Do Language Model Embeddings Encode Goal-Conditioned Alignment?
A Counterfactual Geometric Analysis

Anonymous ACL submission

Abstract

Sentence embedding models effectively capture semantic similarity, yet it remains unclear whether they encode directional structure associated with competing objectives. This question matters for applications involving policy reasoning and objective-conditioned retrieval, where topical similarity alone may obscure whether a text advances or obstructs a given objective. We introduce a counterfactual probing framework that isolates goal-conditioned directional shifts using paired texts that preserve factual content while reversing functional orientation relative to a target objective. Using the Sustainable Development Goals (SDG) and Federal Open Market Committee (FOMC) domains, we evaluate whether frozen embedding spaces organize texts along recoverable goal-conditioned directions. We measure directional separation, cross-goal leakage, and asymmetric subspace overlap across multiple embedding architectures. We find that pretrained embeddings contain weak but measurable directional organization prior to adaptation, though this structure remains highly entangled and exhibits substantial off-target interference. Lightweight projection heads trained over frozen embeddings substantially improve directional selectivity without modifying the underlying encoder. Linear projections recover polarity information, while nonlinear projections appear to act as interference filters that reduce cross-goal leakage and recover more selective organization. Directed subspace analyses reveal persistent asymmetric overlap between related objectives, including oppositional structure within the Federal Reserve dual mandate. Our results suggest that goal-conditioned organization in pretrained embedding spaces is partially recoverable, but remains inherently overlapping rather than fully disentangled.

1 Introduction

Sentence embeddings effectively capture semantic similarity, yet their ability to encode *goal-conditioned alignment* remains poorly understood. By goal-conditioned alignment, we refer to whether a text advances or obstructs a specific objective or policy criterion. This directional relationship is distinct from sentiment: texts may positively describe conditions that nevertheless undermine a target objective, or negatively describe conditions that support it.

Disentangling directional alignment from topical similarity is particularly challenging when texts discuss the same subject matter while advocating opposing functional outcomes. For example, policy proposals may concern the same economic intervention while implying conflicting consequences for inflation, employment, or growth. Standard embedding geometries naturally conflate shared semantic structure with goal-conditioned directionality.

In this work, we investigate whether pretrained embedding models intrinsically encode recoverable directional structure. To isolate directional signals from surface semantics, we introduce a counterfactual probing framework that constructs paired texts preserving topic and factual content while reversing their functional orientation toward a target objective. Using this framework, we study two questions: **RQ1:** Do pretrained embedding models encode a reliable directional signal under counterfactual inversion? **RQ2:** To what extent is this structure selectively recoverable across multiple objectives rather than remaining geometrically entangled?

We evaluate these questions across two policy domains: the United Nations Sustainable Development Goals (SDGs) and Federal Reserve policy communications. Across multiple sentence embedding architectures, we find that pretrained representations contain latent but weakly organized directional structure. Lightweight projection heads trained over frozen embeddings substantially improve directional separation and reduce cross-goal interference without modifying the underlying en-

coder.

However, the resulting geometry does not become fully disentangled. Instead, objectives retain structured and asymmetric overlap relationships, including adversarial trade-offs between inflation and employment within the Federal Reserve domain. These findings suggest that goal-conditioned alignment is already partially encoded within pretrained embedding spaces, but exists as an entangled geometric structure rather than as directly isolated features.

2 Related Work

Our work connects four lines of research: counterfactual evaluation, representation probing, embedding geometry, and controllable latent structure.

Counterfactual Evaluation and Controlled Text Inversion. Counterfactual text pairs are widely used to isolate causal or directional signals from superficial correlations in language models. (Kaushik et al., 2019) demonstrate that counterfactually augmented data improves robustness by reducing reliance on spurious lexical cues. More recently, work on controlled language generation has explored whether Large Language Models (LLMs) can reliably produce minimally edited semantic inversions while preserving local coherence and factual consistency (Gat et al., 2023; Youssef et al., 2024). Our work does not use counterfactuals for supervised robustness training. Instead, we treat paired directional inversions as geometric probes that expose how latent embedding spaces organize competing goal-conditioned signals.

Representation Probing and Recoverable Structure. Prior probing work has shown that substantial latent structure can be recovered from frozen neural representations using lightweight linear or low-rank transformations (Alain and Bengio, 2016; Hewitt and Manning, 2019). These studies demonstrate that pretrained representations often encode globally organized information not immediately accessible in the raw coordinate frame. Unlike traditional probing approaches, which typically evaluate downstream classification accuracy over linguistic attributes, we focus on whether frozen embedding spaces encode recoverable goal-conditioned directional organization under controlled counterfactual perturbations.

