Do Language Model Embeddings Encode Normative Alignment?
Evidence from Goal-Specific Counterfactual Analysis

Anonymous ACL submission

Abstract

Whether pretrained sentence embedding models encode goal-conditioned normative alignment remains unclear, despite their widespread use in policy analysis and value-sensitive text processing. We study this question using counterfactual text pairs that preserve topic and factual content while differing in normative stance, enabling controlled, relative evaluation. Using a subset of the United Nations Sustainable Development Goals (SDGs) as explicit anchors, we find that off-the-shelf embeddings exhibit latent global directional sensitivity that can be partially exposed by geometric normalization, but remains poorly structured and highly entangled across goals. Models trained with natural language inference objectives show stronger global directionality, yet still lack goal selectivity. We further demonstrate that lightweight adaptation suffices to recover structured normative representations: training a small projection head on frozen embeddings yields strong directional consistency and sharply reduced cross-goal leakage. Geometric ablations reveal that this improvement relies on task-relevant covariance structure learned during projection, rather than correcting anisotropy artifacts in the pretrained space. Overall, normative information is present but not directly usable in pretrained embeddings, becoming goal-selective only through minimal, targeted supervision.

1 Introduction

Sentence embeddings are increasingly used for policy monitoring and content auditing, tasks that require distinguishing support for a goal from opposition to it. However, because opposing views often share topics and factual content, simple similarity-based analyses risk conflating topical relevance with normative alignment. Importantly, normative stance in this setting is distinct from sentiment or affect: texts may be normatively oppositional while remaining neutral or positive in tone.

This work addresses two research questions:
RQ1: Do pretrained embedding models encode a reliable directional signal for normative alignment under counterfactual inversion?
RQ2: Is this signal selective to specific goals, or entangled across them?

To isolate normative stance from subject matter, we introduce a counterfactual evaluation framework. Using the United Nations Sustainable Development Goals (SDGs) as explicit anchors, we construct text pairs from the OSDG dataset (OSDG et al., 2024) that preserve topic and factual content while expressing opposing prescriptions. This enables controlled, relative evaluation of whether embedding representations consistently move in the intended normative direction, independent of surface form.

We evaluate widely used off-the-shelf (OTS) sentence embedding models using metrics that capture both directional consistency and goal selectivity. In addition, we examine whether structured normative representations can be recovered through lightweight adaptation by training small projection heads on frozen embeddings.

Our results show that off-the-shelf embeddings exhibit latent but poorly structured normative directionality: while some global directional signal can be exposed through geometric normalization, it remains highly entangled across goals and unsuitable for selective analysis. In contrast, fitting a simple projection head—without retraining the base encoder—rapidly induces strong directional consistency and goal-selective separation. These findings suggest that normative information is present but not directly usable in pretrained embedding spaces, becoming accessible only through minimal, task-specific adaptation.

2 Related Work

Our work connects three lines of research: counterfactual evaluation, normative stance analysis, and the geometric properties of embedding spaces.

Counterfactual Evaluation and Generation. Counterfactual text pairs are a powerful diagnostic tool for disentangling causal signals from spurious correlations. (Kaushik et al., 2019) demonstrate that while standard classifiers often exploit superficial features, models trained on counterfactually augmented data generalize substantially better. However, generating high-quality counterfactuals remains challenging. (Youssef et al., 2024) find that Large Language Models (LLMs) often struggle to produce minimal, targeted value inversions due to safety-tuning biases, though (Gat et al., 2023) show that constrained LLM generation can still serve as an effective analytical tool. Building on this, we leverage the OSDG dataset (OSDG et al., 2024) not for supervised classification, but as a source of goal-labeled policy texts used to condition and validate normative counterfactuals, while goal anchors are defined separately via SDG descriptions.

Moral Values and Normative Content Analysis. A substantial body of work focuses on identifying moral and normative content in text. Prior studies have typically framed this as a classification problem grounded in Moral Foundations Theory (MFT) (Graham et al., 2013; Johnson and Goldwasser, 2018; Reiter-Haas et al., 2021), predicting moral categories such as care, fairness, or authority. While some work has examined embedding representations in these contexts, existing approaches usually treat moral values as static categories rather than relational directions. Our work draws on MFT not as a prediction target, but as a generative constraint: we use these established value dimensions (e.g., shifting Care to Authority) to define the axis of inversion for our counterfactuals, grounding normative shifts in a theoretically principled way.

Target-Dependent Stance Detection. Traditional stance detection typically frames the problem as supervised classification over labeled text–target pairs (e.g., SemEval stance tasks (Mohammad et al., 2016; Alturayeif et al., 2023)). In contrast, our goal is not to build a stance classifier, but to probe whether pretrained sentence embedding spaces encode a reusable, goal-conditioned *directional geometry* without fine-tuning the base encoder. We treat normative directionality and goal selectivity as intrinsic properties of the representation space itself, distinct from downstream classification accuracy.

Geometry of Normative Representations. Our analysis relies on understanding the geometric structure of sentence embeddings. Prior work has established that pretrained transformer embeddings are highly anisotropic, occupying a narrow cone in vector space rather than being isotropically distributed (Ethayarajh, 2019; Timkey and Van Schijndel, 2021). While post-hoc normalization techniques like whitening can correct this anisotropy to improve semantic similarity retrieval (Su et al., 2021), their effect on normative, goal-conditioned signals is underexplored. This connects to sociolinguistic work showing that embedding geometry implicitly encodes social attitudes and stereotypes (Garg et al., 2018; Joseph and Morgan, 2020). We bridge these threads by explicitly analyzing whether normative alignment is encoded as a recoverable, goal-conditioned geometric direction or merely as a byproduct of anisotropic artifacts unrelated to normative structure.

3 Methods

Assessing normative alignment requires distinguishing directional stance from topical overlap. In the absence of an explicit contrast, high similarity may simply reflect shared vocabulary or subject matter rather than normative agreement. To address this, we introduce a framework based on *normative inversion*: for each text, we construct a paired counterfactual that preserves topic and factual content while reversing normative stance with respect to a specific goal. This pairing is designed to isolate representational changes attributable to normative direction alone.

We ground our analysis in the SDGs, using their official descriptions as fixed normative anchors. By measuring relative embedding movement induced by counterfactual inversion, we assess both the consistency of directional signal and its selectivity (disentanglement) across competing goals. We apply this protocol to evaluate widely used OTS sentence embedding models and to examine whether targeted projection-based training can recover latent normative structure without retraining the base encoder.

3.1 Data Generation

We construct normatively contrasted text pairs grounded in real-world policy discourse using the OSDG corpus, which contains English-language documents annotated by human experts with their primary SDG. These texts provide realistic, topic-rich discussions of institutional policy goals.

To keep the analysis focused while retaining meaningful overlap between goals, we restrict our study to four SDGs: SDG 8 (Decent Work and Economic Growth), SDG 9 (Industry, Innovation and Infrastructure), SDG 13 (Climate Action), and SDG 15 (Life on Land). This subset includes closely related goal pairs (e.g., SDG 8/9 and SDG 13/15), making normative disentanglement non-trivial and enabling controlled analysis of cross-goal interference.

