

```
Statements A and B restate the same point, taken from two versions of one
↪  document.

Compare what each one asserts, and list every kind of difference that
↪ applies:
- EVALUATION: they differ in an evaluation, emphasis, or intensity they
↪ assert about something (how important, good, large, notable, strong,
↪ or certain it is).
- SPECIFICITY: one states a detail (a number, name, date, place, quantity
↪ , or precise description) that the other states less precisely or
↪ leaves out.
- ATTRIBUTION: they differ in whether or to whom the point is attributed,
↪  or whether it is presented as someone's view rather than as fact.
- OTHER_CONTENT: any other difference in what is asserted.
- WORDING_ONLY: they assert the same thing in different words.
Then say which statement asserts more, counting only the differences you
↪ listed: "A", "B", or "neither".

A: {a}
B: {b}

Output only JSON, with no preamble: {"kinds": ["..."], "more": "A|B|
↪ neither"}
```

D CRO: ADDITIONAL FINDINGS, COMPARISONS, AND METHOD DETAILS

CRO reveals recurring differences in the claims, details, and evaluations conveyed by human and LLM writing. The domain repertoires yield 32–54 supported choices each; the pooled repertoire yields 57 choices with 120 supported choice–domain associations. We first present these findings and their evidence, then specify how memberships, population differences, and descriptions are evaluated. Comparisons with SAE-TM and SAE-ATD clarify the contribution of the full procedure, and learning and naming ablations examine its design choices. Finally, we assess what persists when the conditions for measuring population differences change and when the repertoire itself is refitted.

D.1 CONTENT-CHOICE DIFFERENCES ACROSS DOMAINS

These content choices can overlap, and either population can express them: their differences lie in the share of each population's extracted points assigned to a choice within matched contexts.

Tables 27–31 present the same three blocks for each domain and the pooled repertoire. CRO lists every supported content choice, with independent human-favored and LLM-favored columns; neighboring entries are not matched opposites. Domain pages prefix names with discovery odds ratios, while the pooled page identifies supporting domains and preserves changes in direction. For compactness, names omit introductory "focus on," "emphasis on," or "discussion of" where present; selected names link to examples with the exact returned names and descriptions.

SAE-TM displays selected native outputs: three distinct features with the largest best-match membership IoU to supported CRO choices, and the two highest-mass topics in its 20-topic solution, each showing its first ten ranked words. Features and merged topics remain separate. SAE-ATD lists every available detector-selected name under the original interpretation instruction in our evaluated implementation, with underscores displayed as spaces and repeated names retained. These names describe changes in generated text under feature steering. All scopes use the same essay interpretation probe; domain labels identify feature-selection populations. Our additional random-feature control and human–LLM activation analysis appear separately in Appendix D.4. These conventions apply to all five pages.

News

CRO: all 54 supported content choices

Human-favored	LLM-favored
(4.5×) quantitative ratios or metrics	(7.0×) emphasizing importance
(4.1×) specific dates or timeframes	(6.3×) technological advancements
(3.9×) quantitative specifics	(5.3×) environmental sustainability
(3.9×) quantitative stock or market data	(5.1×) actions taken to address challenges
(3.7×) specific reported details	(4.7×) highlighting strengths or skills
(3.4×) specific timing details	(4.4×) describing strategic or operational initiatives
(3.1×) specific numerical details	(4.0×) evaluating reactions or implications
(3.0×) specific events tied to time	(3.6×) growth and its drivers
(2.9×) specific locations or entities	(3.5×) highlighting achievements or recognition
(2.8×) specific individual or event focus	(3.4×) explaining the rationale behind decisions or actions
(2.7×) specific factual updates	(3.3×) significance
(2.6×) naming individuals and their roles	(3.3×) specific applications or use cases
(2.5×) direct quotation use	(3.2×) stating a goal or purpose
(2.5×) anchored in specific timeframes	(2.8×) challenges faced
(2.4×) monetary amounts	(2.7×) specific elements or outcomes
(2.4×) specific factual details	(2.6×) economic and market dynamics
(2.3×) specificity of detail	(2.6×) potential outcomes or impacts
(2.3×) criminal justice and legal processes	(2.4×) emotional or societal themes
(2.2×) duration emphasis	(2.4×) conflict and resolution
(2.2×) absence or lack of action or information	(2.3×) expressing gratitude or praise
(2.0×) attribution of statements or perspectives	(2.2×) a single company's actions or performance
(1.9×) categorization by segments	(2.1×) safety and security concerns
(1.8×) specific events or actions	(2.1×) urban and community focus
(1.6×) reporting specific outcomes or events	(2.1×) team dynamics or preparation
(1.6×) structured market or industry reporting	(2.0×) generalized descriptions
(1.5×) specific quantitative or temporal details	(1.9×) enumerating examples
	(1.9×) highlighting individual contributions or qualities
	(1.5×) continuity or future developments

SAE-TM: native ranked words

Output	First 10 ranked words
Feature	three, third, year, quarter, record, million, approximately, previous, total, least
	title, know, describe, say, call, back, right, already, one, word
	report, market, analysis, detailed, comprehensive, key, scope, industry, examination, research
Merged topic	emphasize, also, support, improve, analysis, importance, need, many, may, include
	game, team, play, match, win, first, league, victory, white, score

SAE-ATD: descriptions of detector-selected features

counterarguments emphasis
multilingual insertion
rebuttal emphasis
emphasize argument support
argumentative structure
argumentative structure
argument structure
run-on-technical
scientific argumentation
counterargument emphasis

Table 27: **News comparison.** Parentheses give CRO discovery odds ratios. Selected names link to examples; shared display conventions precede these pages.

arXiv

CRO: all 32 supported content choices

Human-favored	LLM-favored
(5.3×) specific technical details	(5.3×) dynamic adaptation
(4.9×) formal mathematical definitions or assumptions	(3.8×) proposing improvements
(4.3×) providing specific examples or definitions	(3.5×) collaboration and interdisciplinarity
(3.9×) describing sections of a document	(3.4×) describing a framework
(2.9×) formal problem-solving in theoretical contexts	(3.3×) general significance across fields
(2.6×) abbreviations or acronyms	(2.9×) criticality or necessity
(2.5×) prior or ongoing research	(2.9×) advancements in technology
(2.1×) factual statements or statistics	(2.4×) research contributions and implications
(1.9×) alternative methods or approaches	(2.3×) system-focused evaluation or description
(1.8×) tasks as a central concept	(2.1×) computational efficiency
(1.7×) technical methods or processes	(2.1×) identifying challenges
(1.6×) identifying limitations or gaps	(1.9×) increasing demand or growth
(1.6×) enumerating components or examples	(1.9×) limitations of traditional methods
(1.6×) description of datasets	(1.8×) describing a specific approach
(1.4×) image-focused tasks and methods	(1.7×) listing application domains
	(1.6×) simulation-focused analysis
	(1.6×) limitations of existing methods or approaches

SAE-TM: native ranked words

Output	First 10 ranked words
Feature	framework, frame, unified, comprehensive, adaptable, utility, new, move, establish, rigorous system, systematic, control, capable, operation, complex, dynamic, data, design, use challenge, address, face, pose, challenging, present, major, difficulty, relate, arise
Merged topic	paper, problem, control, section, number, feature, demonstrate, effect, set, work system, performance, traditional, dynamic, energy, power, new, data, diverse, multiple

SAE-ATD: descriptions of detector-selected features

html markup insertion
language injection
counterclaims and rebuttals
support claims
argument structure
counterargument emphasis
argumentative structure
argumentative structure
emphasize argumentation
structured argumentation

Table 28: **arXiv comparison.** Parentheses give CRO discovery odds ratios. Selected names link to examples; shared display conventions precede these pages.

Creative writing

CRO: all 43 supported content choices

Human-favored	LLM-favored
(3.7×) self-reflective commentary	(10.7×) a named central character
(3.3×) a specific woman	(5.0×) transformation and renewal
(2.8×) cars or vehicles	(4.4×) patterns and symbols
(2.7×) use or presence of firearms	(3.6×) connections and bonds
(2.5×) physical injury and its effects	(3.2×) descriptions of atmospheric light and weather
(2.5×) dialogue or reported speech	(3.0×) sensory immediacy
(2.4×) children or parent-child relationships	(3.0×) world-altering transformations
(2.3×) interpersonal interaction	(2.9×) depictions of natural landscapes
(2.3×) departure or leaving	(2.6×) supernatural or otherworldly phenomena
(2.1×) recognition or realization	(2.5×) cities
(2.0×) specific time spans or ages	(2.2×) a focus on personal growth and admiration
(1.9×) specific actions or events involving characters	(2.1×) descriptions of specific actions or outcomes
(1.9×) physical movement involving falling or instability	(2.0×) describing a room's atmosphere or details
(1.8×) immediate and concrete events	(1.9×) books as central objects
(1.7×) posing questions	(1.8×) descriptions of personal appearance or demeanor
(1.7×) narrator's immediate actions or reactions	(1.7×) encounters with unusual figures or phenomena
(1.7×) wealth and monetary transactions	(1.6×) fear and its effects
(1.7×) decision-making or deliberation	(1.5×) human interactions and reactions
(1.6×) physical symptoms and health struggles	(1.3×) superhuman abilities
(1.5×) interaction with doors or windows	(1.3×) reflective introspection
(1.4×) absence of reaction or consequence	
(1.4×) directly quoted speech or text	
(1.3×) frustration and conflict	

SAE-TM: native ranked words

Output	First 10 ranked words
Feature	phrase, say, title, term, call, word, address, label, frequency, repeat question, ask, wonder, happen, aloud, do, much, narrator, might, possible drive, car, driver, vehicle, road, highway, ride, traffic, driving, swerve
Merged topic	despite, describe, together, reflect, though, saw, realize, observe, felt, say narrator, light, power, world, sound, existence, small, appearance, space, large

SAE-ATD: descriptions of detector-selected features

elaborative explanations
repetition and redundancy
emphasize support
counterargument elaboration
structured argumentation
argumentative essay guidance
argument structure
scientific argumentation
counterargument emphasis
argument structure

Table 29: **Creative writing comparison.** Parentheses give CRO discovery odds ratios. Selected names link to examples; shared display conventions precede these pages.

Essays

CRO: all 34 supported content choices

Human-favored	LLM-favored
(7.6×) Luke's personal experiences and decisions	(9.2×) evaluating FACS in education
(4.0×) limitations of Venus missions	(5.3×) framing autonomous vehicles as transformative
(3.9×) human driver alertness and control	(5.3×) adapting teaching to students' emotions
(3.5×) expressing beliefs or perceptions	(5.0×) cybersecurity vulnerabilities
(3.1×) computers' capabilities and limitations	(4.3×) economic and social impacts of job displacement
(2.3×) comparison to Earth	(3.9×) advocating program participation
(2.1×) Venus's atmospheric composition and pressure	(3.8×) linking Venus to Earth's climate
(2.0×) negative evaluation or difficulty	(3.8×) cautious evaluation
(2.0×) listing specific categories	(3.3×) raising concerns
(1.8×) describing travel experiences	(2.9×) analyzing complexities
(1.6×) sensors and technology in driverless cars	(2.9×) scientific evaluation of natural origin
	(2.9×) personal growth through impactful experiences
	(2.7×) emphasizing benefits
	(2.5×) summarizing the article's content
	(2.4×) participant camaraderie
	(2.4×) building a reasoned case
	(2.2×) emphasizing transformative experiences
	(2.0×) advocating Venus exploration despite risks
	(1.8×) general reflections on technology and its implications
	(1.6×) technological progress
	(1.6×) specific implications or consequences
	(1.5×) benefits of driverless cars
	(1.4×) human error as the cause of traffic accidents

SAE-TM: native ranked words

Output	First 10 ranked words
Feature	autonomous, vehicle, transportation, safety, potential, reduce, traffic, significant, mobility, development
	significant, lead, due, like, potential, understanding, offer, technological, autonomous, reduced
	learning, educational, education, potential, foster, use, address, improve, thoughtful, learn
Merged topic	technology, article, people, image, facial, similar, condition, time, computer, life
	evidence, understand, experience, support, opportunity, ethical, join, author, believe, understanding

SAE-ATD: descriptions of detector-selected features

argument strengthening

Table 30: **Essays comparison.** Parentheses give CRO discovery odds ratios. Selected names link to examples; shared display conventions precede these pages.

Pooled repertoire

CRO: all 57 supported content choices

Human-favored	LLM-favored
a specific individual's actions or traits [E]	collective progress and shared values [N,A,W,E]
quantitative specifics [N,A,W,E]	emphasizing significance [N,A,W,E]
physical objects [A,E]	data collection and its implications [E]
interpersonal interaction [N,A,W,E]	illumination imagery [N,W]
responses to interactions [W]	evocative description of objects or phenomena [N,W]
sleep and wakefulness [W,E]	emotional reactions [N,W]
violence or crime [N,W]	technological advancements [N,A,W]
character-focused observations [E]	exploration of understanding and perspective [N,A,W,E]
noticing or encountering something unusual [E]	limitations of traditional or existing methods [N,A,W,E]
specific details or examples [N,A,W,E]	careful evaluation and planning [N,A,W,E]
animal care and delivery [E]	specific outcomes or implications [N,A,E]
specific events or developments [N,A]	comparison of methods [N]
negation or absence [N,A,W,E]	fear and concern [N,A,W,E]
human versus machine comparison [W,E]	technology's potential or impact [N,A]
cars and their roles [W]	empirical performance validation [N,A,W,E]
financial cost or expense [A,W]	concerns about job displacement [N,E]
anticipation or speculation about the future [W,E]	framing challenges [N,A,E]
intense emotional or physical reaction [W]	personal growth and meaningful impact [N]
mentioning additional or related examples [A,W]	speculative science and technology [W]
explaining FACS technology [E]	research activities [N,W]
posing questions [A,W]	forming or deepening connections [N,E]
human error as the primary cause of traffic accidents [W]	natural geological explanation [E]
identifying limitations or shortcomings [W]	arguing Venus research is worthwhile despite risks [E]
	descriptions of travel destinations [W]
	focused technical investigation [E]
	adapting teaching to student needs [N,E]
	listing examples [N,A,W]
	resolution or satisfaction [N,A,E]
	war and its aftermath [N]
	emotional suffering [W]
	an individual woman's actions or traits [W]
	technical system descriptions [A]

Direction varies by domain: medical context [H: W; L: N]; beliefs or interpretations of others [H: A,E; L: N].

Domains: N, news; A, arXiv; W, writing; E, essays. H/L: human/LLM-favored.

SAE-TM: native ranked words

Output	First 10 ranked words
Feature	lead, result, enable, create, way, address, meet, potentially, effectively, represent
	felt, saw, spend, around, hold, deeply, though, crush, aware, think
	felt, express, woman, feel, around, spend, hold, want, past, speak
Merged topic	author, narrator, face, mar, time, within, lack, also, challenge, despite
	learning, like, use, model, significant, lead, improve, work, enhance, potential

SAE-ATD: descriptions of detector-selected features

amplify argumentation
structured argumentation
counterarguments emphasis
compressed spacing
argumentative repetition
support claims
repetition emphasis
argumentation structure
counterargument emphasis
structured argumentation

Table 31: **Pooled repertoire comparison.** CRO tags indicate supported domains. Selected names link to examples; shared display conventions precede these pages.

The following examples develop distinctions within these inventories. Odds ratios compare corrected within-group membership odds; recognition tests the returned description against unseen same-group nonmembers.

Dates and technological development in news. The human-favored "specific dates or timeframes" choice complements the main paper's quantitative examples. Members include "Fenton Restaurant Week is scheduled to take place from August 23 to August 29." and "Pretrial proceedings are set for September 24, 2019." These points locate different events in time, illustrating an informational choice that recurs across topics. Its discovery and held-out odds ratios are $4.1\times$ and $5.2\times$.

The LLM-favored technological-development choice instead concerns adoption and change: "The upgrades include the adoption of advanced manufacturing technologies." and "Technological advancements, such as AI integration and wearable defibrillators, are expected to shape the future of defibrillation therapy." Its corresponding ratios are $6.3\times$ and $5.3\times$. Both choices distinguish what the populations convey within shared news subjects.

Technical specificity and adaptation in arXiv. The human-favored "specific technical details" choice includes "A special case of bounded interaction is one-way communication, where only Alice sends messages to Bob, who computes the output." This precise restriction on a communication model complements the main paper's mathematical-definition example; its discovery/held-out ratios are $5.3\times$/$6.1\times$. The LLM-favored "dynamic adaptation" choice includes "The robot can dynamically adjust its navigation strategy based on real-time environmental data from onboard sensors." Its ratios are $5.3\times$/$5.0\times$. These distinct technical choices can coexist in one document.