Embedding Geometry and Anisotropy. Pretrained transformer embeddings are known to exhibit strong anisotropy, concentrating within narrow regions of vector space where a small number of dominant principal components account for a disproportionate fraction of the variance (Ethayarajh, 2019; Timkey and Van Schijndel, 2021). This geometric concentration distorts cosine similarity and contributes to hubness effects, motivating post-processing approaches such as whitening and principal component removal (Su et al., 2021; Ait-Saada and Nadif, 2023). However, while these methods often improve global isotropy, they can simultaneously degrade local semantic structure. Related work further suggests that semantic and functional attributes may correspond to approximately recoverable directions within latent representation spaces (Subramani et al., 2022). Contrastive representation learning objectives additionally encourage a geometry balancing alignment between positive pairs and uniformity across the hypersphere, producing embedding spaces with different structural properties than autoregressive next-token prediction objectives (Wang and Isola, 2020; Gao et al., 2021). Our findings suggest that selectively recovering objective-conditioned structure may require concentrating variance along a smaller number of directional axes rather than uniformly flattening the embedding geometry.

Latent Directionality and Representational Overlap. A growing body of work studies whether high-level semantic attributes correspond to approximately linear or locally recoverable directions within neural representation spaces. Prior research on steering vectors and representation editing demonstrates that latent directions can often be extracted directly from frozen models and manipulated without full parameter fine-tuning (Turner et al., 2023). At the same time, recent work on representational superposition argues that neural systems frequently encode partially overlapping features within shared low-dimensional subspaces rather than as fully disentangled factors (Elhage et al., 2022). Our work differs from these approaches in two respects. First, we evaluate directional organization within sentence embedding spaces rather than autoregressive generation trajectories. Second, rather than focusing solely on the recoverability of individual directions, we study the degree to which objective-conditioned structure remains overlapping, asymmetric, and selectively recoverable across related institutional domains.

3 Counterfactual Probing Framework

We evaluate goal-conditioned directional structure using a controlled counterfactual probing framework. For each text sample, we construct a paired counterfactual that preserves the underlying semantic and factual substrate while reversing its functional orientation relative to a target objective. Measuring the resulting embedding displacement allows us to test whether pretrained embedding spaces encode consistent and selective directional structure beyond broad semantic similarity alone.

Counterfactual pairs are generated using large language models (LLMs) under constrained rewriting rules designed to preserve semantic content while controlling unrelated stylistic variation through generative symmetry. We evaluate the framework across two policy domains: the United Nations Sustainable Development Goals (**SDG**) and Federal Reserve monetary policy objectives (**FOMC**). These domains provide partially overlapping objective systems suitable for studying both directional selectivity and cross-goal interference.

We apply the framework to frozen off-the-shelf (**OTS**) embedding models and evaluate whether lightweight projection heads can recover more selective directional organization without modifying the underlying encoder.

3.1 Counterfactual Probe Data Generation

We construct paired counterfactual datasets using the OSDG (OSDG et al., 2024) corpus together with Federal Reserve meeting notes and policy communications spanning 2006–present. The SDG corpus covers the first 16 Sustainable Development Goals, while the FOMC corpus isolates five primary monetary policy objectives and instruments: employment, inflation, growth, balance sheet policy, and forward guidance.

Counterfactual pairs are generated through a constrained multi-stage LLM pipeline using Gemma3 (12B) for structural parsing and Gemini Flash for controlled directional inversion.

Directional Filtering and Structural Extraction. We first remove texts lacking a clear functional orientation relative to the target objective. Remaining samples are structurally decomposed using principles from Rhetorical Structure Theory (RST), separating the primary goal-conditioned claim (*nucleus*) from supporting contextual material (*satellites*).

Controlled Directional Inversion. The extracted nucleus is rewritten to reverse its directional orientation while preserving the underlying factual substrate. Generation is constrained by four rules: (1) preserve empirical claims, (2) invert evaluative interpretation rather than factual content, (3) avoid introducing new causal arguments, and (4) avoid introducing new policy prescriptions. Supporting satellite content is then updated to remain coherent with the revised nucleus.

Generative Symmetry and Reintegration. To reduce stylistic artifacts introduced by synthetic generation, we additionally paraphrase the original text using the same generation framework while preserving its initial orientation. This equalizes generator-specific characteristics across both sides of the pair while intentionally introducing paraphrastic variation, such that the framework does not rely on minimal lexical edits between originals and counterfactuals. The revised nucleus and satellites are then reintegrated and passed through a final constrained grammatical normalization stage.

The resulting pairs preserve broad semantic and factual structure, approximate length, and generator distribution characteristics while expressing opposing functional orientations relative to the target objective. The relevant evaluation criterion is therefore representational rather than lexical: whether embedding spaces recover selective directional organization from semantically similar but functionally opposed texts.