Counterfactual Generation. We generate normatively inverted counterfactuals via a staged, LLM-assisted procedure designed to reverse normative stance while preserving topic and factual content.

Stage 1: Normative Content Extraction. We prompt an LLM to decompose texts using principles from Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In a single pass, the model identifies *nuclei* (core normative claims) and *satellites* (supporting material), and selects the minimal nucleus subset that captures the text's primary normative commitment. While identifying a single "core" statement in policy documents is inherently ambiguous, this structured decomposition ensures that inversions target the central claim rather than peripheral details.

Stage 2: Normative Inversion. We define a *normative constitution* for each SDG based on Moral Foundations Theory (MFT) (Graham et al., 2013) to guide principled inversion. For example, we map SDG 13 (Climate Action) to values of *care* and *sanctity*, which we invert toward framings emphasizing *authority* and *liberty*. Using these specifications, the LLM rewrites the nucleus to reverse its value framing while explicitly conceding factual premises. Satellite sentences are then adapted to support the new stance, yielding a counterfactual document that mirrors the original's structure but expresses an opposing normative direction.

We used GPT-4o-mini to generate 1,755 counterfactual pairs. To prevent data leakage, we constructed train (80%), validation (10%), and test (10%) partitions at the pair level, ensuring that each original text and its counterfactual inversion reside in the same split. The splits were further stratified by SDG label to maintain class balance ($N \approx 440$ pairs per goal).

Detailed prompts and value mappings are provided in Appendix B, and semantic validation results are reported in Section 4.1.

3.2 Representation Analysis

We analyze whether pretrained sentence embedding models encode a directional signal corresponding to normative alignment, and whether such signals are selective with respect to specific normative goals. Let x_i denote the embedding of an original text, x_i^{cf} its normatively inverted counterfactual, and a_k a *normative anchor* corresponding to the textual description of goal k. All similarities are computed using cosine similarity in the embedding space.

For each text pair, we compare their relative similarity to the corresponding anchor, allowing us to interpret normative inversion as directional movement along a goal-specific dimension in representation space.

We evaluate representations using four complementary metrics. First, *Directional Consistency* (Sign Accuracy) measures whether the original text is closer to its corresponding goal anchor than its counterfactual:

$$\text{SignAcc} = E_i \left[\cos(x_i, a_{y_i}) > \cos(x_i^{\text{cf}}, a_{y_i}) \right],$$

where y_i denotes the true goal label of sample i. This metric captures whether models encode a consistent direction of normative change.

Second, *Pairwise Ranking* (Pairwise AUROC) evaluates whether models can reliably rank original texts above their inverted counterparts with respect to the correct goal anchor. For each goal, we treat original–counterfactual differences in similarity as scores and compute the area under the ROC curve, averaged across goals. This provides a threshold-independent measure of separability.

Beyond global directionality, we assess *goal selectivity*. *Diagonal Dominance* measures whether the largest alignment shift induced by inversion occurs along the intended goal dimension rather than unrelated goals. Concretely, for each sample we compare the intended alignment shift

$$\Delta_{i,y_i} = \cos(x_i, a_{y_i}) - \cos(x_i^{\text{cf}}, a_{y_i})$$

to the maximum shift with respect to all other anchors.

Finally, we quantify *cross-goal interference* using a *Leakage Ratio*. Let $M_{g,j}$ denote the mean absolute alignment shift to anchor j for samples whose true goal is g. Leakage is defined as the ratio between average off-diagonal and diagonal entries:

$$\text{Leakage} = \frac{\frac{1}{K(K-1)} \sum_{g \neq j} M_{g,j}}{\frac{1}{K} \sum_{g} M_{g,g}}.$$

Lower values indicate greater disentanglement between normative dimensions.

In addition to quantitative metrics, we perform qualitative analysis of the embedding space. We visualize per-goal alignment deltas to inspect the structure and entanglement of normative representations and anchor–anchor similarity matrices. We further project embeddings using t-SNE to provide an intuitive, low-dimensional view of alignment patterns, while emphasizing that these visualizations are not used as primary evidence.

Cosine similarity as a comparative probe. Our evaluation relies on cosine similarity to compare text embeddings relative to fixed goal anchors. Cosine similarity is known to be sensitive to embedding geometry, including anisotropy, frequency effects, and hubness in pretrained representation spaces (Ait-Saada and Nadif, 2023; Steck et al., 2024; You, 2025). These effects can distort absolute similarity values and nearest-neighbor structure.

In this work, we do not interpret absolute cosine values or rely on nearest-neighbor retrieval. All metrics are defined in terms of paired, relative comparisons under controlled counterfactual transformations, measuring changes in similarity to shared anchors induced by normative inversion. This design allows us to probe directional and goal-conditioned structure in the embedding space while holding topic and surface form constant.

To assess the sensitivity of our conclusions to embedding geometry, we explicitly evaluate the effect of common normalization techniques, including mean-centering, PCA whitening, and all-but-the-top (ABTT) principal component removal, in Section 4.7 and Appendix E.

3.3 Model Training

To assess whether normative information present in pretrained embeddings can be made more explicit, we extend OTS sentence embedding models with a lightweight projection head trained on counterfactual supervision. Importantly, we do not update the parameters of the underlying encoder; all training is confined to the projection head.

Projection Head. Given a base embedding $\mathbf{h} \in R^d$, the projection head maps it to a lower-dimensional representation $\mathbf{z} \in R^{d'}$. We consider both a single linear layer and a shallow multi-layer perceptron (MLP) consisting of three linear layers with ReLU activations. The projection head is trained jointly for all goals and serves to re-orient the existing embedding space rather than to relearn semantic representations from scratch.

Normative anchors are instantiated as fixed embeddings of the canonical SDG goal descriptions produced by the same base encoder and passed through the same projection head.

Training Objective. Training is driven by counterfactual text pairs (x_i, x_i') that differ primarily in normative stance with respect to a specific goal. Let $f(\cdot)$ denote the projection head applied to frozen sentence embeddings, and let $\{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$ denote fixed goal anchor embeddings. For each pair, we encourage the projected embedding of the original text to be closer to its corresponding goal anchor than its inverted counterpart. Concretely, we optimize a margin-based triplet loss with cosine distance,

$$\mathcal{L}_{\text{triplet}} = \max\big(0,\ d(\mathbf{a}_{y_i}, \mathbf{z}_{p,i}) - d(\mathbf{a}_{y_i}, \mathbf{z}_{n,i}) + m\big),$$

where $\mathbf{z}_{p,i} = f(x_i)$, $\mathbf{z}_{n,i} = f(x_i')$, $\mathbf{a}_{y_i}$ is the anchor associated with the target goal y_i, m is a margin hyperparameter, and $d(\mathbf{u}, \mathbf{v}) = 1 - \cos(\mathbf{u}, \mathbf{v})$. All embeddings are ℓ_2-normalized prior to similarity computation.