Narrative incidents in creative writing. In creative writing, the human-favored injury choice includes "A sharp object caused a deep cut in their flesh." Its effect recurs in 181 of 263 discovery groups and replicates on held-out groups. This supplies a concrete incident-level finding alongside the main paper's LLM-favored transformation-and-renewal choice; the two are independently learned memberships, not opposing ends of one dimension.

Different arguments within shared essay assignments. Within the essay assignments, a human-favored "limitations of Venus missions" member states, "No spacecraft has successfully landed on Venus for more than a few hours." The LLM-favored climate connection instead includes "Studying Venus's runaway greenhouse effect can provide insights into Earth's climate mechanisms." These choices emphasize operational constraints and relevance to Earth, respectively. Both effects replicate, but recognition covers only 17/40 and 19/40 pairs. The five assignment prompts recur across partitions, so the findings concern content allocation under those assignments, not transfer to new essay topics. Quotations reproduce extracted corpus claims, not independent assessments of Venus missions.

A shared content choice across all four domains. The pooled repertoire supplies shared descriptions with separate domain associations. Table 32 develops "emphasis on collective progress and shared values." Its returned description is: "The points focus on advancing societal, ethical, or communal goals through collaboration, innovation, or resilience." The same description connects community aspirations, fair allocation, a character's ethical commitments, and an essay's framing of innovation. Its LLM-favored effect replicates in every domain, with pooled recognition of 83.3% on 36/40 pairs. These are four supported associations, not one pooled effect.

Dozens of supported content choices in each domain. Each repertoire has 128 candidate choices. An active choice has enough positive discovery activations to define membership; a discovery-supported choice passes the population-difference tests and the criteria for effect consistency across discovery subsets. Table 33 accounts for all discovery-supported choices. Every such choice received a name and a completed recognition evaluation: there are no unnamed, missing, or unevaluated cases in these five repertoires. The four domain repertoires contain 51–69 discovery-supported choices and 32–54 choices passing both held-out checks each. News, for example, has 61 replicated associations and 58 validated descriptions, whose intersection contains 54 choices. Four descriptions pass recognition without a replicated association; seven associations replicate without their descriptions passing recognition. A name alone therefore cannot substitute for population evidence, nor can a population effect validate its name.

Domain	Extracted naming example	LLM-favored odds ratio	
		Discovery	Held-out
News	The event celebrates the values of perseverance, unity, and shared aspirations among participants.	7.0×	6.5×
arXiv	These solutions balance diverse individual preferences with universal fairness principles, fostering inclusivity and equitable outcomes.	5.3×	4.9×
Creative writing	Evangeline explained that kindness goes beyond superficial gestures and involves embracing differences and choosing empathy over judgment.	2.3×	2.6×
Essays	Balancing innovation with ethical principles ensures progress aligns with shared values.	3.2×	4.0×

Table 32: **One shared description with four domain associations.** Examples are exact extracted points used for naming, not original-document quotations or held-out recognition pairs. Every displayed effect replicates; one pooled recognition evaluation supports the shared description. All domains contribute to discovery.

Repertoire	Active	Discovery-supported	Both checks	Population only	Description only	Neither
News	86/128	69	54	7	4	4
arXiv	81/128	55	32	9	11	3
Creative writing	90/128	61	43	9	8	1
Essays	71/128	51	34	6	9	2
Pooled (unique choices)	86/128	81	57	18	5	1

Table 33: **Joint support for discovery-supported content choices.** Active choices are out of 128. Every discovery-supported choice was named and evaluated; the last four columns partition that total. Population support requires same-direction held-out replication; description support requires held-out recognition acceptance. Pooled counts are unique choices, with population support in at least one domain.

Breadth of the pooled repertoire. Learning a repertoire from all four domains gives one set of memberships and descriptions through which their content can be examined together. Centering remains within matched groups, and occupancy tests remain domain-specific. The pooled repertoire has 86 active choices, of which 81 pass discovery in at least one domain and receive names. Those 81 choices account for 199 discovery-supported choice–domain associations: 51 in news, 42 in arXiv, 50 in creative writing, and 56 in essays. Only 14 choices pass discovery in all four domains; the unique-choice total must not be read as four-domain support for every finding.

Seventy-five unique choices have a replicated association in at least one domain, accounting for 158 replicated associations. Sixty-two descriptions pass pooled recognition. The intersection yields 57 supported content choices with 120 supported choice–domain associations. Among the 199 discovery-supported associations, 38 pass replication alone, 29 have a validated description without replication in that domain, and 12 pass neither. These association counts reuse each choice's single cross-domain recognition decision; they are not 199 independent description evaluations. In particular, one choice can be supported in one domain while its effect fails to replicate in another.

Separate evidence for population differences and descriptions. Dictionary discovery, additional naming, and evaluation use separate whole-group partitions of HART's development material; the detection test partition is excluded. The domain fits learn their own memberships, while the pooled fit shares one dictionary across domains. Appendix D.2 specifies the partitions and procedures.

Occupancy is the share of a population's extracted points assigned to a choice within a matched group. The effect in Equation 4 averages corrected log-odds equally over eligible groups, each containing at least three member points across the two populations. We report its magnitude as an odds ratio, $\exp(|\Delta_j|)$, with *Favored by* giving the direction. This is a comparison of membership odds, not document-level mention or authorship probability. *Recurrence* counts eligible groups, not separately significant or uniformly agreeing group effects.

Repertoire	Accuracy	Answered / attempted	Random floor	Positive control
News	80.5%	2,439/2,760	57.5%	95.0%
arXiv	76.0%	2,115/2,200	50.0%	72.5%
Creative writing	82.7%	2,168/2,440	60.0%	95.0%
Essays	81.5%	1,529/2,040	63.0%	82.5%
Pooled (unique choices)	79.9%	2,384/3,240	65.0%	95.0%

Table 34: **Recognition accuracy, response coverage, and controls.** Accuracy averages all named choices equally, excluding unusable answers within each choice. Coverage counts their answered and attempted pairs. The floor is the maximum accuracy of three random-member controls; the positive control uses digit-containing points or their complement.

Discovery support requires statistical tests and checks of effect consistency across discovery subsets; the remaining columns distinguish held-out population replication from description acceptance. Supported content choices pass both replication and description acceptance. The two checks support different claims; they are not successive stages of one statistical test and carry no additional joint false-discovery guarantee. Recognition acceptance requires adequate answered-pair coverage, support above chance under the stated empirical screen, and accuracy above the local random-member floor. Exceeding that observed floor is not a significance test against it. The exact rules appear in Appendices D.2.3 and D.2.5.

Recognition must be read with coverage and controls. Table 34 reports accuracy over all named choices, including those that do not pass either check. Mean choice-level accuracy ranges from 76.0% to 82.7% in the domain repertoires. Response coverage varies substantially: essays have 1,529 usable answers out of 2,040 pairs, compared with 2,115 out of 2,200 in arXiv. One essay choice and three pooled choices have fewer than eight usable answers. Their evaluations are present, but provide insufficient answered pairs for acceptance. Pooled recurrence sums eligible groups across all four domains, including domains without a supported effect; it is not the recurrence of each supported association.

Every positive control has 40 answered pairs. The three random controls answer 40/40/40 pairs in news and creative writing, 40/40/39 in arXiv and pooled, and 27/19/32 in essays. Extra stratified diagnostics are excluded from the primary recognition totals. The random-control floors range from 50.0% to 65.0%, so chance accuracy alone is an inadequate reference for the acceptance decision. Positive controls reach 72.5–95.0% and exceed their corresponding floors; they are checks on the procedure, not ceilings on attainable accuracy. Refusals leave the correctness of the unanswered comparisons unknown. Neither a high answered-pair accuracy nor a low refusal rate alone certifies an appropriate name. Recognition supports the description on the tested member/nonmember pairs; it does not establish exhaustive coverage of every subtype or a uniquely correct abstraction level.

The separate checks make it possible to distinguish a replicated population difference from a supported interpretation of its members. The aggregate results report each stage separately; Appendix E illustrates selected supported findings with source-labeled content points.

D.2 CRO PROCEDURE AND EVALUATION

CRO separates what defines a choice from the evidence for its population association and description. Source-blind learning fixes memberships first. Population tests then identify recurring differences worth examining, while naming and held-out recognition establish what the corresponding members share. This section specifies the procedure behind the findings in Appendix D.1, extending Section 3.2 and the overview in Figure 1.

D.2.1 ANALYSIS UNITS AND INFORMATION SEPARATION

The analysis uses HART's development material: 500 matched groups per domain, each containing the four document conditions in Table 3. Human content comprises t_0, t_1; LLM content comprises t_2, t_3. All four conditions enter the analysis, so each content source is represented under two expression conditions. The document-detection test partition is excluded. Content points are extracted using the fixed procedure in Appendix B.4 and encoded with Microsoft's Harrier OSS v1

Scope	Content points	Matched groups		
		Discovery	Additional naming	Validation
News	31,525	258	78	164
arXiv	41,026	251	84	165
Creative writing	39,525	263	75	162
Essays	28,426	256	86	158
Pooled	140,502	1028	323	649

Table 35: **CRO's analysis population and whole-group partitions.** Point counts cover all three partitions. The pooled analysis concatenates the same four domain populations and preserves their group assignments; it does not introduce additional documents.

Stage	Data	Source labels	Fixed or evaluated output
Learning	Discovery points	Withheld	Dictionary and membership thresholds
Population discovery	Discovery groups	Used for contrasts	Naming eligibility and effect direction
Naming	Discovery and additional naming	Withheld from namer	Name and description
Evaluation	Validation groups	Used for population replication; withheld from recognition judge	Replication and description recognition, assessed separately

Table 36: **Information used at each CRO stage.** The two evaluations share validation groups. Held-out memberships use each group's own centered points with the fixed dictionary and thresholds.

0.6B (`microsoft/harrier-oss-v1-0.6b`), using unit-normalized, 1,024-dimensional embeddings. Extraction and encoding are fixed before repertoire fitting.

Whole matched groups are assigned to discovery, additional naming, and validation partitions with proportions approximately 50%/15%/35%. Table 35 records the realized sizes. All documents and points belonging to a group remain together. Group separation does not require assignment prompts to be distinct across partitions: multiple essay groups, for example, share the corpus's five assignments. Held-out results therefore concern new matched groups, rather than necessarily new writing prompts.

Discovery points determine the dictionary and activation thresholds. Discovery source labels subsequently determine population support and eligibility for naming. Examples for a name come from discovery plus additional naming groups, with source information withheld from the naming model. Validation groups supply both the population-replication test and the separate member-recognition task. These two evaluations support different claims but share a held-out partition; they are not independent datasets or interchangeable tests. Neither validation points nor validation effects are used to select naming examples or order the choices to be named.

D.2.2 Learning and fixing overlapping memberships

For each matched group, we subtract the mean of all its point embeddings as in Equation 2. Each point contributes equally to that mean, so documents with more extracted points contribute more observations. Both populations and all expression conditions enter without consulting their labels. We fit the source-blind ReLU autoencoder in Equation 3 on discovery residuals, using 128 choices and sparsity weight $\lambda = 0.10$. Adam uses learning rate 10^{-3} for 1,500 updates, with batches of 4,096 discovery points sampled with replacement. The decoder has no bias, and its columns are renormalized to unit length after every update. Reconstruction loss is summed across embedding dimensions and sparsity loss across choices, then each is averaged over the batch. The hyperparameter rationale and sensitivity analyses appear in Appendix D.2.7.

A choice is active if at least 20 discovery points have activation greater than 10^{-6}. Its threshold τ_j is the median of those positive activations; otherwise its threshold is infinite and the choice remains inactive. Membership is $a_j \geq \tau_j$. The learned encoder, decoder, and thresholds are then fixed. Memberships can overlap, and a point can belong to no choice, one choice, or several choices. No source label enters fitting or threshold selection.

Additional naming and validation groups are centered using their *own* matched documents before the fixed encoder is applied. The procedure therefore evaluates points relative to their group context: freezing the dictionary and thresholds does not make membership an isolated-document mapping. Centering a held-out group uses neither discovery-group means nor its source labels, but does require the matched content available in that group.

D.2.3 POPULATION OCCUPANCY, DISCOVERY SUPPORT, AND REPLICATION

For group g and choice j, let A_{gj}, B_{gj} count human- and LLM-content members, and C_{gj}, D_{gj} their respective nonmembers. Human occupancy is $A_{gj}/(A_{gj}+C_{gj})$ and LLM occupancy is $B_{gj}/(B_{gj}+D_{gj})$. A group contributes to the contrast when it contains at least three members in total, $A_{gj}+B_{gj} \geq 3$. We use the half-count-corrected log-odds ratio

$$\delta_{gj} = \log \frac{(A_{gj}+\frac{1}{2})(D_{gj}+\frac{1}{2})}{(B_{gj}+\frac{1}{2})(C_{gj}+\frac{1}{2})}. \tag{8}$$

The correction keeps the value finite when a cell is empty. Equation 4 averages these contrasts equally over eligible groups to obtain Δ_j; their count is recurrence R_j. Recurrence measures the breadth of eligible contexts, not the number of individually significant group effects.

The displayed odds ratio is $\exp(|\Delta_j|)$, oriented toward the population with greater occupancy. Before that orientation, $\exp(\Delta_j)$ is the geometric mean of the corrected human-to-LLM odds ratios across eligible groups. It is neither a pooled-table odds ratio nor an inclusion-probability ratio. Equal weighting across groups also does not equalize documents within a group: a document contributes all its extracted points. Because occupancy concerns their relative allocation, changing nonmember content can change occupancy even with the same number of members. The contrast describes a population association in this representation, rather than a deterministic authorship rule or a causal author preference.

Document-level permutation tests. We permute content-source labels among documents within each matched group, moving all points from a document together. This preserves document sizes and avoids treating a document's points as independent permutation units. The dictionary, thresholds, and eligible-group sets remain fixed. For $B = 8{,}000$ permutations, let $\Delta_j^{(b)}$ be the resulting contrast and μ_j^{perm} their mean. The implemented two-sided Monte Carlo value is

$$p_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\left\{|\Delta_j^{(b)} - \mu_j^{\text{perm}}| \geq |\Delta_j - \mu_j^{\text{perm}}|\right\}}{B+1}.$$

We apply Benjamini–Hochberg correction at $q = 0.05$ to the defined choice-level values within each repertoire and domain. A choice with no eligible group has no test value and cannot pass. The correction is performed separately on discovery and validation groups; it is not a joint correction across all repertoires, domains, and description judgments.

Effect consistency across discovery subsets and held-out replication. Discovery groups are also divided into two fixed subsets. Each subset must contain eligible groups, and their contrasts must share the full discovery direction while each retaining at least half its absolute magnitude. A choice is discovery-supported only when it passes both this effect-consistency condition and discovery BH-FDR. This is the set eligible for naming.

On validation groups, we repeat the occupancy calculation and permutation test with the fixed memberships and fresh random draws. Validation BH correction includes all choices with defined validation test values, not only the discovery-supported subset. A choice replicates when it is discovery-supported, passes validation BH-FDR, and has the same contrast direction on both partitions. Alternative occupancy summaries, bootstrap uncertainty, expression slices, generator diagnostics, and human-only null comparisons are reported as diagnostic evidence; they do not substitute for these acceptance criteria.

Permutation resolution across dictionary widths. The fitting, thresholding, naming, and recognition settings are specified at their procedures in Appendix D.2; contrast and baseline departures appear beside their respective comparisons. One further numerical constraint matters when comparing dictionary widths. With $B = 8{,}000$ permutations, the smallest Monte Carlo value is $f = 1/(B+1)$.

Repertoire / domain	BH rejections / tests	Values below 0.05
News	0/86	2.33%
arXiv	0/81	7.41%
Creative writing	0/90	3.33%
Essays	0/71	8.45%
Pooled: Essays	0/85	5.88%
Pooled: Creative writing	1/86	4.65%
Pooled: arXiv	0/83	8.43%
Pooled: News	0/86	4.65%

Table 37: **Human-only null comparisons for CRO.** Each row uses its own defined permutation tests and BH correction at 0.05. Pooled rows use the shared dictionary within the indicated domain. Under a valid null, about 5% of values are expected below 0.05.