Dataset Statistics. We apply a stratified $80/10/10$ train-validation-test split directly over matched original-counterfactual pairs to preserve pairwise symmetry and prevent cross-split contamination.

The SDG corpus contains $2{,}481$ paired examples distributed across 16 goals ($1{,}992$ train / 241 validation / 248 test). The FOMC corpus contains 552 paired examples spanning 5 policy objective categories (439 train / 55 validation / 55 test). In total, the framework contains $3{,}033$ counterfactual pairs ($6{,}066$ individual documents). Complete per-category statistics are provided in Appendix B.

3.2 Coordinate Geometry and Probing Metrics

Let $\mathbf{h}_i \in \mathbb{R}^d$ denote the embedding of an original text produced by a frozen encoder $f(x)$, and let $\mathbf{h}_i^{\text{cf}}$ denote its corresponding counterfactual embedding. We evaluate these pairs in both the raw *base space*

($\mathcal{H} \subseteq \mathbb{R}^d$) and a learned *projected space* ($\mathcal{Z} \subseteq \mathbb{R}^{d'}$).

To track directional organization, we define a set of objective anchors $\mathbf{a}_k$. In the base space, anchors are empirical centroids computed over all samples associated with objective k:

$$\mathbf{a}k = \frac{1}{|I_k|} \sum i \in I_k \mathbf{h}_i \tag{1}$$

where I_k denotes the index set for objective k. In the projected space, anchors are learned jointly with the projection head. All proximity comparisons use cosine similarity:

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}|_2, |\mathbf{v}|_2} \tag{2}$$

To characterize directional separation and cross-goal interactions, we evaluate four complementary geometric diagnostics:

Directional Target Margin. We measure directional separation by comparing the relative similarity of each sample and its counterfactual counterpart to the target anchor:

$$\text{TargetMargin} = \frac{1}{N} \sum_{i=1}^{N} \left[\cos(\mathbf{z}_i, \mathbf{a}_{y_i}) - \cos(\mathbf{z}_i^{\text{cf}}, \mathbf{a}_{y_i}) \right] \tag{3}$$

where y_i denotes the objective label of sample i, and $\mathbf{z}$ denotes the evaluated representation ($\mathbf{h}$ in the base space or $\mathbf{z}$ in the projected space). Positive values indicate that the original sample is positioned closer to its target anchor than its counterfactual inversion.

Cross-Goal Leakage. To measure off-target interference, we define the alignment differential induced by inversion:

$$\Delta_{i,j} = \cos(\mathbf{z}i, \mathbf{a}j) - \cos(\mathbf{z}i^{\text{cf}}, \mathbf{a}j) \tag{4}$$

Let Mg, j denote the mean absolute alignment differential toward anchor j across samples belonging to goal g. Leakage is defined as the ratio between the average off-diagonal and diagonal entries of the resulting shift matrix:

$$\text{Leakage} = \frac{\frac{1}{K(K-1)} \sum g \neq jMg, j}{\frac{1}{K} \sum_g Mg, g} \tag{5}$$

where K denotes the total number of objectives. Low leakage indicates that directional shifts remain localized to their intended objective.

Diagonal Dominance. We measure the relative dominance of target-aligned over off-target directional shifts:

$$\text{DiagDom} = \frac{1}{K} \sum_{g=1}^{K} \left(M_{g,g} - \frac{1}{K-1} \sum_{j \neq g} M_{g,j} \right) \tag{6}$$

High diagonal dominance indicates that directional inversions primarily affect their intended objective while remaining comparatively stable across off-target regions.

Directed Subspace Overlap. To measure asymmetric overlap between goal-conditioned regions, we evaluate how well the dominant variance structure of one region explains another. Let $\mathbf{Z}_A$ and $\mathbf{Z}_B$ denote centered coordinate matrices associated with goals A and B.

We compute a low-rank principal component basis $\mathbf{L}_A \in \mathbb{R}^{d \times k}$ over $\mathbf{Z}_A$ and define the associated projection operator:

$$\mathbf{P}_A = \mathbf{L}_A \mathbf{L}_A^\top \tag{7}$$

Projecting $\mathbf{Z}_B$ into the dominant subspace of A, we define the directed subspace overlap as:

$$\text{DSO}(B \to A) = \frac{\text{Var}\left(\mathbf{Z}_B \mathbf{P}_A\right)}{\text{Var}(\mathbf{Z}_B)} \tag{8}$$

High overlap indicates that the dominant subspace associated with goal A explains a substantial fraction of the variance associated with goal B. Because the projection is directional, $\text{DSO}(A \to B)$ and $\text{DSO}(B \to A)$ are not necessarily symmetric.