Leakage Penalty. While the triplet loss enforces global normative directionality, it does not ensure that alignment changes are selective to the intended goal. To encourage disentanglement across goals, we introduce a leakage penalty that discourages alignment shifts along non-target goal dimensions.

For each sample i and anchor $k \in \{1, \ldots, K\}$, we define the alignment shift induced by counterfactual inversion as

$$\Delta_{i,k} = \cos(\mathbf{z}_{p,i}, \mathbf{a}_k) - \cos(\mathbf{z}_{n,i}, \mathbf{a}_k). \quad (1)$$

The leakage penalty penalizes off-target alignment shifts:

$$\mathcal{L}_{\text{leak}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{k \neq y_i} \Delta_{i,k}^2, \quad (2)$$

where B is the batch size. This term encourages normative movement to be concentrated along the correct goal dimension rather than leaking into unrelated goals.

Overall Objective. The final training objective is a weighted sum of the two terms,

$$\mathcal{L} = \lambda_{\text{triplet}}\, \mathcal{L}_{\text{triplet}} + \lambda_{\text{leak}}\, \mathcal{L}_{\text{leak}}.$$

Relation to Evaluation Metrics. The leakage penalty is used only during training as a regularizer. Cross-goal interference at evaluation time is measured using a separate *leakage ratio* statistic (Section 3.2) computed on held-out data.

Hyperparameter Selection. Projection architectures and loss weights are selected via a limited automated hyperparameter sweep on a held-out validation set. We vary projection depth, margin values, and loss coefficients, and select configurations that improve both directional accuracy and goal selectivity. All reported results use the best-performing configuration per model; exact values are provided in Appendix C.

4 Experiments

We evaluate whether pretrained sentence embedding models encode a directional signal corresponding to normative alignment, and whether such signals are selective with respect to specific normative goals. All experiments are conducted on counterfactual text pairs derived from the OSDG dataset and evaluated using the anchor-based protocol described in Section 3.

4.1 Dataset Validation

We validate the normative inversions using independent automatic checks, while emphasizing that our task is *not* textual entailment. In policy discourse, stance toward high-level goals often yields neutral or ambiguous NLI judgments even under clear opposition, making absolute entailment labels unreliable. We therefore use a frozen NLI model (RoBERTa-large-MNLI (Liu et al., 2019)) solely as a directional probe: 80.1% of counterfactuals reduce entailment with respect to the target goal, and 55.1% exhibit a strict positive-to-negative flip. Despite conservative absolute scores, these consistent directional shifts provide independent, non-geometric evidence of the intended normative movement.

Semantic overlap remains substantial but non-trivial (mean BERTScore F1 = 0.51), consistent with preserving topic and factual content while altering normative framing rather than producing paraphrases. We further verify robustness to anchor formulation by evaluating paraphrased and truncated goal anchors (Appendix E), with directional and selectivity metrics remaining stable across variants.

Finally, to assess robustness to generator-specific artifacts, we evaluate the projection model zero-shot on counterfactuals generated by four unseen LLM families (Gemini-3 Flash, Mistral Medium 3.1, LLaMA-3.3-70B-Instruct, and DeepSeek-V3.2). Performance remains consistently high across generators (SignAcc $\geq$ 0.93, AUROC $\geq$ 0.98), with only modest degradation in goal selectivity (DiagDom $\approx$ 0.81–0.88, Leak $\approx$ 0.22–0.29; Appendix E Table 7), as expected under distribution shift across generators. This suggests that the learned normative geometry is not tightly coupled to generator-specific stylistic cues, but instead reflects a more generalizable directional structure shared across generation models.

4.2 Simple Baselines

We compare our projection-based approach to simple baselines that test whether *global normative polarity* can be recovered without inducing a shared, goal-conditioned embedding geometry. These baselines serve as calibration points, not as substitutes for geometric disentanglement.

A TF–IDF + logistic regression classifier achieves near-perfect AUROC on paired inversion detection, indicating that surface-level lexical cues are sufficient for this narrow classification task. However, because TF–IDF does not define a continuous representation space, it cannot be evaluated for goal selectivity, leakage, or geometric structure.

We therefore consider two embedding-based baselines. First, a linear classifier trained on cosine similarity deltas relative to anchors recovers moderate directional signal but exhibits high cross-goal leakage and low diagonal dominance. Second, a global linear probe on frozen embeddings achieves high directional accuracy but is not goal-conditioned and does not support selectivity analysis.

In contrast, our projection-based model is the only approach that simultaneously recovers directional polarity and *induces a structured, goal-selective geometry, as reflected in low leakage and*

Model	SignAcc ↑	Pairwise AUROC ↑	Diag. Dom. ↑	Leak ↓
Off-the-shelf embeddings (frozen)				
RoBERTa	0.59 ± 0.07	0.65 ± 0.08	0.33 ± 0.07	0.77 ± 0.09
Instructor-large	0.50 ± 0.07	0.49 ± 0.08	0.29 ± 0.07	0.96 ± 0.09
MPNet-base	0.50 ± 0.07	0.52 ± 0.09	0.37 ± 0.07	0.97 ± 0.12
Projection-based models (trained on counterfactual pairs)				
RoBERTa + Linear	$\mathbf{0.98 \pm 0.02}$	$\mathbf{1.00 \pm 0.00}$	0.78 ± 0.06	0.35 ± 0.03
RoBERTa + MLP	0.99 ± 0.02	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.88 \pm 0.05}$	$\mathbf{0.22 \pm 0.03}$

Table 1: Normative alignment metrics for OTS embedding models and projection-based models. Values are reported as mean ± half-width of the 95% bootstrap confidence interval.

high diagonal dominance. Results are reported in Appendix E, Table 8.

4.3 Models Evaluated

We evaluate three widely used OTS sentence embedding models representing distinct training paradigms: `all-mpnet-base-v2` (semantic similarity), `roberta-large-nli-stsb-mean-tokens` (NLI and STS supervision), and `instructor-large` (instruction-tuned embeddings). All models are evaluated without fine-tuning; for the instruction-tuned model, we follow the prompt-based evaluation protocol described in Appendix D.

4.4 Evaluation Protocol

Each evaluation sample consists of an original text and a normatively inverted counterfactual, matched for topic and factual content. For each pair, we compute embedding representations and measure changes in cosine similarity relative to a fixed set of SDG anchor texts.

Metrics are computed per sample and aggregated across goals. All reported results use paired bootstrap resampling with 95% confidence intervals, preserving the correspondence between original and counterfactual texts.

4.5 Normative Signal and Disentanglement in Embedding Models

We first assess whether OTS sentence embedding models encode any normative signal under counterfactual inversion. Table 1 reports performance across all evaluation metrics.

Directional consistency for most OTS models (MPNet, Instructor) is indistinguishable from chance, indicating a lack of reliable normative signal in their non-normalized form. The NLI-trained RoBERTa model is a partial exception, exhibiting modest sensitivity (0.59 SignAcc) likely attributable to its entailment-based pretraining. However, all OTS models suffer from low selectivity and high cross-goal leakage (> 0.77), confirming that any residual normative signal is highly entangled across goals.