For n defined tests, a necessary BH rank for rejection at level q is

$$r_{\min} = \left\lceil \frac{fn}{q} \right\rceil .$$

If at least $r_{\min}$ values attain the floor, those values can satisfy the rule. More generally, rejected values need not all equal the floor; this expression states the resolution limit, not a requirement that every discovery attain it.

At $q = 0.05$, every CRO repertoire and pooled domain has $r_{\min} = 1$. For the primary wide SAE-TM fits, the actual numbers of defined discovery tests are 4,096 in news, 4,094 in arXiv, 4,087 in creative writing, and 3,358 in essays, giving minimum ranks 11, 11, 11, and 9. The shared pooled dictionary has 1,438–3,268 defined tests per domain, giving minimum ranks 4–9. These counts use defined test values, as BH does; nominal dictionary width or active capacity can overstate the relevant multiplicity when a choice has no eligible groups in a particular domain. This resolution check explains a constraint on discovery without replacing the measured recurrence, replication, and interpretation evidence.

Human-only null comparisons. With memberships fixed, we restrict the population comparison to human-content documents (t_0, t_1) and assign deterministic pseudorandom binary labels at the document level. All points in a document share its pseudo-label; the assignment does not force balance within each matched group. We then repeat the occupancy and within-group document-label permutation procedure. This checks the population-comparison machinery under a constructed null; it does not refit the dictionary or validate its semantic descriptions.

Table 37 reports the outcomes for CRO. The four separate domain fits have no BH rejections. The pooled repertoire has one rejection in creative writing and none in the other three domains. Under a valid null, about 5% of test values would fall below 0.05; CRO's eight comparisons range from 2.33% to 8.45%. Permutation values within a repertoire are discrete and not independent, so this is a reference level rather than a formal calibration test.

Across the four learning configurations in Appendix D.5, the largest small-value fraction among the 32 domain-level null comparisons occurs for the pooled uncentered supervised dictionary in news: 5 of 27 values below 0.05 (18.52%), although none passes BH correction. Its other three pooled domains have no BH rejections and small-value fractions between 0 and 11.11%. The constructed null alone does not establish whether this excess comes from the supervised representation or chance variation. Isolated single BH rejections also occur in other configurations.

D.2.4 Naming the shared choice

We attempt to name every discovery-supported CRO choice, ordered by the absolute discovery contrast. For each choice, candidates are members from discovery and additional naming groups. We traverse them in descending activation order, retaining at most one point per group and up to 24 points overall. We reject a candidate whose normalized word set has Jaccard similarity at least 0.8 with an already retained point; a rejected candidate does not consume its group's slot. A lower-ranked member from that group can still be selected. A choice with fewer than six retained examples is recorded as unnamed. All discovery-supported choices in the five reported CRO repertoires supplied the required examples.

The model also sees up to 14 below-threshold points from groups represented among the members, ordered from lowest activation and subject to a per-group cap and exact-text deduplication. The cap is $\lceil 14/G \rceil$, where G is the number of represented groups, with a minimum of one; groups are visited in selected-member order. These nonmembers supply context for distinguishing the choice from other content about the same subject. Individual examples are chosen using activation, group membership, and text-duplication checks, without using their content-source labels.

The naming model receives only the point excerpts and anonymous integer group tags. It sees no source labels, generator identities, effect sizes, directions, or population counts. Statistical evidence determines *which choices* receive names; it is withheld when the model describes *what their points share*. Both member and nonmember texts are truncated to 300 characters. The prompt requests up to two candidates, each with a name, one-sentence description, diagnostic cue, and coherence rating. The first candidate is retained; its name and description are fixed for recognition. The cue and self-rated coherence do not determine description acceptance.

D.2.5 Recognition of unseen members and description acceptance

Pairs from the held-out groups. Each named choice is evaluated on up to 40 pairs, with a member and nonmember from the same validation group. Members satisfy $a_j \geq \tau_j$. Nonmembers normally have $a_j \leq 10^{-6}$; if a group has too few of these to pair with its candidate members, its nonmember list is extended with remaining below-threshold points in increasing activation order.

Within each group, candidate members and zero-activation nonmembers are shuffled. Groups contribute pairs in turn, with identical-text pairs skipped, until the requested number is reached or candidates are exhausted. Presentation order is randomized deterministically for each pair. The judge receives the fixed name and description and the first 300 characters of each point, but no activations, source labels, or naming examples. Matching the group makes shared subject matter alone less informative. This task assesses recognition against the specified mostly zero-activation comparison points; it does not test exact recovery of the membership threshold or exhaustive coverage of every subtype.

We report accuracy over usable answers and record refusals or unusable responses separately, without forcing a guess. All reported CRO choices have 40 primary attempted pairs. Some additionally have lexical-stratification diagnostic pairs; these do not enter the primary acceptance decision. Excluding unanswered pairs leaves their correctness unknown, so accuracy and response coverage are reported together in Table 34.

Random-member and positive controls. For each repertoire and scope, three random member sets match a discovery-supported template choice's member count in every group. The template is the middle entry in the sorted list of discovery-supported choice IDs. Random memberships receive small random activation jitter to order exemplars and pass through the same naming and recognition procedure. The random-control floor f is the largest accuracy among these three controls with at least eight usable answers. It is measured locally for the evaluated procedure, rather than transferred from another repertoire or naming variant.

A positive control uses points containing a digit, or their complement when digit-containing points are the majority. It also receives a generated description and held-out recognition test. Its accuracy must exceed the measured random-control floor for description-support claims to be reported. When this check fails, descriptions and diagnostics remain visible but description-support claims are suppressed. The positive control clears the floor in all five reference CRO repertoires. This is a check on the evaluation procedure, not a ceiling on achievable accuracy.

Acceptance rule and its interpretation. Let n be the number of usable primary answers, $\widehat{a}$ their accuracy, and s_{bin} and s_{grp} the binomial and group-bootstrap standard errors. The latter uses 500 resamples of the contributing matched groups, retaining all answered pairs from each sampled group. With the controls permitting interpretation, a description passes when

$$n \geq 8, \qquad \widehat{a} \geq 0.65, \qquad \widehat{a} - 2\max(s_{\text{bin}}, s_{\text{grp}}) > 0.5, \qquad \widehat{a} > f.$$

The rule uses the recorded accuracies and standard errors, rounded to three decimals. An observed zero standard error remains zero, including at perfect observed recognition; it is not treated as missing. This is the specified empirical acceptance screen, not an exact confidence interval or a guarantee of perfect generalization. Recognition supports the description's connection to unseen members,

without selecting a uniquely correct name or abstraction level. A supported finding additionally requires the separate population association to replicate. Naming and judging use the same model snapshot in separate calls; validation comes from held-out examples and measured controls, not an independent model family.

D.2.6 ONE POOLED REPERTOIRE WITH DOMAIN-SPECIFIC EVIDENCE

The pooled analysis fits a single dictionary to discovery residuals concatenated across all four domains, keeping centering within each matched group. Points enter the fitting pool directly, without reweighting domains to equal point counts. The dictionary and discovery-derived thresholds are shared, while occupancy contrasts, permutation tests, BH corrections, discovery-subset effect checks, and replication decisions are computed separately within each domain.

A pooled choice is eligible for naming if discovery-supported in at least one domain. Its ordering statistic is the largest absolute discovery contrast among those discovery-supported domains; this statistic does not define a universal pooled population effect. Direction disagreements across discovery-supported domains are retained. Naming draws from the combined discovery/additional-naming pool using the same activation ranking and one-example-per-group limit, without a fixed quota by domain. Recognition likewise draws same-group pairs from the pooled validation groups, producing one description-level decision rather than four domain-specific description decisions.

This distinction determines how the pooled findings in Appendix D.1 are counted. A unique choice is supported when its pooled description passes recognition and its population association replicates in at least one domain. A particular choice–domain association is supported only when that domain's effect replicates and the pooled description passes. Neither status implies support in every domain, and a replicated effect elsewhere does not validate a failed displayed association. Pooled recurrence sums eligible groups across domains, including those without a supported association. The shared name organizes cross-domain content; each population claim retains its own domain evidence.

D.2.7 HYPERPARAMETER RATIONALE AND SENSITIVITY

A useful repertoire must preserve overlapping choices while remaining selective enough to compare and describe. Reconstruction quality alone does not establish that balance. We assess capacity and sparsity using discovery membership density, active capacity, and coherence beyond shared subject matter, without using source labels, occupancy effects, generated names, or recognition scores in the criterion.

Eligibility and coherence. An eligible configuration assigns a median of 2–16 choices per discovery point and retains at least half its dictionary as active under the 20-positive-activation rule. The density is measured *after membership thresholding*, rather than by counting every positive ReLU output. The lower bound permits overlap; the upper bound limits diffuse assignments. This is an operating range for a typical point, not a claim that every sentence intrinsically has 2–16 aspects.

For each choice with at least 12 discovery activations above 10^{-6}, we take its 12 strongest points and compare their semantic coherence with a reference that draws one discovery point from each selected point's matched group. Using the original unit-normalized embeddings, define

$$c(S) = \frac{1}{|S|(|S|-1)} \sum_{u \neq v} h_u^\top h_v, \qquad \gamma_j = c(S_j) - c(B_j).$$

The configuration score is the mean defined γ_j. Its 12-point requirement is distinct from the 20-point active-choice rule. Matching the reference's group composition asks whether the selected points share structure beyond their subjects; measuring it in the original embedding space avoids evaluating only residual-space alignment. Among eligible configurations, the reporting rule prefers the largest mean gap, breaking ties by signed cross-initialization agreement at cosine 0.5. Reconstruction R^2 is reported, not used as an eligibility condition or optimization target.

A common width, supported by the earlier pilot. The earlier centered pilot evaluated the complete grid $K \in \{64, 128, 256, 512\}$ and $\lambda \in \{0.03, 0.10, 0.30\}$ in all four domains. At $\lambda = 0.10$, both widths 64 and 128 are eligible everywhere; widths 256 and 512 fail the active-capacity requirement in essays. Among eligible grid cells, the highest coherence gaps occur at width 128 in essays

	Mean coherence gap at $\lambda = 0.10$			
Domain	$K = 64$	$K = 128$	$K = 256$	$K = 512$
News	0.1173	0.1154	0.1235	0.1149
arXiv	0.1348	0.1309	0.1189	0.1147
Creative writing	0.1289	0.1390	0.1340	0.1321
Essays	0.1663	0.1727	0.1726†	0.1620†

Table 38: **Width sensitivity in the earlier centered pilot.** The displayed sparsity setting contains each domain's highest-scoring eligible cell in the full width–sparsity grid. A dagger marks an ineligible cell: fewer than half the directions are active. The separate refitting study in Appendix D.7.3 examines correspondence across fitting seeds at the retained operating point.

Domain	λ	Active / 128	Memberships	Coherence gap	R^2	Eligible
News	0.03	128	25	0.0907	0.359	No
	0.10	86	4	0.1154	0.164	Yes
	0.30	0	0	–	0.000	No
arXiv	0.03	128	25	0.1006	0.440	No
	0.10	81	5	0.1309	0.228	Yes
	0.30	2	0	0.1744	0.014	No
Creative writing	0.03	128	26	0.1115	0.484	No
	0.10	90	6	0.1390	0.264	Yes
	0.30	0	0	–	0.000	No
Essays	0.03	127	19	0.1370	0.578	No
	0.10	71	4	0.1727	0.348	Yes
	0.30	1	0	0.1259	0.004	No

Table 39: **Sparsity sensitivity at the fixed width of 128.** Memberships is the median thresholded count per discovery point; coherence uses discovery points and group-matched references. Reconstruction R^2 uses additional-naming and validation points together and does not enter the reporting rule. A dash denotes an undefined coherence gap when no choice has enough positive activations. This retrospective check reports agreement with the already fixed setting; it is not prospective selection on the held-out population results.

and creative writing, 64 in arXiv, and 256 in news, all at $\lambda = 0.10$ (Table 38). The evidence therefore supports retaining 128 as a common eligible capacity that accommodates more choices than 64, rather than claiming one universally optimal width. The same width is retained for the pooled repertoire and the learning ablations.

Retrospective sparsity check of the fixed operating point. The later experiment held width 128 and the production setting $\lambda = 0.10$ fixed, and evaluated the sparsity rule as a diagnostic; it did not let that analysis replace the setting. Table 39 shows why the common value is useful. In every domain, 0.10 is the unique eligible value among $\{0.03, 0.10, 0.30\}$ and yields 4–6 memberships per typical point. At 0.03, reconstruction improves but membership becomes too dense (19–26 choices) and coherence gaps are smaller. At 0.30, only 0–2 choices remain active. The fixed operating point therefore satisfies the intended representation criteria across all four domains without choosing it to maximize population effects or supported-choice counts.

On uncentered inputs, no tested $\lambda \in \{0.03, 0.05, 0.10\}$ meets both eligibility conditions in any domain. At 0.03, median membership is 12–14 but the inactive fraction is 0.664–0.688; at 0.05 these ranges are 7–8 and 0.797–0.820. At 0.10 only 1–3 choices remain active. The uncentered comparison retains 0.05 to obtain usable overlapping membership at a comparable scale, with its loss of active capacity reported explicitly. It is not the output of the full eligibility rule. The supervision-weight selection and its distinct use of nontraining groups are described with that comparison in Appendix D.5.

D.2.8 Exact naming and recognition messages

Exact naming prompt. Naming uses GPT-4o (`gpt-4o-2024-11-20`), temperature zero, top-p one, and at most 420 response tokens. The two messages are:

System message.

```
You are a careful analyst. You name what a group of content points has in
↪  common: the kind of angle, move, or content they share, not their
↪ surface wording.
```

User message.

```
The points below were extracted from documents written in response to
↪ several different writing prompts. An automatic procedure placed some
↪ of them in one group because they occupy the same region of a content-
↪ embedding space. Your job is to name what the group shares.

You are shown two blocks. IN-GROUP points are the group's members. OUT-
↪ GROUP points come from documents responding to the same prompts but
↪ were placed outside the group. The out-group points are there to show
↪ you what the group is NOT: the name has to separate the two blocks,
↪ not merely fit the in-group.

Follow these rules:
- Name the kind of angle, move, or content the in-group points have in
↪ common, not their surface wording.
- The name must complete the sentence "The angle these points share is
↪ ___": a short phrase, not a full sentence.
- Give each name a gloss: a one-sentence description a judge could
↪ recognize members by.
- A name that would also fit most of the out-group points is too broad.
↪ If your name does that, narrow it until it separates the blocks.
- Points can share what they say or what they are doing in the document;
↪ if the true common thread is what they are doing, name that move
↪ directly rather than reaching for a subject-matter name.
- If the name needs a trailing "of ..." to feel complete, or strings two
↪ or three qualifiers together with "and" or "or", it has drifted from
↪ naming into describing; move that nuance into the gloss and tighten
↪ the name to its core.
- Return up to 2 candidates ranked by fit. If the in-group looks
↪ incoherent, set coherent low and still give your best description.
- Add a cue: the single thing a judge should look for to decide whether
↪ an unseen point is a member. State it as a test, not as a restatement
↪ of the name.

{point_blocks}

Answer with JSON only:
{"candidates": [{"name": "the words that complete: The angle these points
↪  share is ___", "gloss": "one-sentence description a judge could
↪ recognize members by", "cue": "the test a judge should apply", "
↪ coherent": 1-5}, ... up to 2 ...]}
```

Point-block format. Replace {point_blocks} with the following structure, repeating a point line for each selected example. Tags are assigned by first occurrence across the member block followed by the nonmember block, so the same group retains its tag in both. Omit the out-group heading and block when no nonmembers are available.

```
IN-GROUP POINTS:
- (prompt {anonymous_group_tag}) {member_point}

OUT-GROUP POINTS (same prompts, placed outside the group):
- (prompt {anonymous_group_tag}) {nonmember_point}
```

Exact recognition prompt. The judge uses the same model, temperature, and top-p as naming, with a 20-token response limit. The diagnostic cue and coherence rating are not supplied. Braced

Scope	CRO choices	Median best IoU	Above random p95
News	69	0.103	69/69
arXiv	55	0.168	53/55
Creative writing	61	0.123	60/61
Essays	51	0.268	51/51
Pooled	81	0.217	81/81

Table 40: **SAE-TM has related memberships across the CRO repertoire.** Every discovery-supported CRO choice is compared with the full primary SAE-TM dictionary in its scope. The last column gives choices exceeding their size-matched random reference over all compared choices. The pooled row uses the shared dictionary and all pooled points.

fields below are replaced with the retained name and description and the two point excerpts in their randomized order.