3.3 Latent Space Projection Head

To recover more selective directional organization from frozen embeddings, we append a lightweight three-layer multi-layer perceptron (**MLP**) projection head over the encoder output while keeping the underlying encoder fixed throughout optimization. The projection head therefore cannot introduce new information into the representation space, but only reorganize information already encoded within the pretrained embeddings.

Given a base embedding $\mathbf{h}_i \in \mathbb{R}^d$, the projection head maps the representation into a projected coordinate space $\mathbf{z}_i \in \mathbb{R}^{d'}$, which is subsequently optimized under the Differential Alignment objective described below.

3.4 Differential Alignment (DA) Loss

To shape the projected space without enforcing strict orthogonality between objectives, we introduce the **Differential Alignment (DA)** loss. The objective operates on relative alignment differentials computed over cosine-normalized representations, jointly optimizing the projection head together with a set of learnable goal anchors $\mathbf{A} \in \mathbb{R}^{K \times d'}$.

Given normalized original embeddings $\tilde{\mathbf{Z}}^{+} \in \mathbb{R}^{N \times d'}$ and counterfactual embeddings $\tilde{\mathbf{Z}}^{-} \in \mathbb{R}^{N \times d'}$, we compute the corresponding similarity matrices against all normalized anchors $\tilde{\mathbf{A}}$:

$$\mathbf{S}^{+} = \tilde{\mathbf{Z}}^{+}\tilde{\mathbf{A}}^{\top}, \qquad \mathbf{S}^{-} = \tilde{\mathbf{Z}}^{-}\tilde{\mathbf{A}}^{\top} \tag{9}$$

The optimization is governed by the resulting differential matrix:

$$\mathbf{\Delta} = \mathbf{S}^{+} - \mathbf{S}^{-} \tag{10}$$

where each element $\Delta_{n,c}$ measures the relative alignment differential between sample n and anchor c.

Target Alignment Objective. For the ground-truth objective y_i, we maximize directional separation using a temperature-scaled soft-margin objective:

$$\mathcal{L}\text{target} = -\frac{1}{N}\sum i = 1^{N} \log \sigma\left(\frac{\Delta_{i,y_i}}{\tau}\right) \tag{11}$$

where τ controls the sharpness of the alignment margin.

Off-Target Regularization. To suppress diffuse cross-goal interference without imposing hard orthogonality constraints, we additionally regularize off-target alignment differentials:

$$\mathcal{L}\text{off} = \frac{1}{N(K-1)}\sum i = 1^{N} \sum_{c \neq y_i} \log!\left(1 + \Delta_{i,c}^{2}\right) \tag{12}$$

Unlike a standard quadratic penalty, the heavy-tailed log-relaxation grows sublinearly for large deviations, suppressing weak background interference while permitting stable overlap between objectives when strongly supported by the underlying representations.

The complete objective is given by:

$$\mathcal{L}\text{total} = \mathcal{L}\text{target} + \lambda \mathcal{L}_{\text{off}} \tag{13}$$

where λ controls the strength of the off-target regularization.

Because the optimization operates on sample-to-anchor alignment differentials rather than symmetric pairwise distances between goals, the resulting geometry is not constrained to be reciprocal. This allows the learned space to preserve asymmetric overlap relationships present in the underlying representations.

4 Experimental Setup

To evaluate how directional structure varies across pretraining paradigms, we study four embedding backbones spanning causal language modeling and contrastive retrieval objectives: two decoder-only models from the `Harrier-OSS` family (`270M`, `0.6B`) and two encoder-only retrieval models from the `E5` family (`E5-Large-V2`, `E5-Unsupervised`). All models are evaluated strictly off-the-shelf with frozen encoder weights, using representations extracted from the final pooling layers of the Hugging Face implementations.

All documents are embedded without instruction prefixes or goal-conditioning prompts. Because modern embedding models are highly sensitive to prompt steering, introducing explicit task instructions would confound latent representational structure with in-context conditioning effects. The resulting metrics therefore probe directional organization already present within the pretrained embeddings.

4.1 Optimization & Evaluation Protocol

The projection head is implemented as a lightweight three-layer MLP jointly optimized with a set of learned goal anchors $\mathbf{A} \in \mathbb{R}^{K \times d'}$ using AdamW. Training is performed for 20 epochs with batch size 128, learning rate 10^{-4}, weight decay 0.01, and dropout 0.1. For the Differential Alignment objective, we fix $\tau = 0.2$ and $\lambda = 1$.

All macro-level geometric metrics are evaluated using paired bootstrap resampling with $B = 1{,}000$ iterations. Resampling is applied directly over matched original-counterfactual pairs to preserve pairwise structure. Reported scores correspond to the bootstrap mean together with their associated 95% confidence interval half-widths.