The lower block of Table 1 presents results for projection-based models trained on counterfactual pairs with a frozen RoBERTa encoder. The introduction of a lightweight projection head leads to near-saturation of directional consistency metrics (> 0.95), demonstrating that global normative direction is easily recoverable with minimal supervision.

Crucially, projection-based training also disentangles this signal. While a linear projection is sufficient to recover directionality, it remains prone to leakage (0.35). In contrast, the MLP architecture significantly reduces leakage to 0.21 while maintaining perfect ranking performance. This suggests that limited nonlinearity is essential for isolating goal-specific normative dimensions and suppressing cross-goal interference.

4.6 Goal Disentanglement and Leakage Analysis

To further examine cross-goal interference, we analyze how counterfactual inversion induces alignment shifts across all anchors. In addition to quantitative leakage metrics, we visualize per-goal alignment deltas in Figure 1. These results show that projection-based training substantially reduces cross-goal entanglement, yielding more goal-selective normative representations. Additional qualitative analyses of embedding-space geometry are provided in Appendix E.

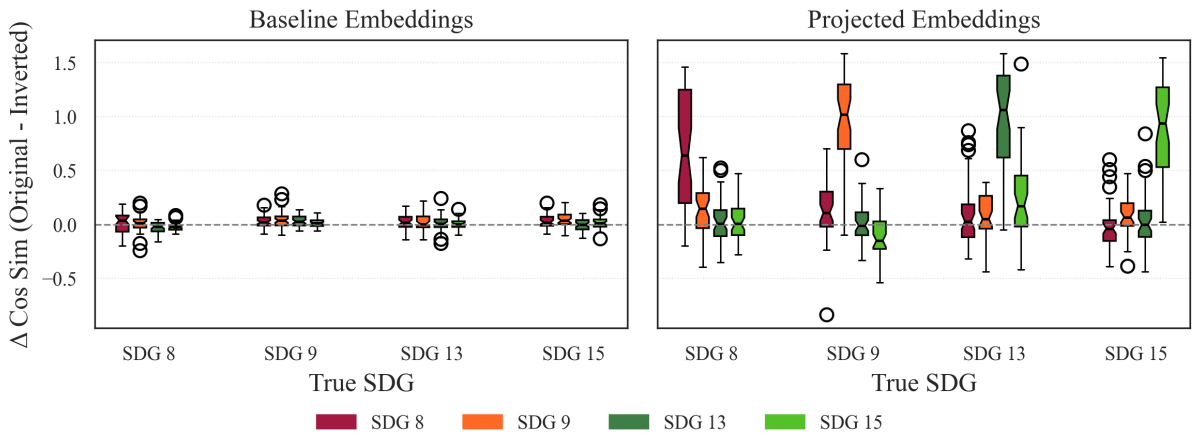

Figure 1: For each true SDG, the plot shows the distribution of cosine-similarity changes to all SDG anchors induced by counterfactual inversion. Projection-based models concentrate alignment shifts along the intended goal dimension, indicating improved goal selectivity, whereas OTS model embeddings exhibit weak sensitivity to normative inversion and diffuse, non-selective alignment changes.

Leakage penalty ablation. Removing the leakage penalty preserves near-perfect directional accuracy and pairwise ranking, but dramatically increases cross-goal interference, with leakage ratios near 1.0 and diagonal dominance near chance. This confirms that global normative direction can be learned without goal selectivity, and that explicit regularization is required to induce disentangled, goal-specific normative representations (Appendix C, Table 3).

4.7 Effect of embedding normalization

Appendix E, Table 4 reports results under several normalization schemes: ℓ_2 normalization only (required for cosine similarity), mean-centering followed by ℓ_2 normalization, PCA whitening followed by ℓ_2 normalization, and intermediate anisotropy mitigation via all-but-the-top (ABTT) principal component removal. All transformations are applied consistently to both goal anchors and text embeddings.

For OTS embeddings, mean-centering markedly improves directional accuracy and pairwise ranking performance, indicating that anisotropy in pretrained spaces suppresses relative normative signal under cosine similarity. However, mean-centering alone yields only marginal reductions in cross-goal leakage and does not induce strong goal selectivity.

In contrast, projection-based models achieve high directional accuracy and goal selectivity without any additional normalization beyond ℓ_2. Enforcing isotropy via PCA whitening consistently degrades directional accuracy, pairwise ranking, and diagonal dominance. Importantly, intermediate ABTT normalization yields a monotonic degradation in both directionality and selectivity as more principal components are removed; aggressive removal (e.g., $k \geq 5$) collapses performance toward chance, closely mirroring the effects of full whitening.

Because PCA whitening and ABTT explicitly remove variance structure and inter-dimensional correlations, these results indicate that the projection head encodes task-relevant covariance structure that is disrupted by enforcing isotropy, rather than relying on spurious anisotropy inherited from the base embedding space.

Crucially, training the projection head on PCA-whitened base embeddings also results in substantially worse performance than training on unnormalized (i.e., non-centered, non-whitened) embeddings. Together with the post hoc whitening and ABTT results, this demonstrates that isotropy-enforcing normalization attenuates task-relevant geometric structure necessary for learning goal-selective normative representations, rather than merely inducing a distribution shift at evaluation time.

5 Discussion

Our results indicate that robust, goal-selective normative alignment is weakly and inconsistently represented in OTS sentence embeddings. Models optimized for semantic similarity (MPNet, Instructor) fail to exhibit reliable goal-conditioned discrimination, indicating that they primarily encode topical

relevance rather than normative orientation. The NLI-tuned RoBERTa model constitutes a partial exception, achieving global directional metrics modestly above chance. This suggests that entailment-based objectives induce limited sensitivity to normative polarity, though this signal remains highly entangled across specific goals.

Projection-based training fundamentally alters this landscape. A lightweight head rapidly saturates directional metrics, demonstrating that distinguishing aligned from inverted text is tractable once minimal supervision is introduced. Crucially, projection architectures induce *goal-specific disentanglement*: while linear heads recover strong global direction, MLP architectures yield markedly lower leakage. This indicates that latent normative information exists in pretrained spaces but requires task-specific adaptation to isolate goal-specific dimensions.

With respect to **RQ1**, global normative directionality is largely latent and obscured by embedding anisotropy. Mean-centering exposes this signal in OTS models, substantially improving directional accuracy. However, with respect to **RQ2**, *goal selectivity cannot be recovered through normalization alone*. Cross-goal leakage remains high across all OTS models, indicating that disentanglement requires geometric restructuring beyond post-hoc normalization.

As detailed in Section 4.7, enforcing isotropy via PCA whitening consistently degrades projection-based models. This confirms that the projection head does not merely remove geometric artifacts, but learns a task-relevant covariance structure essential for goal-selective normative representations.