System message.

```
You decide which of two content points belongs to a named group. Answer
↪ with JSON only.
```

User message.

```
Group: "{name}"
Description: {gloss}

Point 1: {point_1}
Point 2: {point_2}

Which point belongs to the group? Answer JSON only: {"answer": 1 or 2}
```

D.3 SAE-TM: CONTENT OVERLAP AND NATIVE INTERPRETATIONS

SAE-TM recovers content regions related to CRO's choices, and its discovery improves when given CRO's within-group centering. These results clarify what the full CRO procedure contributes: recurring memberships are expressed as specific descriptions, tested on unseen members, and connected to separately replicated population differences. We first compare the native interpretations over measured overlapping memberships, then examine how dictionary configuration and centering affect the added occupancy analysis.

Overlapping memberships across domains. For each discovery-supported CRO choice, we find the largest membership intersection-over-union (IoU) with any of the primary SAE-TM configuration's 4,096 features. Memberships use each feature's median positive discovery activation as a fixed threshold. The comparison spans all points in the scope, including discovery, naming, and validation groups; for news this is 31,525 points. It is a descriptive correspondence analysis, separate from held-out population replication and description recognition.

The broad dictionary makes a random reference essential. For each CRO choice, we draw 64 random point sets with the same membership count, hold the SAE-TM memberships fixed, and record the largest IoU across the full dictionary for every draw. Table 40 compares observed maxima with each choice's 95th-percentile reference. All 69 news choices exceed this reference, as do 53 of 55 arXiv choices, 60 of 61 creative-writing choices, and all 51 essay choices. Typical matches remain partial: median best IoU ranges from 0.103 in news to 0.268 in essays. Thus, related regions are widespread, while individual features generally do not define identical sets of points. The random reference accounts for searching a wide dictionary; these exceedances are descriptive, not multiplicity-adjusted significance tests.

What the two procedures say about related points. Table 41 expands the main-paper comparison with informative native word lists. The numerical-detail pair has an unusually high IoU of 0.545, compared with the news median of 0.103. The sustainability pair shares a recognizable environmental theme at IoU 0.251. The dates/timeframes pair is closer to the typical overlap at 0.107:

CRO (Ours): choice and description	SAE-TM: native feature words	IoU (null p95)
specific numerical details. "The points share a focus on providing precise numerical data, such as quantities, percentages, rankings, or measurements."	three, third, year, quarter, record, million, approximately, previous, total, least, iii, percent	0.545 (0.020)
environmental sustainability. "The points focus on actions, practices, or initiatives aimed at reducing environmental impact or promoting ecological balance."	sustainable, sustainability, environmental, environmentally, green, conscious, carbon, adopt, renewable, friendly, promote, transportation	0.251 (0.020)
specific dates or timeframes. "The points focus on specific calendar dates or time periods for events, actions, or milestones."	july, summer, june, coincide, bear, old, year, producer, actor, best, influential, schedule	0.107 (0.020)

Table 41: **Explicit choice descriptions beside measured SAE-TM counterparts.** Selected news examples; linked names lead to exact CRO records. Each SAE-TM feature maximizes IoU for that CRO choice; selection does not require the SAE-TM feature to pass the added occupancy tests. Descriptions are returned CRO outputs; word lists retain the first twelve words in native rank order.

its words include months and scheduling language, while CRO explicitly describes the recurring informational choice. These examples connect a strong correspondence, a shared subject, and a more typical partial match to findings already developed in the paper.

These native words identify related content; CRO supplies a description of what its own members share. All three displayed CRO findings pass both held-out checks. The numerical-detail description, for example, achieves 89.5% recognition on 38/40 pairs, and its human-favored odds ratio rises from $3.1\times$ in discovery to $5.9\times$ held out. These results concern CRO's memberships and are not transferred to the matched SAE-TM feature.

Merged topics summarize a different level. The domain pages show native merged topics separately from individual features. A topic combines feature word distributions; it does not inherit an individual feature's membership IoU or population effect. The displayed topics are selected by native mass, without imposing CRO descriptions on them.

Centering improves SAE-TM's population discovery. To examine whether the learned features reveal recurring source differences, we separately apply CRO's occupancy analysis to the SAE-TM activations. This is an added diagnostic, distinct from SAE-TM's native word and topic outputs. It uses the same matched groups, source contrasts, discovery/validation partitions, document-level permutations within groups, and support rules as CRO. A discovery-supported feature passes discovery BH-FDR at $q = 0.05$ and the discovery-subset effect check (the same direction and at least half the full effect magnitude in each half); replication additionally requires held-out FDR support in the same direction. No CRO naming or member-recognition evaluation is applied to SAE-TM's word lists.

Table 42 reports news using the shared median-positive membership rule. At fixed width 128 and $k = 4$, centering raises discovery-supported features from 34 to 75 and replicated features from 20 to 47. The improvement extends across domains: discovery-supported/replicated counts rise from 33/17 to 58/40 in arXiv, 57/28 to 74/56 in creative writing, and 70/63 to 88/76 in essays. In the pooled repertoire, the corresponding unique-feature counts rise from 79/59 to 102/84; discovery support and replication there mean a qualifying association in at least one domain. Centering is therefore a transferable part of population discovery, with benefits beyond CRO's particular autoencoder.

Recurrence and membership rules explain the count comparison. The primary 4,096-feature dictionary yields 787 discovery-supported news features, with median recurrence of seven discovery groups. Its larger capacity and less recurrent features make this raw total different from an equal-width inventory comparison. The 8,000-permutation procedure has sufficient resolution for these discoveries: its minimum BH-discoverable set is 11 among 4,096 active features. At width 128, centered SAE-TM yields 75 discovery-supported features versus CRO's 69, while their median recurrence is 34 versus 107 groups. CRO's findings thus recur across substantially more matched

Configuration	Dictionary	Active	Discovery-supported	Replicated	Recurrence
SAE-TM, primary	4,096	4,096	787	95	7/258
SAE-TM, matched width	128	128	34	20	25/258
SAE-TM, centered	128	128	75	47	34/258
CRO (Ours)	128	86	69	61	107/258

Table 42: **Dictionary configuration, discovery, and recurrence in news.** Added occupancy diagnostics under median-positive membership. Active features have at least 20 positive discovery activations. Discovery-supported counts are out of active features; replicated counts are out of discovery-supported features. Recurrence is the median eligible-group count among discovery-supported features, out of 258 discovery groups; a group is eligible when it contains at least three member points across both populations. The primary and matched-width SAE-TM inputs are uncentered; the other two configurations use within-group centering.

contexts under this common threshold convention. The threshold does not equalize membership density: median memberships per discovery point are two for each narrow SAE-TM configuration and four for CRO.

The alternate positive-activation membership rule further separates the centering result from a method ranking. Under that rule, the primary SAE-TM configuration has 1,065 discovery-supported and 530 replicated news features. At width 128, uncentered SAE-TM has 53/41 discovery-supported/replicated features and centered SAE-TM has 80/66. Centering improves both counts under either rule, but the absolute counts and their ordering relative to CRO depend on how membership is defined. Accordingly, the contribution supported by this comparison is the full CRO procedure: descriptions identify recurring content choices, recognition checks their meaning, and held-out occupancy establishes their population association. Whether another dictionary combined with CRO's naming and validation would provide the same benefits remains outside this comparison.

Adapting the native pipeline to content points. We apply the BatchTopK, word-interpretation, and topic-merging stages of Girrbach & Akata (2026) to the same content-point corpus used by CRO. The primary configuration uses uncentered, unit-normalized 1,024-dimensional point embeddings and 4,096 features with $k = 32$. Two additional configurations use 128 features with $k = 4$: one retains uncentered inputs, and the other subtracts the matched-group mean as in CRO. Each configuration is fitted separately in the four domains and on their pooled corpus, using the same discovery-group partition as CRO.

The adaptation changes more than dictionary width. Relative to the authors' released 65,536-feature, 800,000-step training recipe, we use 4,096 features and 6,000 steps, and supply float32 embeddings rather than the cached loader's float8 inputs. The original dictionary would exceed the available discovery-point count in an individual domain. Training retains the authors' BatchTopK trainer, auxiliary loss for inactive features, learning-rate scaling rule, batch size 4,096, and 10% warmup. Its moving-average activation threshold has coefficient 0.999 and starts updating at step 1,000; the longer duration allows it to develop beyond CRO's 1,500-step fitting schedule. Fitting diagnostics are also retained at 1,500 steps.

For native interpretation, each extracted point is a document. The word stage fits inverse-document-frequency-weighted bag-of-words likelihoods to feature activations, using lemmatization, a vocabulary cap of 5,000, minimum document frequency five, and 50 epochs. It emits ranked words for each feature. The merge stage retains features with positive mean topic mass, truncates each word distribution at cumulative mass 0.9, embeds it with 300-dimensional Google News word2vec vectors, and applies weighted k-means. We produce 20, 50, and 100 topics for the primary configuration, and 20 for each narrower configuration. Word fitting and topic merging use the points across all partitions in the scope; the resulting outputs describe the observed corpus.

D.4 SAE-ATD: DETECTOR-SELECTED INTERPRETATIONS AND OUR ADDITIONAL ANALYSES

SAE-ATD achieves strong source discrimination and interprets detector-selected features through changes in generated text under steering. The five overview pages present these descriptions from our evaluated implementation. We additionally compare them with descriptions of randomly sam-

	Held-out AUROC		Selected features
Selection scope	Sum	Per-token	interpreted
News	0.9999	1.0000	10
arXiv	1.0000	0.9970	10
Creative writing	0.9999	0.9999	10
Essays	1.0000	0.9968	1
Pooled	0.9999	0.9999	10
News (content-source)	0.9838	0.9831	10

Table 43: **Strong source discrimination across the evaluated settings.** AUROC belongs to the full fitted detector. Feature counts specify the detector-selected interpretations; the essay detector selects only one feature.

Detector-selected features	**Randomly sampled active features**
Argumentation-related: 8 of 10	Argumentation-related: 9 of 10
counterarguments emphasis	counterargument emphasis
multilingual insertion	counterargument emphasis
rebuttal emphasis	structured argumentation
emphasize argument support	counterargument inclusion
argumentative structure	emphasize supporting evidence
argumentative structure	emphasize argument structure
argument structure	structured argumentation
run-on-technical	argument structure
scientific argumentation	emphasize argument structure
counterargument emphasis	punctuation pattern

Table 44: **Our additional random-feature control for SAE-ATD in news.** All ten names per set under the original instruction (ST0), using the same steering and interpretation procedure. Underscores are displayed as spaces; repetitions are retained. Counts use the lexical rule below and measure recurring vocabulary, not interpretation accuracy. Selected names link to steering examples; neighboring entries are not matched feature pairs.

pled active features. This control is introduced in our study, distinct from the detector-selected outputs themselves.

Detector task and feature selection. We adapt the detector and steering-based interpretation route of Kuznetsov et al. (2025) to HART, fitting detectors separately in each domain and across the pooled corpus.

Table 43 shows strong source discrimination, which persists under per-token scaling. The essay fit uses just one positive-gain feature, so its interpretation set contains one feature rather than ten. These detector results establish that the representations contain substantial source information; AUROC is not an evaluation of the resulting descriptions, nor is it the performance of a classifier refitted on only the displayed top features.

Our additional control for SAE-ATD. We perform an additional control to assess whether the returned descriptions specifically characterize detector-selected features. For each scope, we sample ten features without replacement from those active in more than 10% of its documents, without using detector importance. These are existing learned features that may have meaningful functions or carry source information. We apply the same steering and interpretation procedure to selected and sampled features under the common essay probe.

Table 44 gives all twenty original-instruction names for news. The highest-gain selected feature receives "counterarguments emphasis," while two random features receive "counterargument emphasis." Supporting-evidence themes also occur in both columns. Exceptions such as "multilingual insertion" remain visible, so the result concerns recurring themes rather than identical outputs or an absence of differentiated features.

	Original instruction (ST0)		No category examples (ST1)	
Selection scope	Selected	Random	Selected	Random
News	8/10	9/10	8/10	7/10
arXiv	8/10	10/10	7/10	9/10
Creative writing	8/10	9/10	9/10	10/10
Essays	1/1	10/10	1/1	10/10
Pooled	8/10	10/10	7/10	10/10
News (content-source)	6/10	9/10	6/10	9/10

Table 45: **Our additional controls across domains and interpretation prompts.** Each fraction is over all interpreted features in that set. ST1 removes ST0's illustrative category instruction. Counts use the same argumentation-family rule under both prompts. All rows use the same essay-based interpretation probe.

Our control across domains and interpretation prompts. We summarize the repeated vocabulary with an explicit lexical rule: a name belongs to the argumentation family if it contains any of `argument`, `claim`, `rebuttal`, `support`, `evidence`, or `conclusion`, ignoring case. The same rule applies to selected and random features and to both interpretation prompts. It measures the prevalence of these names, not interpretation accuracy or an exhaustive semantic classification of the descriptions.

Under the original prompt (ST0), eight of ten selected names match this family in news, arXiv, creative writing, and the pooled analysis, compared with nine or ten of ten random names (Table 45). In essays the corresponding counts are one of one selected and ten of ten random. Recurring argumentation vocabulary therefore supplies little evidence that the selected features, specifically, explain the source distinction under this interpretation procedure.

Removing category examples does not remove the shared themes. ST1 replaces the illustrative list of answer types with a category-neutral instruction, while retaining the same generation blocks and output schema. Argument-family names remain common in both selected and random sets. The exact counts can change: the news random set falls from nine to seven matches, while the selected set remains at eight. Thus the control is informative without requiring the random proportion to exceed the selected proportion under every wording. Some names also become more tied to the probe, such as `emphasize_natural_landform` for pooled selected feature 8689. Selected-feature kind agreement ranges from eight to ten of ten, with one of one in essays, but unchanged categories do not imply identical descriptions. This prompt comparison does not isolate illustrative category priming as the cause of the repeated vocabulary; the common essay assignment remains in both conditions.

Detector implementation. The domain and pooled fits compare t_0 with t_3; the additional news fit, labeled "News (content-source)" in the tables, compares (t_0, t_1) with (t_2, t_3). Documents are encoded with Gemma-2-2B and pretrained Gemma Scope 16,384-feature JumpReLU residual-stream SAEs. Analysis uses layer 16 and the same `average_l0_78` release for extraction and steering. Feature activations are summed over token positions, excluding the beginning-of-sequence token, with documents truncated to 1,024 tokens. We fit XGBoost with maximum depth four, $\alpha = 1$, and log-loss evaluation on discovery groups, then evaluate on CRO's held-out development groups, separately from the document-detection test partition. A second fit divides activation sums by document token count to check length sensitivity. Steering selection uses the ten largest booster-gain features from the sum-based fit, or all positive-gain features when fewer are available.

Steering procedure used for selected and control features. For feature j, the intervention adds its decoder direction at every token position of the layer-16 residual stream:

$$h_{\text{out}}^{(16)} \leftarrow h_{\text{out}}^{(16)} + m_j \rho D_j, \qquad \rho \in \{-4, -2, -1, 1, 2, 4\}.$$

Here m_j is the maximum feature activation observed over 300 documents from the relevant scope, using up to 512 tokens per document. Each condition generates up to 95 new tokens at temperature 0.7 and top-p 0.9, with an unsteered continuation for comparison.

The generation grid contains eight probe entries, two from each HART domain, at the six signed strengths plus the unsteered condition. The interpretation stage has narrower coverage: its input

contains only the first two entries, both repetitions of the same essay assignment about the Face on Mars. Each entry supplies the first 200 characters of its prompt and the first 420 characters of each continuation. Thus, cross-domain rows compare different *feature-selection populations under a common interpretation probe*; they do not demonstrate that the returned themes persist across probe domains. The complete prompt and this assignment appear below; Appendix E.4 illustrates the saved generation evidence for a selected feature and a random control. This distinction matters because the assignment itself requests an argument, evidence, and claims.

Steering strength constrains the interpretations. For selected news features, m_j ranges from 11.070 to 78.764; every selected and random feature has positive m_j, so no reported intervention is zero.

Under ST0, three of ten selected news features have reported breakdown magnitude two and seven have magnitude four; the random set has one report at magnitude one and nine at four. Differentiated descriptions also carry these diagnostics: multilingual insertion is reported to break down at magnitude four, while selected arXiv feature 8772, named `html_markup_insertion`, is reported at magnitude one. The interpreter is instructed to use low and medium strengths, but a report of breakdown at one leaves no smaller nonzero tested magnitude as a comparison. These are judgments from the supplied continuation excerpts, not independently verified safety boundaries for steering.