5 Results and Analysis

We evaluate whether the learned projection reorganizes frozen embedding spaces into more direction-

Model Framework / Space	Target Margin ↑		Diag. Dom. ↑		Leakage ↓	
	Base	Proj	Base	Proj	Base	Proj
Domain: UN Sustainable Development Goals (SDG)						
Harrier OSS 0.6b	$.03_{\pm.00}$	$\mathbf{.71}_{\pm.05}$	$.20_{\pm.05}$	$\mathbf{.73}_{\pm.05}$	$.84_{\pm.06}$	$\mathbf{.17}_{\pm.01}$
Harrier OSS 270m	$.02_{\pm.01}$	$.60_{\pm.05}$	$.25_{\pm.05}$	$.71_{\pm.06}$	$1.01_{\pm.11}$	$.21_{\pm.02}$
E5-Large-V2	$.01_{\pm.00}$	$.60_{\pm.05}$	$.40_{\pm.06}$	$.68_{\pm.06}$	$.77_{\pm.09}$	$.19_{\pm.02}$
E5-Unsupervised	$.02_{\pm.00}$	$.61_{\pm.05}$	$.24_{\pm.05}$	$.72_{\pm.06}$	$.92_{\pm.08}$	$\mathbf{.17}_{\pm.02}$
Domain: FOMC Central Bank Mandates						
Harrier OSS 0.6b	$.01_{\pm.01}$	$\mathbf{.89}_{\pm.10}$	$.42_{\pm.14}$	$.82_{\pm.11}$	$.94_{\pm.56}$	$\mathbf{.30}_{\pm.07}$
Harrier OSS 270m	$.01_{\pm.01}$	$.83_{\pm.08}$	$.40_{\pm.13}$	$\mathbf{.84}_{\pm.11}$	$.96_{\pm.54}$	$.37_{\pm.08}$
E5-Large-V2	$.01_{\pm.00}$	$.87_{\pm.09}$	$.55_{\pm.13}$	$.80_{\pm.11}$	$.45_{\pm.36}$	$.33_{\pm.09}$
E5-Unsupervised	$.02_{\pm.01}$	$.76_{\pm.10}$	$.38_{\pm.13}$	$.75_{\pm.11}$	$.85_{\pm.41}$	$.35_{\pm.10}$

Table 1: Macro-level geometric metrics across evaluated embedding configurations. Subscripts denote 95% bootstrap confidence interval half-widths. Target Margin measures the signed directional separation between original and counterfactual samples relative to their target anchors. Bold values indicate the strongest projected performance within each domain.

ally selective and geometrically coherent regions. Alongside the quantitative metrics, we visualize the resulting geometry using UMAP to provide qualitative intuition regarding large-scale structural changes.

5.1 Directional Reorganization Under Projection

Introducing the learned projection produces consistent improvements across all evaluated architectures and domains (Table 1). Projection simultaneously increases directional separation while reducing cross-goal interference.

For `Harrier-OSS-0.6B` in the SDG domain, Target Margin increases from .03 to .71, Diagonal Dominance from .20 to .73, and Leakage decreases from .84 to .17. Similar improvements appear across all evaluated architectures, including the specialized FOMC domain.

To characterize the associated structural changes, we analyze the singular value spectrum of the evaluation representations extracted from `Harrier-OSS-0.6B`. Following projection, the variance explained by the leading principal component increases from 7.5% to 22.2% in SDG and from 18.3% to 31.8% in FOMC. Rather than flattening the representation geometry, the learned projection concentrates variance along a smaller number of dominant directional axes while reducing diffuse cross-goal interference.

These quantitative changes are reflected qualitatively in Figures 1 and 2. In the projected space, positive and counterfactual distance distributions separate more cleanly, while the embedding geometry reorganizes from diffuse topic-centered clusters into more localized goal-conditioned regions surrounding the learned anchors.

Importantly, several base models already exhibit moderate directional sensitivity while still maintaining substantial cross-goal interference. The projection therefore improves not only directional accessibility, but also the selectivity and concentration of the resulting geometry.

5.2 Architectural Divergence Across Policy Domains

The SDG and FOMC domains exhibit substantially different native directional organization. Within the SDG corpus, all evaluated base models exhibit moderate directional accessibility, suggesting that broad development-oriented policy language already induces partially recoverable goal-conditioned structure.

In contrast, the specialized FOMC domain produces a clear architectural split. The decoder-only `Harrier` models exhibit weak baseline organization, with near-zero Target Margins and elevated cross-goal leakage, whereas the encoder-only `E5` models retain substantially stronger directional separation prior to projection.