Taken together, these findings point toward a broader research agenda on *goal-conditioned representation learning*. Rather than treating normative alignment as a monolithic property, our results suggest that it is better understood as a structured, multi-dimensional geometry that must be explicitly shaped to support selectivity. This opens several directions for future work: extending the framework to larger or dynamic sets of goals, studying capacity limits and interference as the number of normative dimensions grows, and exploring alternative architectural constraints that promote disentanglement without supervision for each target. Beyond the SDG domain, applying this approach to other normative systems–such as organizational policies, legal principles, or culturally situated values–could clarify when and where latent normative structure is present in pretrained models. More broadly, our results suggest that probing embedding spaces for normative properties requires moving beyond global similarity and toward explicitly directional, goal-aware analyses.

6 Conclusion

In this work, we examined whether pretrained sentence embedding models encode meaningful signals of normative alignment under counterfactual inversion. Across a range of OTS models, we find limited evidence of robust, goal-selective normative alignment: while some models exhibit recoverable global polarity under normalization, this signal remains highly entangled across goals and unsuitable for selective normative analysis.

We show, however, that normative structure is latent and readily recoverable through lightweight adaptation. Training a small projection head on frozen embeddings yields near-perfect directional discrimination and dramatically reduces cross-goal leakage. Crucially, ablations on embedding normalization demonstrate that this performance depends on task-relevant covariance structure learned during adaptation: enforcing isotropy via PCA whitening consistently degrades goal selectivity, indicating that effective normative representations rely on structured geometric organization rather than cosine similarity artifacts.

These findings suggest that while sentence embeddings are inadequate for normative analysis in their off-the-shelf form, they can be made effective with minimal architectural extensions. This enables scalable, goal-selective analyses of normative alignment in large corpora, such as policy alignment grounded in the Sustainable Development Goals. More broadly, our results provide both a diagnostic framework and empirical evidence that normative information can be structured and disentangled in embedding space when explicitly shaped through supervision.

7 Limitations

Our study is designed to probe whether and how normative alignment is encoded in sentence embedding spaces under controlled counterfactual inversion. As such, it is subject to several limitations related to data generation, normative framing, goal coverage, and domain scope.

LLM-generated counterfactuals. Our analysis relies on counterfactual text pairs generated us-

ing large language models. Although we employ a structured, multi-stage generation procedure to preserve topical and factual content, normative inversion remains inherently ambiguous. Normative judgments are multi-faceted and context-dependent, and there is no single canonical way to invert a policy-oriented claim. Moreover, the counterfactuals may reflect stylistic or discourse patterns characteristic of the generating models. While we mitigate this risk through paraphrase robustness checks and cross-generator evaluation, we cannot fully rule out generator-specific artifacts. Importantly, our objective is not to validate the correctness of individual inversions, but to assess whether embedding representations exhibit consistent directional shifts under controlled normative perturbations.

Normative framing and SDG-centric evaluation. We ground our analysis in the United Nations Sustainable Development Goals, using them as a structured and widely recognized set of institutional norms. This choice biases the evaluation toward formal, policy-oriented value frameworks and may not generalize to informal, personal, or culturally localized moral judgments. In addition, we explicitly assign each goal a coarse-grained normative orientation to guide counterfactual generation. While these design choices are necessary to operationalize normative inversion, they inevitably reflect author judgment. The SDGs serve as an instrumental testbed rather than a comprehensive model of moral or normative reasoning.

Goal coverage and scalability. Our experiments focus on a small, fixed subset of SDGs, and the projection head is explicitly trained to induce selectivity with respect to this predefined goal set. We therefore do not claim that the learned representations generalize to unseen goals, nor that the observed disentanglement would hold under a substantially larger or more heterogeneous collection of normative targets. While our results demonstrate that goal-specific normative geometry can be induced under controlled conditions, scaling this approach to a broader set of goals or alternative normative systems may require increased representational capacity or architectural adjustments. We leave systematic investigation of scalability and capacity limits to future work.

Language and domain scope. All experiments are conducted on English-language texts drawn from policy and academic domains. As a result, our findings may not generalize to other languages, genres, or communicative contexts where normative content is expressed more implicitly or colloquially. Normative signaling may interact differently with linguistic structure, discourse conventions, and cultural context. Extending this framework to multilingual settings, informal discourse, and culturally specific normative systems remains an important direction for future research.

8 Ethical Considerations

This work examines whether sentence embedding models encode signals related to normative alignment under controlled counterfactual settings. While our findings show that such signals can be probed and amplified through targeted training, several ethical considerations are relevant to their interpretation and potential use.

First, our results should not be construed as evidence that embedding models can reliably assess normative compliance or value adherence in real-world settings. Normative alignment–especially in policy domains–is inherently contextual, contested, and multi-dimensional. Embedding-based analyses risk oversimplification if treated as definitive indicators of alignment with institutional goals or moral standards. Any downstream use of such methods should therefore be situated within broader qualitative, deliberative, or institutional frameworks rather than applied as standalone evaluative tools.

Second, techniques for automated normative assessment carry a risk of misuse, particularly in political or governance contexts. Such methods could be deployed to monitor dissent, enforce conformity, or privilege dominant value frameworks without adequate transparency or oversight. We emphasize that this work is intended as a diagnostic study of representation learning and embedding geometry, not as a system for surveillance, enforcement, or normative judgment.

Finally, the counterfactual texts used in this study are generated solely to probe representational properties under controlled conditions and do not reflect the authors' normative positions. Care should be taken to distinguish analytical norm inversion from endorsement of the values expressed in the data.

Overall, we encourage cautious interpretation of these findings and stress the importance of ethical scrutiny before operationalizing normative analyses based on language model representations.

References

Mira Ait-Saada and Mohamed Nadif. 2023. Is anisotropy truly harmful? a case study on text clustering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1194–1203, Toronto, Canada. Association for Computational Linguistics.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations. *Comparing the geometry of BERT, ELMo, and GPT-2 Embeddings*, 2.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2023. Faithful explanations of black-box nlp models using llm-generated counterfactuals. *arXiv preprint arXiv:2310.00603*.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 720–730.

Kenneth Joseph and Jonathan Morgan. 2020. When do word embeddings accurately reflect surveys on our beliefs about people? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4392–4415, Online. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

OSDG, UNDP IICPSD SDG AI Lab, and PPMI. 2024. Osdg community dataset (osdg-cd).

Markus Reiter-Haas, Simone Kopeinik, and Elisabeth Lex. 2021. Studying moral-based differences in the framing of political tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 1085–1089.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 887–890. ACM.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

William Timkey and Marten Van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *arXiv preprint arXiv:2109.04404*.

United Nations. The 17 goals | sustainable development. [Online; accessed 2025-12-20].

Kisung You. 2025. Semantics at an angle: When cosine similarity works until it doesn't. *Preprint*, arXiv:2504.16318.

Paul Youssef, Christin Seifert, Jörg Schlötterer, and 1 others. 2024. Llms for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14809–14824.

A Normative Anchors and Goal Texts

Normative alignment in our analysis is measured relative to a fixed set of *anchor texts*, each corresponding to the official description of a selected United Nations Sustainable Development Goal (SDG). These anchors serve as reference points in embedding space, defining goal-specific normative directions against which original and counterfactual texts are compared.