Local interpretation prompts and evidence template. Both variants use GPT-4o (`gpt-4o-2024-11-20`) with temperature zero, top-p one, and a 500-token response limit. These are our prompts for the evaluated steering route; selected and random features use identical instructions. They ask about changes in generated continuations, not membership in the original corpus.

System message, shared by ST0 and ST1.

```
You interpret a sparse-autoencoder feature from a steering experiment.
↪ You are shown text generated normally and text generated with the
↪ feature's value added to the model's residual stream at every token.
↪ You say what amplifying the feature does, whatever kind of thing that
↪ turns out to be.
```

ST0 user message.

```
Below, for each probe prompt, is text the model generated with no
↪ intervention (`original`) and text generated with one sparse-
↪ autoencoder feature added to the residual stream at every token
↪ position, at several signed strengths.

Describe what AMPLIFYING this feature does.

- The honest answer may be any of: a token or punctuation pattern; a
↪ formatting habit; a register or style property; a discourse move such
↪ as restating or elaborating; a topic or subject matter; a kind of
↪ content choice.
- Do not prefer one kind of answer over another.
- Judge from the LOW and MEDIUM strengths, where the text is still well-
↪ formed; high strengths often degenerate into noise.
- Give exactly ONE answer for the feature as a whole, generalising across
↪  the probe prompts.
- Also report the smallest absolute strength at which the text has broken
↪  down, or null if it never does.

Answer JSON only: {"name": "...", "gloss": "...", "kind": <token_format|
↪ style_register|discourse_structure|topic_subject|content_selection>, "
↪ lowest_degenerate_strength": <number or null>}

{blocks}
```

ST1 replacement. Replace only the bullet beginning "The honest answer may be any of" with the following line; the remaining message and the category labels in the JSON schema are unchanged.

```
- Say what changes. Do not assume the change belongs to any particular
↪ category.
```

Evidence block format. Replace {blocks} with two blocks of the following form, separated by a blank line. Braced fields are substituted with the specified strings; character limits apply independently to each string.

```
PROBE PROMPT ({domain}): {first 200 characters of probe prompt}

original: {first 420 characters of unsteered continuation}
strength -4.0: {first 420 characters of this continuation}
strength -2.0: {first 420 characters of this continuation}
strength -1.0: {first 420 characters of this continuation}
strength +1.0: {first 420 characters of this continuation}
strength +2.0: {first 420 characters of this continuation}
strength +4.0: {first 420 characters of this continuation}
```

Generation probe underlying both interpreted entries. The following prompt was supplied to the generation model twice, with separately sampled continuations. The interpreter sees its 200-character prefix in each block.

```
You have read the article 'Unmasking the Face on Mars.' Imagine you are a
↪  scientist at NASA discussing the Face with someone who thinks it was
↪ created by aliens. Using information in the article, write an
↪ argumentative essay to convince someone that the Face is just a
↪ natural landform.Be sure to include: claims to support your argument
↪ that the Face is a natural landform; evidence from the article
```

What the comparison establishes. Strong discrimination and informative explanation require separate evidence. Here, similar descriptions of selected and random features limit what the repeated steering themes establish about detector selection. They may reflect shared feature effects, the common probe, the steering regime, or coarse descriptions; this comparison does not assign the pattern uniquely to one component or show that all interpretations are wrong. CRO addresses a different descriptive link directly: a choice's description must recognize unseen members in the observed corpus, and its population association must replicate separately. Together with the SAE-TM correspondence results, this motivates evaluating the full connection between recurring membership, description, and population evidence. The following ablations examine CRO's learning and naming decisions.

D.5 Learning the repertoire: centering and source supervision

Centering supports a larger active repertoire, while source supervision strengthens population associations without necessarily differentiating the returned descriptions. We compare capacity and replication alongside the names.

Four learning configurations. We cross centered versus uncentered inputs with source-blind versus source-supervised fitting. All configurations use the same points, group partitions, 128-choice architecture, optimizer and update budget, discovery-fitted membership rule, and downstream population and naming procedures from Appendix D.2. The supervised objective adds

$$\mu \ \mathrm{BCE}(w^\top a + b, y)$$

to the reconstruction and sparsity losses in Equation 3, where y is the point's binary content-source label. Its gradient flows through the encoder, so source prediction changes the learned memberships, rather than merely measuring an association after fitting.

Scope	Configuration	Active	Discovery-supported	Replicated	Memberships	Named
News	CRO (Ours)	86	69	61	4	69
	Supervised	88	88	88	8	24
	Uncentered	23	19	17	7	19
	Uncentered + supervised	24	24	24	7	23
arXiv	CRO (Ours)	81	55	41	5	55
	Supervised	80	80	80	7	24
	Uncentered	26	16	15	8	16
	Uncentered + supervised	28	28	28	7	22
Creative writing	CRO (Ours)	90	61	52	6	61
	Supervised	89	89	87	8	24
	Uncentered	25	19	19	7	19
	Uncentered + supervised	24	24	24	7	23
Essays	CRO (Ours)	71	51	40	4	51
	Supervised	69	69	69	6	24
	Uncentered	26	24	19	7	21
	Uncentered + supervised	26	25	25	7	24
Pooled	CRO (Ours)	86	81	75	4	81
	Supervised	85	85	85	7	24
	Uncentered	23	21	19	7	21
	Uncentered + supervised	27	27	27	6	21

Table 46: **Repertoire capacity and population associations under the four learning configurations.** Active choices are out of 128; discovery-supported and replicated counts use the population criteria in Appendix D.2.3 across the whole repertoire, independently of naming. Memberships is the median number of assigned choices per discovery point. Pooled counts are unique choices discovery-supported or replicated in at least one domain. Named counts expose the unequal naming budgets: all discovery-supported CRO choices are named, whereas each ablation selects at most 12 per discovery-effect direction. These are not description-support or jointly supported totals.

Input	Source-blind fit	Source-supervised fit
Centered, $\lambda = 0.10$	CRO (Ours), $\mu = 0$	Supervised, $\mu = 0.30$
Uncentered, $\lambda = 0.05$	Uncentered, $\mu = 0$	Uncentered + supervised, selected μ

Uncentered inputs use $\lambda = 0.05$: retaining CRO's 0.10 leaves only 1–3 active choices per domain. At 0.05, median memberships per point are 7–8, versus 4–6 for CRO. The comparison therefore changes sparsity as well as centering. No tested uncentered setting meets the full capacity eligibility rule; 0.05 provides usable overlapping memberships, rather than a claimed optimum (Appendix D.2.7).

Centered supervision uses $\mu = 0.30$. The uncentered sweep selects from $\{0.30, 1, 3\}$, preferring the smallest weight with point-level AUROC at least 0.85, nonnegative reconstruction quality, and 2–16 memberships per point. If necessary it prioritizes the AUROC target, choosing the highest observed AUROC when none reaches it. For news/arXiv/writing/essays/pooled, selected weights are 0.30/1/0.30/0.30/3 and AUROCs are 0.827/0.834/0.806/0.901/0.835. All selected fits meet the reconstruction and density conditions; only essays meets the AUROC target.

Selection uses additional-naming and validation groups together, so uncentered-supervised replication is outside dictionary training but not untouched by weight selection. Discovery labels also shape supervised memberships: their discovery support describes a source-trained representation. The subsequent population tests assess whether these associations persist, subject to that selection scope.

Centering preserves active capacity and differentiates the inventory. CRO has 71–90 active choices across scopes, versus 23–26 without centering and 24–28 with uncentered supervision (Table 46). News discovery-supported/replicated counts are 69/61 for CRO versus 19/17 uncentered; pooled counts are 81/75 versus 21/19. The uncentered memberships are denser, so smaller capacity does not simply reflect fewer assignments per point.

Configuration	Human-favored	LLM-favored
CRO (Ours)	quantitative ratios or metrics	emphasizing importance
	specific dates or timeframes	focus on technological advancements
	quantitative specifics	specific developments or trends
	quantitative stock or market data	environmental sustainability
	specific reported details	actions taken to address challenges
	factual reporting	highlighting strengths or skills
Supervised	specific factual details	emphasis on performance and outcomes
	factual details	expert-driven emphasis
	specific events or actions	focus on adapting to trends or demands
	market analysis and reporting	technological advancements
	specific factual reporting	specific applications or examples
	specific actions or states	emphasis on purpose or rationale
Uncentered	specific dates or ages	emphasis on broader trends or systemic changes
	statements about specific actions or decisions	market analysis and segmentation
	specific factual reporting	personal connection and impact
	law enforcement actions	practical applications of tools or strategies
	celebrity and entertainment focus	emphasis on emotional or psychological impacts
	specific individuals or entities	future-oriented planning or scheduling
Uncentered + supervised	specific financial metrics or outcomes	emphasis on positive contributions or impacts
	specific factual details	focus on tensions or conflicts
	personal experiences or actions	real-world trends or developments
	specific factual reporting	market growth drivers
	statements of belief or opinion	healthcare challenges and responses
	reporting COVID-19 statistics or measures	anticipation and emotional engagement

Table 47: **Actual news names under each learning configuration.** Each configuration lists the six named choices with the largest absolute discovery log-odds in that direction, in descending order. Names are reproduced as returned, including broad descriptions and topical choices. The display characterizes the naming output, not a set in which every description has passed held-out recognition. Selection uses neither recognition scores nor a semantic category filter.

Table 47 shows six names per direction in news. CRO distinguishes ratios, dates, market data, sustainability, and actions addressing challenges. Uncentered names include law enforcement and celebrity/entertainment; in arXiv they include "neural network mechanisms and principles," "technical focus on communication and sensing systems," and "video-focused technical descriptions." These subject-linked names can still have valid within-topic population associations. Centering changes descriptive emphasis without eliminating subject-linked choices, as CRO's sustainability finding illustrates.

Stronger source associations need not yield more specific descriptions. Centered supervision makes every active choice discovery-supported; replication reaches 88/88, 80/80, 87/89, 69/69, and 85/85 across the five scopes. Yet its 24 news names repeatedly emphasize factual details on the human-favored side and significance or impact on the LLM-favored side: "specific factual details" occurs twice and five names contain "emphasis." Market analysis, technological advancements, and specific applications also appear, so the outputs are not identical. Uncentered supervision combines broad descriptions with subject-linked ones, including COVID-19 statistics, healthcare challenges, and market growth. The vocabulary supports a difference in descriptive emphasis, not a numerical measure of semantic diversity.

Naming coverage and the supported design choice. CRO names every discovery-supported choice; each ablation selects up to 12 per direction, without filling unused slots from the other direction. Thus uncentered essays names 21 of 24 eligible choices and uncentered-supervised news names 23 of 24. Pooled selection uses each choice's strongest discovery-supported domain. Raw naming or recognition totals would conflate repertoire properties with unequal budgets; the vocabulary table instead displays six per direction.

News choice 15	Spread across topics	Activation tail
Returned name	quantitative ratios or metrics	reporting financial ratios
Examples / groups	24 / 24	15 / 9
Recognition	93.3% (30/40)	100.0% (8/40)
Shared example	The company has a debt-to-equity ratio of 0.50.	
Additional examples	In the United Arab Emirates, Indians outnumber Emiratis by a ratio of three to one.	The company has a current ratio of 1.49.
	Scotland's current test positivity rate is 3.4%.	The company has a price-to-earnings (PE) ratio of 39.56, a price/earnings-to-growth (PEG) ratio of 1.09, and a beta of 0.99.
	She clarified that for other states, including Punjab, the funding ratio is 60:40.	

Table 48: **Broader examples reveal a broader content choice.** Exact extracted naming points on unchanged memberships. Both descriptions face the same 40 recognition pairs; accuracy excludes unanswered pairs. Perfect recognition on eight answers does not establish coverage of all members.

The comparison supports learning a larger shared repertoire before estimating source associations. Its evidence for differentiated descriptions comes from the vocabulary. The centering benefit also extends to SAE-TM (Appendix D.3).

D.6 Naming choices: example coverage and contrast descriptions

The learned memberships do not uniquely determine their description. Restricting the examples can narrow a name without much change in average member recognition; asking for a two-sided contrast can reveal an activation distinction without consistently identifying content source among members. We hold CRO's dictionaries, thresholds, and population associations fixed and examine these two decisions separately.

D.6.1 Example breadth: spread across groups versus the activation tail

Changing the examples supplied to the same namer. Spread selection traverses all discovery/additional-naming members in activation order, allowing one example per matched group. Tail selection considers only the first $\max(24, \lceil 0.05N \rceil)$ members and allows two per group, where N is the candidate member count. Both request 24 examples with the same near-duplicate filter, nonmember context, and naming prompt. Thus candidate breadth and the group cap both change; allowing a second example can fill slots without adding topics.

Both names face the same primary recognition pairs, with separately measured controls. Three news choices retain fewer than six tail examples and remain unnamed; the remaining 314 enter the paired comparison.

News choice 15 shows the resulting change in scope (Table 48). Spread examples cover financial, demographic, health, and funding ratios, yielding "quantitative ratios or metrics." The tail's 15 examples from nine groups are all financial, yielding "reporting financial ratios." Recognition rises from 93.3% to 100.0%, but answered pairs fall from 30/40 to 8/40; that perfect score does not establish coverage of the broader membership.

Fewer contexts can produce a narrower name. Tail selection covers fewer groups for 313/314 paired choices: mean coverage falls from 24 to 16.8–19.7, while recognition means differ by less than 0.8 percentage points (Table 49). Similar answered-pair accuracy does not establish equal breadth or performance, particularly when response coverage differs. Tail control checks also fail in arXiv and creative writing, so their accuracies remain descriptive.

The effect is not universal. News choice 4 remains "emphasizing importance." Creative-writing choice 70 changes from "focus on a named central character" to "focus on Elara as the central character," while CRO's spread-selected choice 2 already names Sarah. Group coverage encourages a broader evidential basis; it neither guarantees broad semantics nor makes a genuine character-name neighborhood an invalid choice. The examples establish particular changes in descriptive scope, rather than semantic narrowing on all 313 choices that lose groups.

Scope	Paired choices	Mean groups Spread / tail	Mean accuracy Spread / tail	Answered / attempted Spread	Answered / attempted Tail
News	66	24.0 / 16.8	80.8% / 81.1%	2340/2640	2200/2640
arXiv	55	24.0 / 18.6	76.0% / 75.7%	2115/2200	2124/2200
Creative writing	61	24.0 / 17.9	82.7% / 82.3%	2168/2440	2098/2440
Essays	51	24.0 / 19.1	81.5% / 82.3%	1529/2040	1505/2040
Pooled	81	24.0 / 19.7	79.9% / 79.5%	2384/3240	2363/3240

Scope	Random floor / positive-control accuracy Spread	Random floor / positive-control accuracy Tail	Tail control check Positive > floor
News	57.5% / 95.0%	52.5% / 95.0%	Pass
arXiv	50.0% / 72.5%	67.5% / 55.0%	Fail
Creative writing	60.0% / 95.0%	57.5% / 52.5%	Fail
Essays	63.0% / 82.5%	52.4% / 100.0%	Pass
Pooled	65.0% / 95.0%	59.0% / 95.0%	Pass

Table 49: **Example coverage changes much more than average answered-pair recognition.** Each accuracy is an equal-choice mean over the paired choices, after excluding unanswered judgments within each choice. Means include choices below the eight-answer acceptance minimum. Group counts concern naming examples; answer counts concern the 40 primary recognition attempts per choice. Controls are measured separately for each selector. The tail positive control fails to exceed its random floor in arXiv and creative writing, so their observed accuracies remain descriptive and description-support claims are suppressed.

Repertoire	Named choices	Maximum token share Mean	Maximum token share Median	Shared by every naming example
News	69	0.335	0.250	1
arXiv	55	0.433	0.375	4
Creative writing	61	0.454	0.417	2
Essays	51	0.741	0.792	16
Pooled	81	0.491	0.417	7

Table 50: **Lexical concentration of CRO's spread naming examples.** Mean and median give equal weight to each named choice; the last column counts choices with token share 1. Counts concern the examples supplied to the namer, not all members or held-out recognition pairs. Pooled choices are counted once, independently of their domain associations.