These differences suggest that directional accessibility depends not only on domain structure, but also on the underlying pretraining objective. The contrastive bidirectional objectives used by the `E5` models appear to preserve more linearly accessible structure within specialized institutional language, whereas the causal decoder representations require additional restructuring before comparable selectivity emerges. Following projection, however, all evaluated models converge toward substantially im-

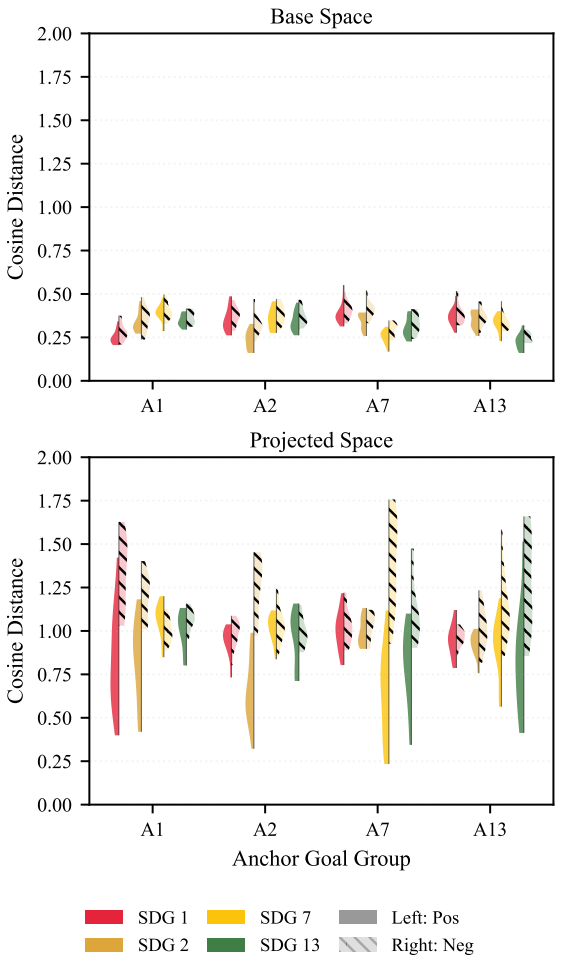

Figure 1: Pairwise cosine distance distributions for selected a SDG subset evaluated in the base space (top) and projected space (bottom). For each split violin, the solid left half corresponds to the positive sample distribution, while the hatched right half corresponds to the counterfactual distribution.

proved directional separation and reduced cross-goal interference.

5.3 Linear Accessibility vs Nonlinear Selectivity

To determine whether the recovered directional structure is fundamentally linear, we replace the nonlinear projection head with a single linear transformation while preserving the same optimization objective (Table 2). The linear projection recovers substantial directional separation in both domains, indicating that goal-conditioned signals are partially linearly accessible within the frozen embeddings. However, it fails to adequately suppress cross-goal leakage, particularly in the SDG domain.

This pattern suggests that goal-conditioned regions remain partially superposed within overlapping representational subspaces. Under this interpretation, the nonlinear projection acts less as a feature creator than as an interference filter, separating coupled directional structure that cannot be cleanly isolated through linear projection alone (Elhage et al., 2022).

Projection	Margin ↑	DiagDom ↑	Leakage ↓
SDG			
Linear	.38	.42	.53
3L MLP	**.71**	**.73**	**.17**
FOMC			
Linear	.70	.69	.53
3L MLP	**.89**	**.82**	**.30**

Table 2: Linear projection ablation on `Harrier-OSS-0.6B`. A single linear layer recovers substantial directional structure but retains considerably higher cross-goal leakage than the nonlinear projection head.

5.4 Asymmetric Goal Overlap Structure

Although projection substantially improves directional selectivity, the resulting geometry does not converge toward fully independent goal axes. Figure 3 shows that several objectives retain asymmetric overlap relationships, where the dominant variance structure of one objective explains substantially more of another objective's local variance than vice versa.

Rather than uniformly suppressing cross-goal structure, the learned projection reduces diffuse interference while preserving concentrated regions of shared organization. In the SDG domain, overlap remains strongest among closely related socioeconomic objectives. In the FOMC domain (Appendix A), inflation-, employment-, and growth-oriented regions remain strongly coupled following projection, whereas balance sheet operations remain comparatively isolated.

Notably, inflation and employment exhibit not only overlap but directional opposition: positive and counterfactual distributions shift in negatively correlated directions relative to their anchors. This suggests that the projected space partially captures the trade-off structure embedded in the Federal Reserve's dual mandate rather than representing objectives as fully independent dimensions.