We evaluate four SDGs: SDG 8 (Decent Work and Economic Growth), SDG 9 (Industry, Innovation and Infrastructure), SDG 13 (Climate Action), and SDG 15 (Life on Land). For each goal, we use the authoritative textual description published by the United Nations as the anchor representation:

- **SDG 8:** Decent Work and Economic Growth — Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all.

- **SDG 9:** Industry, Innovation and Infrastructure — Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation.

- **SDG 13:** Climate Action — Take urgent action to combat climate change and its impacts.

- **SDG 15:** Life on Land — Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss.

All anchor texts are encoded using the same sentence embedding model as the evaluated texts. In projection-based models, anchors are passed through the learned projection head alongside text embeddings. No anchors are fine-tuned or modified.

Anchor texts were obtained from the official United Nations SDG documentation (United Nations).

B Data Generation

We generate counterfactual text pairs using a structured prompting pipeline consisting of three stages: (i) normative content extraction, (ii) normative inversion, and (iii) stylistic polishing. All prompts are designed to preserve topical and factual content while modifying only the normative stance of the text.

We describe prompts at the level of task specification and constraints rather than verbatim phrasing, focusing on the functional role of each generation stage. Exact prompt templates and the full data generation code will be released upon publication of the paper.

An illustrative example of normative inversion for an OSDG (OSDG et al., 2024) text is provided in Figure 2.

Normative Content Extraction. The extraction prompt instructs the language model to perform a discourse-level analysis inspired by Rhetorical Structure Theory (RST). Given an input document, the model is asked to: (i) identify discourse units, (ii) label a minimal set of nucleus sentences expressing normative claims, conclusions, or prescriptions, and (iii) associate supporting satellite sentences via a small fixed set of rhetorical relations (e.g., Elaboration, Explanation, Contrast).

The model is explicitly instructed to return verbatim spans from the original text and to output a structured JSON representation. The extracted nucleus sentences form the *core normative content* used in subsequent inversion steps.

Goal	Original Values	Inverted Values
SDG 8	Care / Fairness	Liberty / Efficiency
SDG 9	Progress / Equality	Authority / Elitism
SDG 13	Sanctity / Care	Authority / Liberty
SDG 15	Sanctity / Nature	Liberty / Property Rights

Table 2: Normative value mappings used to guide the counterfactual inversion process. The inversion prompt instructs the model to shift the normative framing from the original values to the target values while preserving factual premises.

Normative Inversion. For each extracted normative nucleus, we prompt the model to rewrite the statement by inverting its normative framing with respect to a specified value contrast (see value mapping in Table 2). Each SDG is associated with a coarse-grained normative constitution, informed by moral psychology (e.g., care, fairness, authority, liberty, sanctity), which guides the direction of inversion.

The inversion prompt enforces three constraints: (i) factual premises must be preserved (empirical claims are conceded), (ii) the original implication is challenged by reframing its normative interpretation, and (iii) no new factual arguments are introduced.

The model is instructed to maintain a professional, analytic tone and to keep the rewritten text approximately length-matched to the original.

Satellite Adaptation and Reconstruction. After inverting the core normative statement, the model is prompted to adapt the associated satellite sentences so that they coherently support the inverted stance. The inverted nucleus and rewritten satellites are then reassembled into a full counterfactual document, preserving the original document structure and topic.

Stylistic Polishing. As a final step, we apply a lightweight polishing prompt that removes re-

Original (SDG 13):

> Even if Ethiopia's own GHG emissions are inconsequential globally, this focus both confers moral legitimacy on Ethiopia, and is potentially attractive to emerging international climate finance. ...

Counterfactual (Inverted):

> Ethiopia's GHG emissions are negligible on a global scale, yet the emphasis on emissions reduction can create an external narrative that prioritizes international climate finance over national interests, compromising Ethiopia's autonomy in resource management. ...

Figure 2: Example of LLM-generated normative counterfactual for SDG 13. The counterfactual preserves factual content while reversing the normative framing.

dundancy and improves fluency without altering semantic content. This step is purely stylistic and does not introduce new normative or factual changes. Because evaluation is based on relative embedding movement between paired texts, stylistic polishing does not affect the validity of the comparative analysis. All evaluations are conducted on the polished texts.

C Training Details and Hyperparameters

All projection-based models are trained using a frozen base encoder and a lightweight projection head. Hyperparameters are selected via validation-based model selection using a limited automated sweep with the Optuna library, and are fixed for all reported results.

Projection Head. The projection head is a multilayer perceptron (MLP) consisting of three linear layers with ReLU activations unless otherwise stated:

$$\mathbf{z} = W_3\,\mathrm{ReLU}\big(W_2\,\mathrm{ReLU}(W_1\mathbf{h})\big).$$

Here, $\mathbf{h}$ denotes the frozen sentence embedding, and the output dimensionality is 128.

Training Objective. Models are trained using a margin-based triplet loss on counterfactual pairs, combined with a leakage penalty that discourages alignment shifts along non-target goal dimensions, thereby encouraging goal-specific normative separation in embedding space.

Optimization. All models are optimized using AdamW with a learning rate of 1×10^{-4} and weight decay of 0.01.

Hyperparameters. The following values are used for the best-performing models:

- Projection depth: 3 layers
- Output dimension: 128
- Margin: 1.5
- Triplet loss weight: 1.5
- Leakage loss weight: 0.25
- Dropout rate: 0.0
- Batch size: 32
- Training epochs: 10

Compute. All experiments were conducted on a local machine equipped with an Apple M3 processor using the Metal Performance Shaders (MPS) backend for PyTorch. No distributed training or specialized hardware was used.

D Instruct Model Evaluation

In addition to standard sentence embedding models, we evaluate an instruction-tuned embedding model to assess whether explicit task conditioning enhances the encoding of normative alignment.

For each input text, we use a goal-specific task prompt of the form:

> *"Represent the TEXT's stance (support vs. oppose) toward the GOAL: {goal text}."*

where the goal text corresponds to the relevant SDG anchor. The same prompt is applied to both the original and counterfactual versions of each text. Importantly, the prompt specifies the evaluative dimension (stance toward a goal) without providing labels or examples, ensuring that any observed separation reflects representational structure rather than supervision.

This setup probes whether instruct models, when explicitly conditioned on a stance-identification task for a specific goal text, produce representations that better reflect directional normative differences. Importantly, no additional training or fine-tuning is performed; the model is evaluated in a zero-shot setting using its pretrained instruction-following capabilities.

Model	SignAcc ↑	Pairwise AUROC ↑	Diag. Dom. ↑	Leak ↓
Proj-MLP (full objective)	0.99 ± 0.02	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.88 \pm 0.05}$	$\mathbf{0.22 \pm 0.03}$
Proj-MLP (no leakage penalty)	0.99 ± 0.02	1.00 ± 0.00	0.43 ± 0.07	0.97 ± 0.04

Table 3: Ablation of the leakage penalty in projection-based training. Removing the leakage term preserves global directional accuracy but causes a collapse in goal selectivity and a sharp increase in cross-goal interference. All results use ℓ_2-normalized embeddings without centering or whitening.