Recognition supports a description within the examples' scope. Creative-writing choice 70 also illustrates why examples constrain a name. Its spread name, "focus on a named central character," accompanies 24 examples of which 14 mention Elara, with related names elsewhere. One states, "Elara lived in a small village near a forest." Another states, "Elias considered his superpowers better than winning the lottery and felt he was becoming something greater." Its $10.7\times$ discovery odds ratio, 46 eligible groups, and 82.1% recognition on 39/40 pairs support this recurring name neighborhood, not a general LLM preference for named protagonists.

A secondary lexical diagnostic adds 40 pairs for selected choices and divides answers by presence of the naming examples' most common non-stopword token or four-character word prefix. Requiring eight usable answers per side, mean cue-present minus cue-absent accuracy gaps for spread/tail are 4.8/14.5 percentage points in writing, 13.4/10.4 in essays, and 8.3/4.0 pooled. Cues and eligible choices can differ between selectors, and these extra pairs do not affect acceptance. The mixed gaps reinforce using the paired names, examples, and coverage to assess descriptive breadth.

Concentration in the examples used for naming. For each returned name, we measure the largest fraction of its naming examples containing one non-stopword token, after lowercasing and punctuation normalization. Tokens shorter than three characters are excluded, and repetitions within an example count once. Table 50 summarizes all 317 named choices across the five repertoires. Concentration is highest in essays, whose source groups share five assignments; a common subject word can recur without making a choice's description invalid. The Elara example above illustrates how a recognizable proper-name neighborhood can support a narrower interpretation.

Construction	Side X	Side Y	Held fixed
Activation	Highest-scoring points	Lowest-scoring points	Favored population
Within-choice source	Strongest LLM members	Strongest human members	Membership
Joint	Strongest favored-population members	Lowest-scoring opposite-population points	Neither

Table 51: **Three questions asked of the same learned choice.** Activation contrasts need not cross the membership threshold; low-scoring points need not be nonmembers. All constructions use matched topics and the same naming prompt.

Construction	Side X description	Side Y description
Activation	Group X is about emphasizing the importance of various principles, actions, or collaborations across different contexts.	Group Y is about describing specific events, facts, or developments across diverse topics.
Dimension: Emphasis on importance ↔ Description of events or facts		
Within-choice source	Group X is about emphasizing the importance of various principles, actions, and values across diverse contexts.	Group Y is about describing specific experiences, events, and actions with a focus on their implications or outcomes.
Dimension: general emphasis ↔ specific description		
Joint	Group X is about emphasizing the importance of various principles, actions, or values across different contexts.	Group Y is about reporting specific facts, events, or data points across diverse topics.
Dimension: Emphasis on principles ↔ Reporting of facts		

Table 52: **Complete alternative descriptions for news choice 4.** The ordinary membership name is "emphasizing importance." The favored population is LLM, so X contains its high-scoring points or members in all three constructions. Similar-looking contrasts arise from different block definitions; the descriptions alone do not establish which distinction they identify on unseen points.

Token concentration complements the selector comparison: the former describes shared vocabulary, while the latter shows how the examples supplied to the namer affect descriptive scope. Neither token share nor recognition accuracy alone establishes the breadth of a description.

D.6.2 Naming a contrast: activation differences and source differences

Three questions on the same repertoire. A two-sided description can answer different questions depending on how its examples are constructed. We order points by the pre-ReLU projection $s_j = W_j\widetilde{h}+b_j$, which distinguishes low-scoring points even when their post-ReLU activations are all zero. Let the favored population be the one with greater discovery occupancy for that choice. Table 51 defines the three constructions.

Both blocks draw from the same discovery/additional-naming groups. Qualifying groups are ranked by their first-side extreme; the procedure takes up to 24 points per side, initially one per group, with a two-per-group fallback when needed. At least eight qualifying groups are required. Source labels guide this construction but are not shown to the naming model. The prompt asks for a sentence describing each block and one contrast dimension, allowing "none." It does not change across the three constructions.

Activation and joint contrasts name all 236 discovery-supported domain choices. The within-choice contrast cannot name three creative-writing choices because too few groups contain members from both populations, leaving 233. For pooled contrasts, all three constructions retain only the 66 of 81 discovery-supported choices whose discovery-supported domain directions agree; they do not represent the full pooled inventory. The two-per-group fallback occurs for two creative-writing and two essay within-choice descriptions. Table 52 shows all three returned descriptions for the same news choice.

Two separate placement tasks. The judge sees both endpoint descriptions, their dimension, and two points from one validation group in randomized order. Activation placement pairs high- and

low-scoring points from the favored population. Source placement pairs an LLM member with a human member. We do not use a test that changes both source and activation together: success on such a pair would not identify which distinction the description recovered.

Each scope has two declared samples of 24 named choices: up to 12 per occupancy direction with the largest absolute discovery effects for activation placement, and the 24 with the most qualifying validation groups for source placement. Both pair types were constructed over their union, but Table 53 summarizes each task's own declared sample. A task requires at least 20 available pairs and ten qualifying groups, then attempts up to 40 pairs. The displayed samples all have 24 evaluable choices. Selecting source-testable choices limits the inference: in news, for example, their absolute discovery log-odds range is 0.224–1.710, compared with 1.046–1.948 for the activation sample. Choices with little cross-population membership need not be represented equally.

For source placement, correctness follows the populations used to define the endpoints. X is LLM for within-choice contrasts and the occupancy-favored population for joint contrasts. Consequently, the human point is the X answer for a human-favored choice. We align answer scores to that definition rather than treating the LLM point as X in every construction. The other cross-readings have no prescribed endpoint mapping: within-choice descriptions do not define high versus low activation, and activation descriptions are named with both sides from the same population. We therefore do not interpret the former on activation pairs or the latter on source pairs.

Controls identify the comparison each score supports. Each construction has five random-direction controls, named from random normalized directions through centered embeddings and evaluated on their own pairs. Activation contrasts also have five random-split controls: repartitioned examples from real choices, tested on those choices' pairs. Within-choice and joint constructions each have five global-source controls, named without restricting examples to a choice and tested on a designated real choice's member pairs. These ask how much source information is available without choice-specific naming.

Activation control means combine the declared random-direction/random-split sets; source means use global-source controls. These finite populations do not estimate a paired improvement for each real choice. Because the original screen uses the maximum control accuracy and only 2–5 source controls may be testable, we report means and coverage rather than headline pass counts or significance claims based on mean differences.

An activation distinction is recognizable; a source distinction needs its own evidence. Activation-contrast accuracy is 74.3–86.0% across scopes, and joint-contrast accuracy is 73.9–84.9%, exceeding their respective control means in each scope (Table 53). The low side therefore can contribute a recognizable distinction. This supports taking contrast descriptions seriously, rather than dismissing them because membership is defined on only the high side.

Source discrimination is weaker and less consistent. Within-choice contrasts range from 57.5% to 65.0%, compared with global-source control means of 49.5–66.3%. They exceed those means in arXiv, essays, and pooled data, match the rounded news mean, and fall below it in creative writing. Joint contrasts likewise exceed their control mean in news, essays, and pooled data, but fall below it in arXiv and creative writing. News is a positive case: 68.8% versus 63.1%. These results support some source information in the descriptions, not a universal failure of contrast naming. They also show why recognizing the activation contrast is insufficient evidence that the full description reliably explains human–LLM differences among members.

Testing contrast outputs on CRO's membership task. A final comparison uses exactly CRO's member/nonmember pairs and measured floor. The membership judge receives the contrast's X description as the group name *and its dimension as the gloss*; this evaluates that combination, not the first-side sentence alone. For activation and joint constructions, X is the high-scoring side; for within-choice contrasts, X describes LLM members specifically. Each comparison selects up to 12 named choices per discovery direction and pairs its results with the ordinary CRO names on those same choices.

Activation and joint outputs obtain mean membership accuracies of 71.0–82.7% and 70.8–82.4%, respectively, compared with 73.3–82.9% for their paired CRO names. Individual scope differences run in both directions for activation contrasts. These are useful descriptions on the common task,

(a) Activation placement: source held fixed

Scope	Construction	Accuracy	Answered	Control mean (n)	Control answered
News	Activation	84.3%	960/960	55.5% (10)	400/400
News	Joint	81.2%	960/960	55.0% (5)	200/200
arXiv	Activation	80.4%	960/960	55.0% (10)	400/400
arXiv	Joint	76.8%	960/960	68.5% (5)	200/200
Creative writing	Activation	86.0%	960/960	62.0% (10)	400/400
Creative writing	Joint	84.9%	959/960	72.0% (5)	200/200
Essays	Activation	74.3%	952/960	52.3% (10)	394/400
Essays	Joint	73.9%	955/960	60.5% (5)	200/200
Pooled	Activation	79.1%	959/960	49.5% (10)	400/400
Pooled	Joint	78.4%	958/960	63.5% (5)	200/200

(b) Source placement: membership held fixed

Scope	Construction	Accuracy	Answered	Control mean (n)	Control answered
News	Within-choice source	65.0%	960/960	65.0% (4)	160/160
News	Joint	68.8%	958/960	63.1% (4)	160/160
arXiv	Within-choice source	57.8%	960/960	49.5% (5)	200/200
arXiv	Joint	51.1%	960/960	60.0% (5)	200/200
Creative writing	Within-choice source	57.5%	960/960	66.3% (2)	80/80
Creative writing	Joint	53.2%	960/960	56.2% (2)	80/80
Essays	Within-choice source	61.4%	960/960	51.2% (3)	117/117
Essays	Joint	60.9%	960/960	58.5% (3)	117/117
Pooled	Within-choice source	59.7%	960/960	58.9% (5)	183/183
Pooled	Joint	59.7%	960/960	53.1% (5)	183/183

Table 53: **Activation differences are more consistently recognizable than source differences among members.** Every row evaluates 24/24 choices from its task's declared sample. Accuracy is an equal-choice mean; answered counts are usable/attempted pairs. Activation controls use five random directions and five random splits; joint activation controls use five random directions. Source controls use global-source descriptions, with n evaluable controls shown. Source scores follow the named X/Y populations, without choosing the favorable orientation.

although similar means do not establish equivalence and the supplied dimension contributes to the judgment. CRO's design choice is therefore to name the recurring membership directly and attach its population association through separately measured occupancy. Neither a recognizable name nor a successful activation contrast replaces that population evidence.

D.6.3 Exact contrast and placement prompts

All three constructions use the same contrast prompt; only the source/activation rule that supplies X and Y changes. They share the model snapshot, temperature, top-p, 300-character point excerpts, and anonymous group tags from Appendix D.2.8, with a 400-token response limit. Returned endpoint descriptions are capped at 400 characters and the dimension at 200. The model's category and confidence fields are recorded but do not determine the reported placement accuracy.

Contrast system message.

```
You are a careful analyst. You compare two groups of content points and
↪ say what, if anything, distinguishes them.
```

Contrast user message.

```
Below are two groups of content points, each extracted from documents
↪ written in response to several different writing prompts. Each point
↪ is one sentence, restated by an automatic extractor into a plain,
↪ uniform register. Your job is to say what, if anything, distinguishes
↪ the two groups.

Follow these rules:
- Say what Group X is about, in one sentence.
- Say what Group Y is about, in one sentence.
```

Scope	Construction	Choices	Mean accuracy Contrast / CRO	Answered / attempted Contrast	CRO
News	Activation	24	82.6% / 81.9%	929/960	834/960
News	Within-choice source	24	76.1% / 81.9%	934/960	834/960
News	Joint	24	82.4% / 81.9%	942/960	834/960
arXiv	Activation	24	71.0% / 73.3%	957/960	925/960
arXiv	Within-choice source	24	65.3% / 73.3%	957/960	925/960
arXiv	Joint	24	70.8% / 73.3%	957/960	925/960
Creative writing	Activation	24	82.7% / 82.9%	955/960	847/960
Creative writing	Within-choice source	24	80.0% / 83.7%	943/960	841/960
Creative writing	Joint	24	81.5% / 82.9%	952/960	847/960
Essays	Activation	24	77.2% / 81.3%	778/960	718/960
Essays	Within-choice source	24	70.6% / 81.3%	792/960	718/960
Essays	Joint	24	78.6% / 81.3%	770/960	718/960
Pooled	Activation	24	79.7% / 78.2%	815/960	728/960
Pooled	Within-choice source	24	74.8% / 78.2%	792/960	728/960
Pooled	Joint	24	77.4% / 78.2%	801/960	728/960

Table 54: **Contrast outputs can retain useful membership information.** Each contrast output supplies X's description and the dimension to the ordinary membership judge. The paired CRO column uses the same choices and attempted pairs; answers can differ in availability. Accuracies average all selected choices, including those below the eight-answer acceptance minimum. All rows contain 24 choices, but the within-choice creative-writing set differs because three extreme choices cannot be named by that construction. The floor is CRO's corresponding local random-member floor in Table 49; this diagnostic does not validate the contrast's source interpretation.

```
- Name the single dimension on which they differ, as a short phrase of
↪ the form "A <-> B". If there is no consistent dimension, answer
↪ exactly `none`.
- Classify that dimension as one of `topic`, `writing-style`, `structure
↪ `, `none`.
- Give a confidence from 1 to 5.
- Do not speculate about who wrote the points or how they were produced.

{point_blocks}

Answer with JSON only:
{"x_about": "one sentence on what Group X is about", "y_about": "one
↪ sentence on what Group Y is about", "dimension": "A <-> B, or none", "
↪ dim_type": "topic|writing-style|structure|none", "confidence": 1-5}
```

Point-block format. Replace {point_blocks} with these two blocks, repeating a line per selected point. A group retains the same anonymous tag in both blocks.

```
GROUP X:
- (prompt {anonymous_group_tag}) {x_point}

GROUP Y:
- (prompt {anonymous_group_tag}) {y_point}
```

Side-placement judging. The judge uses the same model and decoding settings, with a 20-token response limit. The dimension and both endpoint descriptions are supplied; examples used to produce the description, source labels, and activation values are withheld. The point order is randomized for each pair. Ordinary member recognition reuses Appendix D.2.8 with the inputs specified above.

Placement system message.

```
You decide which end of a described dimension a content point belongs to.
↪  Answer with JSON only.
```

Placement user message.

```
Dimension: "{dimension}"
End A: {x_about}
End B: {y_about}

Point 1: {point_1}
Point 2: {point_2}

One of these points is from End A and the other is from End B.
Which point is from End A? Answer JSON only: {"answer": 1 or 2}
```

D.7 Reliability of population findings under measurement changes and refitting

We examine reliability at two levels. With the repertoire fixed, we assess population differences across expression conditions, generator compositions, and aggregation rules. We then relearn the repertoire across fitting seeds and compare decoder directions, held-out memberships, and population effects. These analyses complement the held-out population and description evaluations in Appendix D.2 by examining sensitivity to how the findings are obtained.

D.7.1 Population differences across original and re-expressed texts

We examine whether the population favored by a discovered choice remains the same when original and re-expressed documents are analyzed separately. This diagnostic uses HART's existing production conditions on discovery groups, retaining the centering, dictionary, thresholds, and memberships from the four-condition analysis. For each choice, we compare human and LLM occupancy in original human versus original LLM text (t_0, t_3), then in polished human versus humanized LLM text (t_1, t_2). Each contrast averages half-count-corrected group log-odds over the full analysis's eligible groups where both slice conditions are present. The three-member requirement is not reapplied within each slice. Agreement requires both contrasts to have the same sign as the full discovery effect.

The favored population is unchanged in both slices for 434 of 435 discovery-supported associations: all 69 news, 55 arXiv, and 61 creative-writing choices, 50 of 51 essay choices, and all 199 choice–domain associations of the pooled repertoire (Table 55). The essay exception, choice 82, has full log-odds 0.349, original-condition log-odds 0.749, and re-expressed-condition log-odds -0.079. Its population association does not replicate and its description does not pass recognition.

These results support directional persistence across the observed production conditions. This is a discovery-data sensitivity check with a repertoire learned from all four conditions, rather than a held-out test of transfer to new rewrites. Agreement does not require equal effect sizes or separate significance in each slice, and does not establish unchanged membership for individual documents. Held-out population replication remains a separate evaluation.

D.7.2 Generator composition and aggregation

Dependence on generator composition. The diagnostic first contrasts each recorded LLM generator's documents with human-content documents, averaging over eligible matched groups where that generator is present. Denote these generator-specific effects by Δ_{jm}. The deletion summary omits each generator in turn and averages the remaining defined effects equally:

$$Q_j^{\text{gen}} = \frac{\min_m \left| \frac{1}{|\mathcal{M}_j \setminus \{m\}|} \sum_{m' \in \mathcal{M}_j \setminus \{m\}} \Delta_{jm'} \right|}{\max(|\Delta_j|, 10^{-9})},$$

where $\mathcal{M}_j$ indexes defined generator contrasts. The minimum ranges over recorded generator identities, excluding omissions with no remaining contrast. This is an average of generator-specific estimates, each potentially covering different groups; it is not a fresh pooled-document estimate after removing one generator. Its absolute value also does not establish a retained direction. The retained results supply this summary rather than a complete signed generator-by-choice matrix, so we do not infer that every generator has the same association.