Overall, the learned projection appears to amplify structured directional organization already latent within the frozen representations rather than enforcing artificial orthogonality between objec-

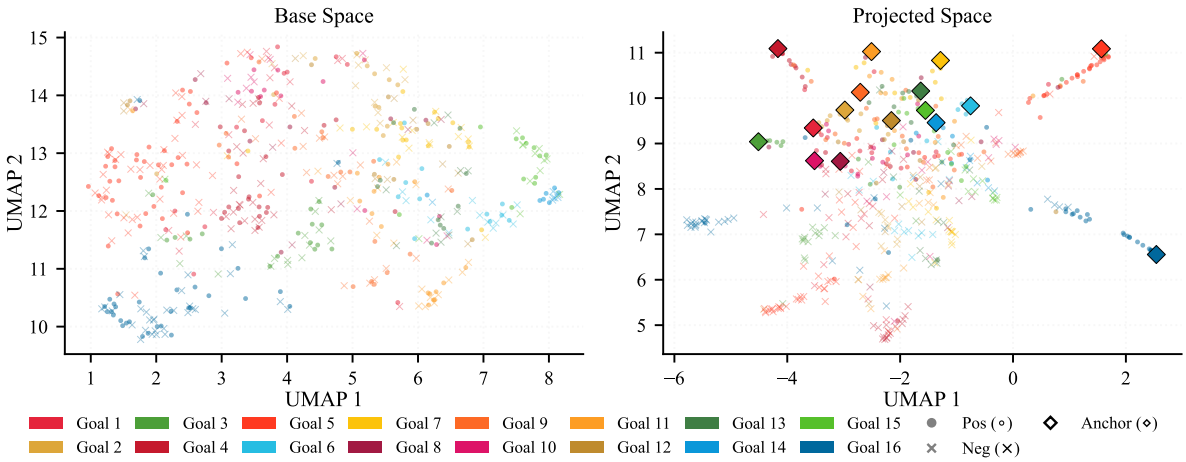

Figure 2: Two-dimensional UMAP visualizations of the SDG embedding space across the base parameters (left) and projected parameters (right). Circle markers (∘) denote positive samples, cross markers (×) denote counterfactual samples, and diamonds (⋄) indicate learned anchor coordinates.

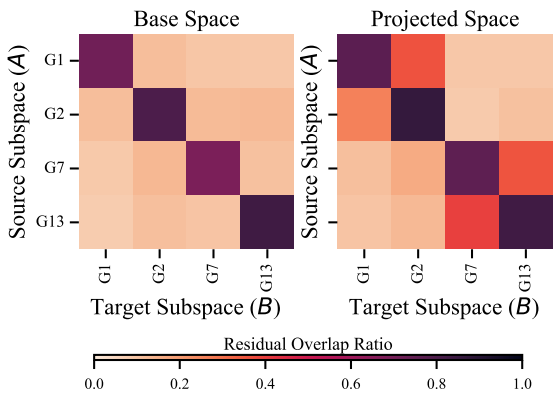

Figure 3: Directed subspace overlap matrices for SDG 1, 2, 7 & 13 across the base space (top) and projected space (bottom). The vertical axis denotes the source subspace (A) and the horizontal axis denotes the target subspace (B). Darker cells indicate stronger shared geometric structure.

tives. These overlap patterns remain stable across PCA rank choices (Appendix A.4).

6 Conclusion

We introduced a counterfactual probing framework for evaluating whether pretrained embedding models encode goal-conditioned directional structure within their latent representations. By constructing tightly controlled counterfactual pairs that preserve factual and topical content while reversing functional orientation, we isolate directional geometric shifts across both cooperative and adversarial policy domains.

Our findings show that frozen embedding spaces contain weak but measurable directional structure prior to adaptation. However, this organization is diffuse and poorly selective in the raw representation space. While linear projections recover substantial directional separation, they retain high levels of cross-goal leakage. In contrast, lightweight nonlinear projection heads substantially improve directional selectivity and reduce off-target interference without modifying the underlying encoder weights.

Importantly, the recovered geometry does not converge toward fully independent objective axes. Directed subspace overlap analyses instead reveal persistent asymmetric dependencies between objectives, suggesting that pretrained representations encode partially shared and hierarchical directional structure rather than strictly orthogonal dimensions. In the SDG domain, this overlap frequently reflects cooperative socioeconomic relationships, whereas the FOMC setting surfaces explicit trade-off structure between institutional objectives.

Taken together, these results suggest that pretrained embedding models natively encode latent directional organization that can be selectively amplified through lightweight geometric restructuring. Rather than emerging as isolated symbolic features, goal-conditioned directions appear embedded within a broader, partially overlapping representational geometry shaped by the statistical structure of the pretraining corpus.

7 Limitations

Several limitations should be considered when interpreting the results of this study.

LLM-generated counterfactuals. Our framework relies on counterfactual text pairs generated using large language models. Although the generation pipeline constrains factual preservation and minimizes topical drift, directional inversion remains inherently non-unique. Multiple valid rewrites may reverse the functional orientation of a text while preserving its semantic content. In addition, generated samples may contain stylistic regularities specific to the underlying models. While our generative symmetry procedure reduces these confounds, we cannot fully eliminate model-specific artifacts. Importantly, the goal of the framework is not to establish canonical inversions, but to measure whether embedding spaces respond consistently to controlled directional perturbations.