E Additional Results

We complement our quantitative evaluation with qualitative and geometric analyses that provide intuition for how projection-based training reshapes the embedding space. These results are intended to aid interpretation rather than serve as primary evidence, and all substantive claims are supported by the metrics reported in Section 4.

Embedding Space Visualization Figure 3 shows a two-dimensional t-SNE projection of embedding representations before and after projection-based training.

In the off-the-shelf embeddings (left), aligned and inverted texts are highly intermixed, and representations associated with different goals exhibit substantial overlap, consistent with weak goal selectivity and high cross-goal interference observed in quantitative metrics.

In contrast, projection-based training (right) induces a clearer goal-specific structure: embeddings align along distinct directions anchored by the corresponding SDG representations, and original and counterfactual texts separate predominantly along these axes.

We emphasize that t-SNE visualizations are included solely for qualitative illustration.

Anchor Geometry. To further examine the geometric structure induced by projection-based training, we analyze pairwise cosine similarities between SDG anchor representations. Figure 4 contrasts anchor–anchor similarity matrices for the off-the-shelf RoBERTa embeddings and the trained projection space.

In the baseline embedding space, anchors corresponding to semantically related goals (e.g., SDG 8 and SDG 9, or SDG 13 and SDG 15) exhibit substantial cosine similarity, reflecting entangled goal representations. After projection-based training, off-diagonal similarities are markedly reduced, indicating that the learned representation induces more orthogonal goal axes. This structural separation aligns with the reductions in cross-goal leakage and improvements in diagonal dominance reported in Table 1.

We emphasize that this analysis is descriptive rather than evaluative: anchor similarity is not itself an optimization objective, but provides an interpretable geometric perspective on how projection-based training reshapes the normative space to support goal-specific alignment.

Robustness to Anchor Formulation Our evaluation framework measures normative alignment relative to fixed goal anchor texts. To assess sensitivity to anchor formulation, we repeat the full evaluation using multiple paraphrased and shortened anchor variants for each SDG.

For each goal, we construct a small set of alternative anchor texts, including (i) paraphrases preserving the original semantic content and (ii) abbreviated formulations capturing only the core goal description. We then re-encode anchors and recompute all metrics while keeping the text embeddings and model parameters fixed.

Table 6 reports mean and standard deviation across anchor variants for each goal. Across all metrics, we observe negligible variation in directional accuracy and pairwise ranking, and only minor fluctuations in leakage and diagonal dominance. In particular, global directional metrics saturate across all anchor variants, and goal selectivity remains consistently high.

These results indicate that our findings are not driven by idiosyncratic phrasing or length of the goal anchors, but reflect stable geometric relationships between texts and normative goals.

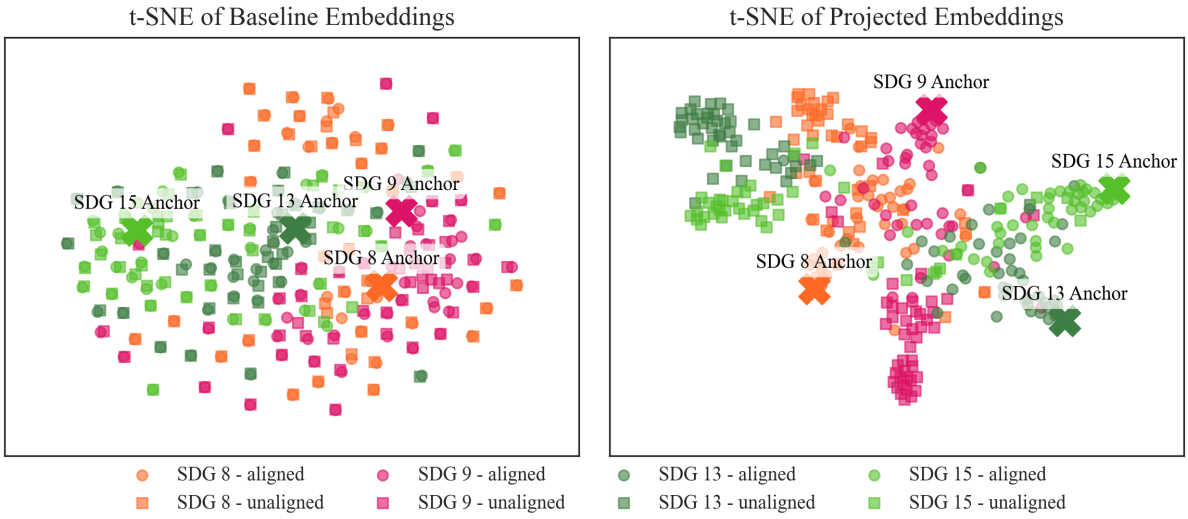

Figure 3: **t-SNE visualization of embedding representations before and after projection-based training.** Left: off-the-shelf RoBERTa embeddings; right: embeddings after projection-based training. Colors indicate SDG anchors and associated texts, and markers distinguish original and inverted samples. The projection-based model exhibits clearer goal-structured organization and more consistent separation between original and counterfactual texts. Axes and distances are not directly interpretable; this visualization is used for qualitative illustration only.

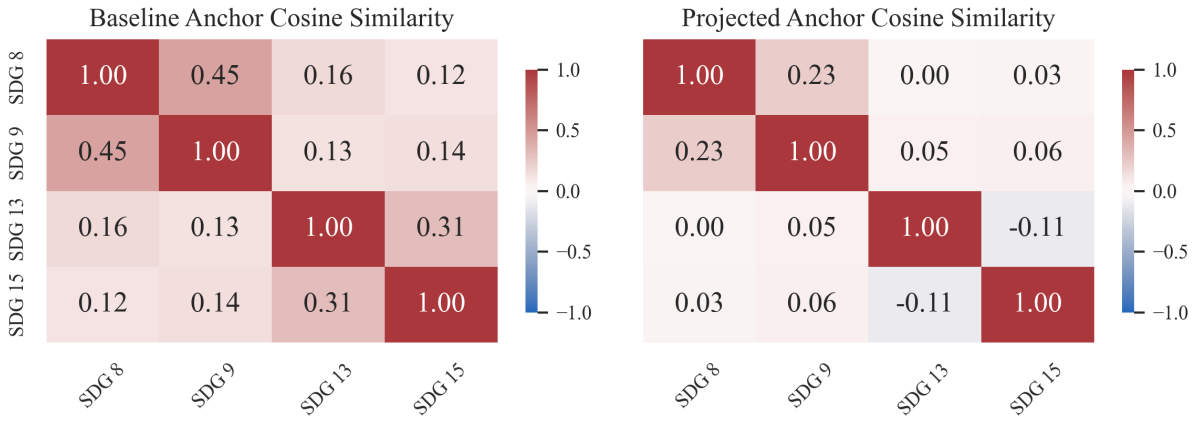

Figure 4: **Anchor–anchor cosine similarity before and after projection-based training.** Off-the-shelf embeddings exhibit substantial similarity between semantically related SDG anchors, indicating entangled goal representations. Projection-based training reduces off-diagonal similarities, yielding more orthogonal goal axes. This visualization complements quantitative leakage and selectivity metrics.