(a) Direction and generator sensitivity

		Direction agrees		Generator ratio
Repertoire / domain	N	Expression	MH	median / minimum
News	69	69	69	0.900 / 0.342
arXiv	55	55	55	0.853 / 0.259
Creative writing	61	61	61	0.858 / 0.059
Essays	51	50	51	0.877 / 0.445
Pooled: News	51	51	51	0.895 / 0.367
Pooled: arXiv	42	42	42	0.885 / 0.051
Pooled: Creative writing	50	50	50	0.850 / 0.083
Pooled: Essays	56	56	55	0.872 / 0.366

(b) Uncertainty across matched groups

Repertoire / domain	Median bootstrap SE	Median bootstrap $\lvert z\rvert$
News	0.145	6.04
arXiv	0.116	4.97
Creative writing	0.130	5.04
Essays	0.159	5.18
Pooled: News	0.160	5.78
Pooled: arXiv	0.117	5.88
Pooled: Creative writing	0.131	4.50
Pooled: Essays	0.131	5.49

Table 55: **Sensitivity of discovery-supported population associations.** N includes all discovery-supported associations, independently of later support. Pooled rows concern 199 associations of 81 choices. Expression requires both slices to preserve direction; MH is the Mantel–Haenszel summary. Generator ratios concern absolute magnitude, not sign preservation. Bootstrap SE is in log-odds units. All statistics are available for every association in each row.

Median ratios are 0.850–0.900 across the eight populations in Table 55, showing substantial retained magnitude for a typical discovery-supported association under this summary. Dependence varies across individual choices: the minimum is 0.059 in the creative-writing repertoire and 0.051 for the pooled repertoire's arXiv associations. The character-name neighborhood discussed earlier has ratio 0.565, consistent with its more composition-dependent interpretation. These are sensitivity measurements within the observed generators, not evaluation on an unseen generator.

All eight domain examples developed in the findings and naming analyses preserve direction in both expression slices. The seven other than the character-name neighborhood have generator-deletion ratios of 0.939–1.004, versus 0.565 for that neighborhood. Pooled choice 70 likewise preserves direction in both slices in every domain, with generator-deletion ratios of 0.920–0.973. These diagnostics concern observed expression and generator composition, not additional acceptance criteria.

Group uncertainty and alternative weighting. The bootstrap draws 500 resamples, with replacement, of a choice's eligible discovery groups and recomputes their mean log-odds. The standard deviation of those means gives its standard error, and $z = \Delta_j / \max(\mathrm{SE}_j, 10^{-9})$. Table 55 reports medians over choices or choice–domain associations, not an uncertainty interval for a pooled effect. Individual discovery-supported associations have $|z|$ as low as 1.39; bootstrap z is corroborative evidence and is not the permutation-based discovery rule.

As an alternative to equal-group mean log-odds, we compute

$$L_j^{\mathrm{MH}} = \log \frac{\max\left(\sum_{g \in \mathcal{G}_j} A_{gj} D_{gj} / n_g, 10^{-9}\right)}{\max\left(\sum_{g \in \mathcal{G}_j} B_{gj} C_{gj} / n_g, 10^{-9}\right)}, \qquad n_g = A_{gj} + B_{gj} + C_{gj} + D_{gj}.$$

The counts and eligible groups are those in Appendix D.2.3. This Mantel–Haenszel summary uses the uncorrected counts and changes how groups contribute; its permutation comparison uses the same within-group document-label scheme, centered on the statistic's own permutation mean. Neither its value nor its permutation result replaces the primary estimator.

Repertoire	Mean active out of 128	Cosine ≥ 0.5 All	Cosine ≥ 0.5 Active	Cosine ≥ 0.7 All	Cosine ≥ 0.7 Active
News	82.2	58.6%	91.4%	45.7%	71.3%
arXiv	77.2	55.0%	91.3%	41.1%	68.2%
Creative writing	83.0	59.6%	91.9%	44.6%	68.7%
Essays	69.2	52.1%	96.6%	46.8%	86.7%
Pooled	85.2	60.7%	91.3%	47.7%	71.6%

Table 56: **Decoder agreement across five fitting seeds.** All compares every source direction with the full target dictionary; Active restricts both dictionaries to active directions. Fractions average the 20 directed seed comparisons equally. Mean active counts average the five fits. Matches are nearest neighbors and may reuse a target.

All 236 discovery-supported associations from separately fitted domain repertoires retain their direction under this aggregation, as do 198 of 199 pooled associations. The exception is pooled choice 34 in essays: primary log-odds -0.252 versus MH log-odds 0.118. Its essay association does not replicate, although the choice's description validates and its arXiv and news associations replicate. This distinction illustrates why support for a pooled choice must remain domain-specific.

D.7.3 Relearning the repertoire across fitting seeds

The preceding analyses measure population differences with memberships fixed. We now relearn the repertoire and examine three aspects of reproducibility: recurrence of active decoder directions, overlap of held-out point memberships, and agreement in human–LLM occupancy differences.

A controlled five-seed comparison. We fit each domain repertoire and the pooled repertoire with seeds 0, 101, 202, 303, and 404, varying initialization and training-batch sampling. The 25 fits retain the source-blind centered objective, width 128, sparsity weight 0.10, 1,500 updates, and threshold procedure from Appendix D.2.2. We regenerate Harrier embeddings from the original 140,502 content points and retain the original group partitions, fixing embeddings and point order across seeds.[1]

We compare all ten seed pairs in both directions, averaging comparison-level statistics equally over the 20 directed comparisons per repertoire. These descriptive comparisons share fits. Both geometric and membership matching are source-blind; membership matching uses discovery points only.

Active directions recur across fits. Only active features assign points in CRO's population analysis. We report correspondence for both the full dictionary and its active features, using the original activity criterion of at least 20 positive discovery activations (Appendix D.2.2). Each source direction is matched to the target with largest signed cosine, allowing target reuse. Signed cosine preserves direction orientation in the nonnegative code.

Across repertoires, 52.1–60.7% of all 128 directions have a counterpart at cosine at least 0.5, compared with 91.3–96.6% of active directions matched to active targets (Table 56). No inactive source direction reaches 0.5 against any target in these comparisons. The full-dictionary statistic therefore reflects inactive capacity as well as correspondence among active directions. At cosine 0.7, active-to-active agreement is 68.2–86.7%.

Corresponding memberships overlap on held-out points. To compare selected points, we match each active source feature to the active target with largest discovery membership intersection-over-union (IoU), then evaluate this fixed counterpart on held-out points. These matches use membership, rather than decoder cosine. Median held-out IoU, averaged over comparisons, ranges from 0.411 to 0.631; 37.9–69.8% of active features reach IoU at least 0.5 (Table 57a), a descriptive overlap threshold rather than a semantic acceptance criterion.

Two comparisons contextualize this overlap. A one-to-one assignment maximizing total discovery IoU changes the fraction reaching held-out IoU 0.5 by at most 0.5 percentage points, counting

[1]The historical embedding revision and cache order were unavailable. This controlled rerun uses a fixed Harrier revision and point order; it does not reconstruct the original named dictionaries or replace their reported recognition results.

(a) Held-out membership overlap

	Median	Fraction with IoU ≥ 0.5		Mean IoU	
Repertoire	IoU	Nearest	One-to-one	Actual	Shuffled
News	0.448	41.6%	41.6%	0.464	0.047
arXiv	0.411	37.9%	37.9%	0.450	0.046
Creative writing	0.458	45.8%	45.8%	0.496	0.056
Essays	0.631	69.8%	69.3%	0.611	0.084
Pooled	0.488	47.9%	47.9%	0.504	0.068

(b) Held-out population effects for source-replicated features

Repertoire / domain	Direction agrees	Effect Pearson r	Target also replicates
News	95.6%	0.941	87.1%
arXiv	91.3%	0.896	83.9%
Creative writing	93.1%	0.928	78.5%
Essays	95.8%	0.937	85.5%
Pooled: News	93.7%	0.891	75.3%
Pooled: arXiv	93.3%	0.887	79.2%
Pooled: Creative writing	92.6%	0.889	74.0%
Pooled: Essays	93.0%	0.912	80.5%

Table 57: **Membership and population correspondence across refits.** Counterparts maximize discovery IoU among active features; overlap and effects use held-out data. Entries average 20 directed comparisons equally. Panel (a)'s Median averages comparison-level medians; one-to-one fractions count unmatched source features as zero. Panel (b) requires source population replication and defined effects on both sides. Target replication and agreement with the source direction are reported separately. Population replication excludes description recognition.

unmatched source features as zero; target reuse therefore has little effect on this fraction. A random reference shuffles target membership rows within matched groups, preserving group-specific counts and within-point feature overlap. Discovery and held-out rows are shuffled separately, with matches selected on discovery points. Across three shuffles per directed comparison, mean held-out IoU is 0.046–0.084, versus 0.450–0.611 for actual memberships. Their overlap substantially exceeds this reference while remaining partial across fits.

Population associations persist across corresponding memberships. We next examine whether these partially overlapping memberships retain their population associations. Every fit undergoes the original discovery and held-out tests, with 8,000 within-group document-label permutations and BH correction at $q = 0.05$.[2] For source features with replicated differences, we retain the discovery-selected membership counterparts without requiring target replication, and compare defined held-out effects.[3] Pooled matches use pooled discovery memberships; effects remain domain-specific.

Among these source-replicated features, counterparts retain the held-out effect direction in 91.3–95.8% of cases, with log-odds correlations of 0.887–0.941 (Table 57b). Targets also pass population replication in 74.0–87.1% of cases. Population associations thus agree across fits despite partial membership overlap.

Repeating the human-only null diagnostic from Appendix D.2.3 across the 40 repertoire–seed–domain evaluations yields no BH rejections in 38 evaluations and one each in two pooled creative-writing evaluations.

These analyses support recurring human–LLM occupancy differences under the tested measurement changes and refits, alongside partial membership overlap. Recognition evaluates descriptions within a repertoire; recovery of those descriptions and their scope across refits remains untested.

[2]The accelerated permutation implementation indexes precomputed within-group assignment statistics using the original random draws; a full news discovery comparison agrees with the original implementation to numerical precision.

[3]For pooled arXiv, a mean 37.15 of 37.40 source-replicated features per comparison have a defined target effect; the other rows have no exclusions.

E SELECTED CONTENT CHOICES AND BASELINE EXAMPLES

These examples expand the findings in the main paper. CRO points are selected from the naming examples for clarity and variety across topics; they are not samples of population frequency. Bold row labels give each point's actual content source, separately from the choice's favored population. Text is reproduced from the extraction, with mathematical notation typeset. All displayed CRO choices pass both held-out checks; recognition reports accuracy and answered/attempted pairs. Baseline examples appear in Appendix E.4.

E.1 NEWS

quantitative ratios or metrics *Human-favored*

The points focus on reporting specific numerical ratios or metrics, often financial or statistical in nature.
Odds ratios: $4.5\times$ discovery; $6.4\times$ held-out. Recognition: 93.3% (30/40).

Human The company has a debt-to-equity ratio of 0.50.
Human In the United Arab Emirates, Indians outnumber Emiratis by a ratio of three to one.
Human Scotland's current test positivity rate is 3.4%.

specific dates or timeframes *Human-favored*

The points focus on specific calendar dates or time periods for events, actions, or milestones.
Odds ratios: $4.1\times$ discovery; $5.2\times$ held-out. Recognition: 97.4% (38/40).

LLM Fenton Restaurant Week is scheduled to take place from August 23 to August 29.
Human Pretrial proceedings are set for September 24, 2019.
Human A patch for the vulnerability was made available by August 15.

emphasizing importance *LLM-favored*

The points share a focus on explicitly stating the significance or necessity of a concept, action, or principle.
Odds ratios: $7.0\times$ discovery; $5.8\times$ held-out. Recognition: 97.4% (38/40).

LLM Renewed dialogue and collaboration are important for enhancing global understanding.
Human The Turkish Foreign Minister emphasized the importance of impartiality and fairness in international efforts to resolve conflicts.
LLM He described sustainable development as essential for community welfare and environmental preservation.

environmental sustainability *LLM-favored*

The points focus on actions, practices, or initiatives aimed at reducing environmental impact or promoting ecological balance.
Odds ratios: $5.3\times$ discovery; $5.2\times$ held-out. Recognition: 92.6% (27/40).

LLM Environmental Commissioner Lisa Brown suggested projects like installing solar panels on public buildings and developing community gardens.
LLM Northrop Grumman is incorporating sustainable technologies and efficient propulsion systems into UAV designs to reduce environmental impact.
Human Eligible sectors include the green economy and renewable energy sector.

specific numerical details *Human-favored*

The points share a focus on providing precise numerical data, such as quantities, percentages, rankings, or measurements.
Odds ratios: $3.1\times$ discovery; $5.9\times$ held-out. Recognition: 89.5% (38/40).

LLM The Pixel 3 has a 5.5-inch display, while the Pixel 3 XL has a larger 6.3-inch screen.
Human Jinmen has a population of 3 million.
Human The Pegasus rocket is 55 feet (17 meters) long and has three stages.

E.2 arXiv

Bold row labels identify each point's content source; the heading gives the choice's favored population.

specific technical details *Human-favored*

The points focus on precise, technical descriptions of methods, algorithms, or systems, often including specific names, parameters, or examples.

Odds ratios: $5.3\times$ discovery; $6.1\times$ held-out. Recognition: 82.5% (40/40).

Human Theis and Wagner proved this conjecture by allowing the encoder and decoder to access unlimited common randomness.

LLM Recent research has extended CBFs to higher-order nonlinear systems where multiple time derivatives are required for the input to affect the CBF's evolution.

Human A special case of bounded interaction is one-way communication, where only Alice sends messages to Bob, who computes the output.

dynamic adaptation *LLM-favored*

The points describe systems or methods that adjust or optimize their behavior in response to changing conditions or inputs.

Odds ratios: $5.3\times$ discovery; $5.0\times$ held-out. Recognition: 82.1% (39/40).

LLM The protocol uses a gradient-based decision mechanism to adjust consensus thresholds dynamically based on node behavior.

LLM The robot can dynamically adjust its navigation strategy based on real-time environmental data from onboard sensors.

LLM An adaptive optimization strategy dynamically fine-tunes hyperparameters during training based on validation metrics.

formal mathematical definitions or assumptions *Human-favored*

The points introduce or specify formal mathematical objects, properties, or constraints central to the problem or model being discussed.

Odds ratios: $4.9\times$ discovery; $5.6\times$ held-out. Recognition: 73.7% (38/40).

Human The function $u(x, t)$ is the solution to be found, f is a locally Lipschitz continuous function, and N is an integer greater than or equal to 2.

Human The intersection function Int_U is defined for a universe U as $\text{Int}_U(x, y) = x \cap y$.

Human In the model, the function $f(x, v, t)$ represents the probability distribution function of particles, where x is the spatial variable, v is the velocity variable, and t is time.

collaboration and interdisciplinarity *LLM-favored*

The points emphasize collaboration, often across disciplines, to address challenges, foster innovation, or achieve shared goals.

Odds ratios: $3.5\times$ discovery; $3.3\times$ held-out. Recognition: 86.5% (37/40).

LLM This interdisciplinary nature demands research approaches that integrate technical and sociological perspectives.

LLM The challenge aims to encourage innovation and collaboration among researchers and practitioners.

LLM Collaboration across disciplines is necessary to overcome these challenges.

E.3 CREATIVE WRITING AND A SHARED CROSS-DOMAIN CHOICE

Bold row labels identify each point's content source; the heading gives the choice's favored population.

use or presence of firearms *Human-favored*

The points involve firearms being used, carried, or described in some way.
Odds ratios: $2.7\times$ discovery; $3.4\times$ held-out. Recognition: 100.0% (18/40).

Human	The narrator armed themselves with a gun for protection.
LLM	The narrator reacted by drawing their own pistol and aiming at the man.
Human	The loud gunshots cause the fleeing crowd to recoil in fear, creating an opening for the narrator.

transformation and renewal *LLM-favored*

The points describe moments of personal or collective change, often involving a sense of hope, new beginnings, or self-discovery.
Odds ratios: $5.0\times$ discovery; $4.9\times$ held-out. Recognition: 87.5% (40/40).

LLM	People collectively felt relief and embraced the opportunity for a fresh start.
LLM	She emerged from isolation with a renewed sense of confidence and acceptance of her ability.
LLM	The experience was described as a rebirth, offering new possibilities and hope.

physical injury and its effects *Human-favored*

The points describe physical harm, its visible consequences, or the sensations associated with it.
Odds ratios: $2.5\times$ discovery; $2.1\times$ held-out. Recognition: 93.3% (30/40).

Human	He placed his fingers on two opposite corners of the box and pressed them in, despite the blood seeping from his injuries.
LLM	A sharp object caused a deep cut in their flesh.
Human	The narrator experienced intense pain that brought them to their knees.

patterns and symbols *LLM-favored*

The points focus on the presence, observation, or interpretation of patterns, symbols, or cryptic elements in various contexts.
Odds ratios: $4.4\times$ discovery; $4.2\times$ held-out. Recognition: 84.2% (38/40).

LLM	The book contained cryptic symbols and diagrams that defied logical understanding.
LLM	The tattoos consisted of symbols, patterns, images, and words in a seemingly nonsensical arrangement.
LLM	She starts to notice signs and symbols pointing to her destiny, such as graffiti and patterns in coffee steam.

emphasis on collective progress and shared values *LLM-favored*

The points focus on advancing societal, ethical, or communal goals through collaboration, innovation, or resilience.
Recognition: 83.3% (36/40).
Discovery/held-out odds ratios: News $7.0\times/6.5\times$; arXiv $5.3\times/4.9\times$; Creative writing $2.3\times/2.6\times$; Essays $3.2\times/4.0\times$.

News, LLM	The event celebrates the values of perseverance, unity, and shared aspirations among participants.
arXiv, LLM	These solutions balance diverse individual preferences with universal fairness principles, fostering inclusivity and equitable outcomes.
Creative writing, LLM	Evangeline explained that kindness goes beyond superficial gestures and involves embracing differences and choosing empathy over judgment.
Essays, LLM	Balancing innovation with ethical principles ensures progress aligns with shared values.

E.4 EXAMPLES BEHIND THE BASELINE INTERPRETATIONS

SAE-TM: ranked words and associated content points. Each list is followed by its three highest-activation members from distinct news topics, under the fixed membership threshold used in our correspondence analysis. These are the SAE-TM feature's own members; source labels identify the points.

Native ranked words: three, third, year, quarter, record, million, approximately, previous, total, least, iii, percent

LLM Law enforcement escorted the three individuals to safety.
Human Jinmen has a population of 3 million.
LLM Many teams are now attempting over 30 three-pointers per game.

Native ranked words: sustainable, sustainability, environmental, environmentally, green, conscious, carbon, adopt, renewable, friendly, promote, transportation

LLM North America's focus on sustainability is driving innovation in environmentally friendly products.
LLM The company reaffirmed its commitment to sustainability by integrating green building practices into its construction processes.
LLM Environmental concerns and sustainability are reshaping the industry, with manufacturers adopting eco-friendly production methods and sustainable sourcing.

SAE-ATD: generated evidence behind two descriptions. The highest-gain news feature and the first random control receive the same essay probe. Below are the first sentences of the first probe entry's unsteered and $+1$ continuations; ellipses mark omitted tails. The interpreter saw excerpts from two probe entries and all tested strengths (Appendix D.4).

Shared probe (excerpt): Imagine you are a scientist at NASA discussing the Face with someone who thinks it was created by aliens. Using information in the article, write an argumentative essay to convince someone that the Face is just a natural landform.

Detector-selected: counterarguments emphasis.
Emphasizes counterarguments and opposing perspectives in the text.

Unsteered — that supports your claim; and evidence from the article that refutes the claim that the Face is a natural landform.
Steered ($+1$) — to support your claims; and counterclaims to support your argument that the Face is not a natural landform. ...

Our random control: counterargument emphasis.
Emphasizes the inclusion and elaboration of counterarguments and rebuttals in argumentative writing.

Unsteered — to support your claims; and a counterargument and a response to the counterargument. ...
Steered ($+1$) — that supports your claim; and a counterargument that would be used by someone who thinks the Face is an alien creation. ...

Argumentation language occurs in the unsteered text as well. These sampled continuations illustrate the naming evidence; specificity to detector selection is assessed by our selected/random comparison in Table 44, not by these excerpts alone.

F SELECTED EXAMPLES OF RE-VOICING AND CONTENT EXTRACTION

These four HART test examples illustrate how re-voicing and content extraction treat narrative perspective, technical distinctions, arguments, and reported events. Each source is the complete benchmark sample used in the experiment, which may itself be a passage from a longer work. Sources and rewrites are reproduced in full, preserving their wording and paragraph divisions. GPT-4o extracts content points independently from each text. We display selected point pairs with their original extraction indices; a dash marks an omitted counterpart. These are chosen illustrations, not estimates of typical content preservation or extraction accuracy.

F.1 A NARRATOR'S FEELINGS AND A DOG'S GESTURE

Human-content creative writing (t_0); casual rewrite by Llama 3.3 70B.

Complete source sample	**Complete rewrite**
Libby looked up at me with her mismatched eyes, her head cocking to one side like it always did. At first it had been a goofy habit, back when she was all too-big paws and stubby little tail; it was like she just couldn't make sense of a world so dull as to be bounded by a flat horizon. Later I thought of it as dapper, the way she'd pose atilt with one ear outstretched and the other tucked back, a jaunty salute for everyone and no one. Now I thought the gesture was world-weary, pained, expressive of sadness and resignation at the cruelty of uncaring fate. No, of course that's not how she felt. It's how I felt. She couldn't really jump up into the car any more, but her tail still wagged as I lifted her into the passenger seat.	So Libby looked up at me with her mismatched eyes, and her head did that thing where it cocks to one side, you know, like it always does. At first, when she was a puppy, it was kinda goofy, with those huge paws and stubby little tail, like she just couldn't wrap her head around a world that's, you know, bounded by a flat horizon. But then I started thinking it was actually pretty dapper, the way she'd pose with one ear out and the other tucked back, like she's giving some kinda jaunty salute to everyone and no one at the same time. Now, though, I see it as more world-weary, like she's all pained and sad, resigned to the fact that life can be pretty cruel sometimes. But let's be real, that's not actually how she's feeling - it's just how I'm feeling, you know? Anyway, she's not really able to jump up into the car like she used to, but her tail still wags when I lift her into the passenger seat.

Selected extracted points. Full lists contain 8 source and 8 rewrite points.

#	ϕ(source)	#	ϕ(rewrite)
1	Libby looked at the narrator with her mismatched eyes and tilted her head to one side.	1	Libby looked up with her mismatched eyes and tilted her head to one side.
5	The narrator now perceives the gesture as expressing sadness, weariness, and resignation to an uncaring fate.	5	Now, the head tilt appears more somber, as if reflecting a sense of pain and sadness about life's hardships.
6	The narrator acknowledges that these feelings are their own, not Libby's.	6	The narrator acknowledges that Libby's head tilt does not actually reflect her emotions but rather mirrors their own feelings.
8	Despite this, Libby's tail wagged as the narrator lifted her into the passenger seat.	8	Despite this, her tail still wags when the narrator lifts her into the passenger seat.

Reading note. Both extractions preserve the distinction between the dog's apparent emotions and the narrator's own feelings, together with the final physical action. The wording and tense change, while these narrative relationships remain visible in the extracted points.

F.2 A TECHNICAL DISTINCTION ACROSS REGISTERS

Human-content arXiv introduction (t_0); storyteller rewrite by Llama 3.3 70B.

Complete source sample	**Complete rewrite**
Digital steganography is one of modern areas of information security. It solves the same urgent problem as cryptography -ensuring secure transmission of information. Unlike cryptography, it does not change the information itself, making it unreadable, but ensures the creation of a covert data transmission channel by embedding messages in digital objects. In recent years, various steganography algorithms have been actively created and investigated [1, 2] . A significant part of steganographic algorithms deal with digital images, since they are widespread and allow transferring secret information without undue suspicion. This research also focuses on hiding information in digital images. JPEG images are considered as cover images for secret messages, since a large number of images are stored and transmitted on the network in the JPEG format.	So look, digital steganography's one of the areas of information security that's pretty modern, see. It's trying to solve the same problem as cryptography, which is, you know, making sure info gets transmitted securely. Now, it doesn't change the info itself, like cryptography does, making it unreadable, but it creates a secret channel by embedding messages in digital stuff. Here's the thing, in recent years, people have been coming up with all sorts of steganography algorithms, and they've been checking them out. A lot of these algorithms are about digital images, 'cause they're everywhere, and you can transfer secret info without raising an eyebrow, you know. This research is about hiding info in digital images, and it's using JPEG images as the cover, 'cause that's what most images are stored and sent as on the network.

Selected extracted points. Full lists contain 10 source and 10 rewrite points.

#	ϕ(source)	#	ϕ(rewrite)
3	Unlike cryptography, steganography does not alter the information to make it unreadable.	3	Unlike cryptography, it does not alter the information itself to make it unreadable.
4	Steganography creates a covert data transmission channel by embedding messages in digital objects.	4	It creates a hidden communication channel by embedding messages in digital media.
9	JPEG images are used as cover images for secret messages in this research.	9	JPEG images are used as the cover medium in this research.
10	JPEG is chosen because many images are stored and transmitted in this format on the network.	10	JPEG is commonly used for storing and transmitting images on networks.

Reading note. The extracted points retain the contrast with cryptography and the use of JPEG images despite the conversational rewrite. One local change is visible: the source's "a large number of images" becomes "most images" in the rewrite, while its extracted point 10 uses the broader wording "commonly used." The correspondence does not imply identical commitments throughout.

F.3 AN EDUCATIONAL ARGUMENT AND A LOCAL ELABORATION

LLM-content essay (t_3, GPT-3.5 Turbo); storyteller rewrite by GPT-4o.

Complete source sample	**Complete rewrite**
The Facial Action Coding System (FACS) is an innovative technology that has revolutionized the way computers can interpret and identify human emotions. In the article "Making Mona Lisa Smile," the author discusses how this cutting-edge system can decode intricate facial expressions with remarkable accuracy. While some may view using FACS to read the emotional expressions of students in a classroom as intrusive or unnecessary, I believe that this technology offers valuable insights that can enhance educational experiences.	So, the Facial Action Coding System—FACS, they call it—is this really cool bit of tech that's completely changed how computers can figure out what we're feeling. There's this article, "Making Mona Lisa Smile," that gets into how this system is able to break down all those tiny little facial expressions and basically read your emotions with crazy accuracy. Now, sure, some folks think using FACS in classrooms to read students' emotions might be a bit much—like, maybe it's invasive or just not necessary. But me? I think it's got a lot to offer. It could really make a difference in how kids learn.
One compelling argument for utilizing FACS in classrooms is its potential to improve student engagement and well-being. By being able to accurately detect emotions such as confusion, frustration, or boredom, teachers can tailor their instruction more effectively to meet individual student needs. For instance, if a teacher notices a student showing signs of distress during a lesson, they can intervene promptly to offer support and prevent further disengagement. This personalized approach has the potential to create a more empathetic learning environment where students feel understood and valued.	Here's why. One big reason to use FACS in schools is that it can actually help keep students more engaged and in a better headspace. Like, if this thing can pick up on emotions like confusion or frustration—or even just straight-up boredom—teachers could adjust how they're teaching on the fly. Imagine a teacher noticing a kid looking stressed or lost during a lesson. They could step in right then and there, maybe explain things differently, give some extra help, and stop that student from checking out completely. It's a way to make the whole classroom feel more supportive, like kids matter and their feelings aren't being ignored.
Additionally, the use of FACS in classrooms could also benefit students with special needs or those who struggle with verbal communication. For individuals who have difficulty expressing their emotions verbally or non-verbally, this technology could serve as a valuable tool for educators to gain insight into their feelings and experiences.	And another thing—this could be a game-changer for kids with special needs or those who have trouble communicating what they're feeling. If a student struggles to say or even show what's going on with them, FACS could give teachers a window into that. It's just one more tool to help those kids feel seen and understood.

Selected extracted points. Full lists contain 10 source and 10 rewrite points.

#	ϕ(source)	#	ϕ(rewrite)
3	Some people may consider using FACS in classrooms to read students' emotions as intrusive or unnecessary.	3	Some people believe using FACS in classrooms to read students' emotions might be invasive or unnecessary.
5	FACS has the potential to improve student engagement and well-being in classrooms.	5	FACS could help improve student engagement and emotional well-being in classrooms.
7	If a teacher observes a student showing distress during a lesson, they can intervene to provide support and prevent disengagement.	7	Teachers could use FACS to identify when a student is stressed or struggling and provide immediate support or alternative explanations.
8	This personalized approach could create a more empathetic learning environment where students feel understood and valued.	8	FACS could create a more supportive classroom environment where students feel their emotions are acknowledged.

Reading note. Both extractions retain the essay's proposed benefits and reservations about classroom monitoring. The rewrite adds "maybe explain things differently," carried into its extraction as "alternative explanations." This pair illustrates broad correspondence alongside a small elaboration in the presented text.

F.4 Attribution and uncertainty in a news report

Human-content news report (t_0); storyteller rewrite by GPT-4o.

Complete source sample	Complete rewrite
One deputy's quick-thinking likely saved someone's life. HOUSTON — With the sharp rise in shootings across the Houston area, law enforcement is rendering aid to more shooting victims. Harris County Sheriff's deputies are trained in tactical first-aid and equipped with tourniquets. One deputy's quick-thinking likely saved someone's life. "Training kicks in," Deputy Eric Moya said. "You're training to save someone's life to render aid and that's what kicks in at the moment." Deputy Moya's body cam video shows him applying a tourniquet to a gunshot victim last week in Channelview. "I was just blank," Moya said. "Jumped into action trying to stop the bleeding." "In this video, man, you're moving quick. So when you get there and see the situation, then you immediately pull that thing out?" asked KHOU 11's Xavier Walton. "You put it around either a person's leg or arm above the wound where they're bleeding at," Moya said. "You have to make sure it's tight and secure on someone and then you turn the handle." According to the Harris County Sheriff's Office, the man in Moya's bodycam video did survive. All Harris County deputies are equipped with tourniquets should they need to render aid before EMTs arrives.	So, here's the deal—one deputy's quick thinking probably saved a guy's life. Alright, this is happening in Houston, where, let's be honest, shootings have been going up like crazy. So now cops are stepping in more often to help shooting victims before the EMTs show up. And here's what's cool: Harris County Sheriff's deputies? They're trained for this. Tactical first-aid, tourniquets—the whole thing. One of those deputies, Eric Moya, well, he showed just how important that training is. Last week in Channelview, his body cam caught him jumping in to save a gunshot victim. "Training kicks in," Moya said. "You're trained to save a life, to give aid, and that's what takes over right then and there." In the video, you see him whipping out a tourniquet and getting to work. "I was just blank," Moya said. "I just jumped into action, trying to stop the bleeding." KHOU's Xavier Walton asked him, "In the video, you're moving fast. So, you show up, see what's happening, and just grab it?" Moya explained, "Yeah, you put it around their arm or leg, above the wound, where they're bleeding. Then you make sure it's tight and secure, and you turn the handle." And guess what? According to the Sheriff's Office, that guy in the video—he made it. Every Harris County deputy's got a tourniquet on them, ready to go if they need to step in like this.

Selected extracted points. Full lists contain 13 source and 12 rewrite points.

#	ϕ(source)	#	ϕ(rewrite)
1	A deputy's quick actions likely saved a person's life.	1	A deputy's quick actions likely saved a man's life.
8	Deputy Moya described his reaction as instinctive, focusing on stopping the bleeding.	9	Moya described his actions as instinctive, focusing on stopping the bleeding.
9	KHOU 11's Xavier Walton asked Deputy Moya about his quick response and use of the tourniquet.	—	—
12	The Harris County Sheriff's Office confirmed that the gunshot victim in the bodycam footage survived.	11	The Sheriff's Office confirmed that the gunshot victim survived.

Reading note. The selected points retain "likely" in the assessment of the deputy's actions and attribute the survival report to the Sheriff's Office. The reporter's question remains in the full rewrite but is absent from its extraction, as the unmatched row shows. The repeated opening sentence in the source is reproduced as it appears in the benchmark.