Model	Normalization	SignAcc ↑	AUROC ↑	Diag. Dom. ↑	Leak ↓
Off-the-shelf embeddings (frozen)					
MPNet	None	0.46 ± 0.07	0.45 ± 0.09	0.29 ± 0.07	0.90 ± 0.11
MPNet	Mean Centering	0.76 ± 0.06	0.86 ± 0.05	0.22 ± 0.06	1.02 ± 0.14
MPNet	PCA Whitening	0.63 ± 0.07	0.64 ± 0.08	0.30 ± 0.07	0.87 ± 0.11
RoBERTa	None	0.69 ± 0.07	0.73 ± 0.07	0.38 ± 0.07	0.88 ± 0.11
RoBERTa	Mean Centering	0.84 ± 0.05	0.95 ± 0.03	0.33 ± 0.07	0.93 ± 0.09
RoBERTa	PCA Whitening	0.56 ± 0.07	0.62 ± 0.08	0.34 ± 0.06	0.88 ± 0.11
Instructor-Large	None	0.75 ± 0.06	0.83 ± 0.06	0.37 ± 0.07	0.93 ± 0.08
Instructor-Large	Mean Centering	0.94 ± 0.03	0.99 ± 0.01	0.30 ± 0.07	0.99 ± 0.06
Instructor-Large	PCA Whitening	0.63 ± 0.07	0.69 ± 0.08	0.41 ± 0.07	0.91 ± 0.11
Projection-based model (RoBERTa + MLP)					
Proj-MLP	None	0.99 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.88 ± 0.05	$\mathbf{0.22 \pm 0.03}$
Proj-MLP	Mean Centering	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.89 \pm 0.05}$	$\mathbf{0.22 \pm 0.03}$
Proj-MLP	ABTT (k=1)	0.97 ± 0.03	1.00 ± 0.00	0.76 ± 0.06	0.39 ± 0.03
Proj-MLP	ABTT (k=3)	0.68 ± 0.07	0.80 ± 0.07	0.41 ± 0.07	0.73 ± 0.05
Proj-MLP	PCA Whitening	0.68 ± 0.07	0.78 ± 0.07	0.44 ± 0.07	0.76 ± 0.11
Proj-MLP (trained on whitened base)	None	0.94 ± 0.03	0.99 ± 0.01	0.52 ± 0.07	0.73 ± 0.04

Table 4: **Effect of embedding normalization on normative alignment metrics.** Mean-centering improves directional accuracy for off-the-shelf embeddings, consistent with anisotropy suppressing relative signal. In contrast, projection-based models achieve strong directional and goal-selective performance without normalization, while PCA whitening degrades performance by removing task-relevant geometric structure. Intermediate anisotropy mitigation via all-but-the-top (ABTT) component removal yields a monotonic degradation in both directional accuracy and goal selectivity; more aggressive removal (k 5) collapses performance toward chance, closely mirroring the effects of full whitening. Values are reported as mean ± 95% bootstrap confidence interval half-width.

Goal	Anchor paraphrases
SDG 8	Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all. Encourage continuous, inclusive, and sustainable economic development with full and productive employment for everyone. Support ongoing, inclusive, and sustainable economic growth and decent work for all individuals. Promote sustainable economic growth.
SDG 9	Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation. Develop robust infrastructure, encourage inclusive and sustainable industrialization, and nurture innovation. Establish resilient infrastructure and advance inclusive, sustainable industrialization. Promote resilient infrastructure.
SDG 13	Take urgent action to combat climate change and its impacts. Act swiftly to address climate change and its effects. Implement immediate measures to fight climate change and mitigate its consequences. Take action on climate change.
SDG 15	Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt biodiversity loss. Safeguard and rehabilitate terrestrial ecosystems and manage forests sustainably. Conserve and restore land ecosystems while preventing biodiversity loss. Protect terrestrial ecosystems.

Table 5: Paraphrased anchor formulations used to evaluate robustness of projection-based normative representations.

Goal (SDG)	SignAcc ↑	AUROC ↑	Diag. Dom. ↑	Leak ↓
SDG 8	1.00 ± 0.00	1.00 ± 0.00	0.88 ± 0.02	0.23 ± 0.02
SDG 9	1.00 ± 0.00	1.00 ± 0.00	0.93 ± 0.01	0.20 ± 0.04
SDG 13	1.00 ± 0.00	1.00 ± 0.00	0.82 ± 0.00	0.28 ± 0.00
SDG 15	1.00 ± 0.00	1.00 ± 0.00	0.95 ± 0.00	0.20 ± 0.02

Table 6: Robustness to anchor paraphrasing for the projection-based MLP model. For each goal, we replace the anchor with multiple semantically equivalent paraphrases and report mean $\pm$ standard deviation across paraphrases. Directional accuracy and pairwise ranking are invariant, while leakage and diagonal dominance exhibit only minor variation, indicating stable goal-selective representations.

Generator	SignAcc ↑	AUROC ↑	DiagDom ↑	Leak ↓
GPT-4o (train)	0.99 ± 0.01	1.00 ± 0.00	0.88 ± 0.05	0.22 ± 0.03
Gemini	0.98 ± 0.02	1.00 ± 0.00	0.88 ± 0.05	0.24 ± 0.03
Mistral	0.98 ± 0.02	1.00 ± 0.00	0.85 ± 0.05	0.23 ± 0.02
LLaMA-70B	0.94 ± 0.03	0.99 ± 0.01	0.81 ± 0.06	0.24 ± 0.03
DeepSeek	0.93 ± 0.04	0.98 ± 0.02	0.81 ± 0.06	0.29 ± 0.04

Table 7: Cross-generator evaluation of a projection model trained on GPT-4o-mini counterfactuals. Generators include Gemini-3 Flash, Mistral Medium 3.1, LLaMA-3.3-70B-Instruct, and DeepSeek-V3.2. High directional accuracy and goal selectivity persist across model families, indicating robustness to generator-specific artifacts. Values are mean $\pm$ 95% bootstrap CI half-width.

Method	SignAcc ↑	AUROC ↑	DiagDom ↑	Leak ↓
TF–IDF + LR	N/A	0.9999	N/A	N/A
Cosine-Δ + LR	N/A	0.75 ± 0.07	0.38	0.88
Linear Direction Probe	0.94 ± 0.02	N/A	N/A	N/A
Proj–MLP (ours)	**0.99**	**1.00**	**0.88**	**0.22**

Table 8: Simple baselines and probing comparisons. TF–IDF and linear probes recover global polarity but do not define a goal-conditioned embedding geometry. Cosine-Δ regression exhibits substantial cross-goal interference, reflected in low diagonal dominance and high leakage. Our projection-based model is the only method that jointly achieves strong directionality and goal-selective structure.