Objective framing and institutional scope. Our evaluation domains are grounded in institutional policy frameworks, specifically the United Nations Sustainable Development Goals and Federal Open Market Committee mandates. These domains provide structured and operationally well-defined objective systems, but they do not capture the full diversity of moral, political, or culturally situated normative reasoning. The resulting directional structure should therefore be interpreted as goal-conditioned organization within these specific institutional contexts rather than as a general theory of human values or normative cognition.

Representation-level interpretation. The observed geometric patterns reflect properties of embedding representations rather than explicit symbolic reasoning or causal understanding. Recoverable directional organization within a latent space does not necessarily imply that the underlying model possesses coherent normative beliefs, stable preferences, or internally consistent world models. Our analysis is therefore diagnostic rather than interpretive: it evaluates representational structure without making broader claims about model intentionality or reasoning capacity.

Language and domain coverage. All experiments are conducted on English-language policy and institutional texts. The resulting findings may not generalize to multilingual settings, informal discourse, or domains where directional signaling is expressed more implicitly or culturally variably. Extending the framework across languages, discourse styles, and non-institutional objective systems remains an important direction for future work.

8 Ethical Considerations

The methods introduced in this work could potentially be misused in political, institutional, or governance settings to monitor ideological conformity, suppress dissent, or privilege dominant value frameworks under the appearance of quantitative neutrality. Because the framework measures directional organization relative to predefined institutional objectives, any downstream deployment would inherit the assumptions and biases embedded within those objective systems.

In addition, the counterfactual texts used throughout this study are synthetic analytical constructions generated for controlled geometric evaluation. These inversions do not represent the views of the authors and should not be interpreted as endorsements of the positions expressed within the generated samples.

Our work is intended strictly as a diagnostic study of representation geometry rather than a normative scoring, surveillance, or decision-making system. We therefore encourage caution before deploying embedding-based directional analyses in real-world contexts.

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.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on*

empirical methods in natural language processing, pages 6894–6910.

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*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

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*.

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

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

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581.

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*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR.

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 Additional FOMC Geometric Diagnostics

This appendix provides additional geometric diagnostics for the FOMC objective domain discussed in Section 5.4. Unlike the SDG domain, the projected FOMC representations retain substantial overlap between institutionally coupled objectives, particularly among inflation-, employment-, and growth-oriented regions.

A.1 FOMC Anchor Distance Distributions

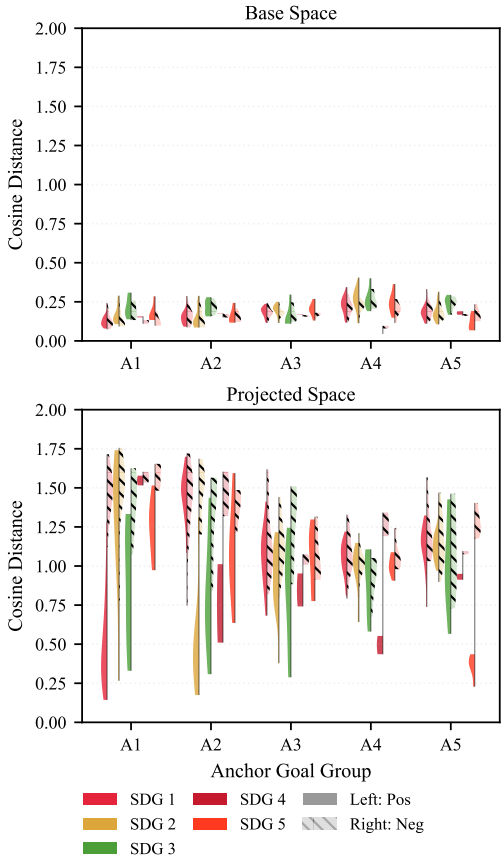

Figure 4: Anchor-relative cosine distance distributions for selected FOMC objectives evaluated under the frozen base space (left) and projected space (right). Following projection, positive and counterfactual distributions become substantially more separated while preserving structured overlap between inflation-, employment-, and growth-oriented regions. The inflation and employment objectives additionally exhibit inverse directional displacement patterns consistent with the trade-off structure of the Federal Reserve dual mandate.

A.2 FOMC UMAP Visualizations

A.3 FOMC Directed Subspace Overlap

A.4 DSO Rank Sensitivity

DSO Rank Sensitivity. We evaluated Directed Subspace Overlap (DSO) across PCA ranks $k \in 4, 8, 10, 16, 32$, using k=10 in the primary experiments. Pairwise overlap structure remained highly stable across ranks (Spearman $\rho = .89$ at k=4 and

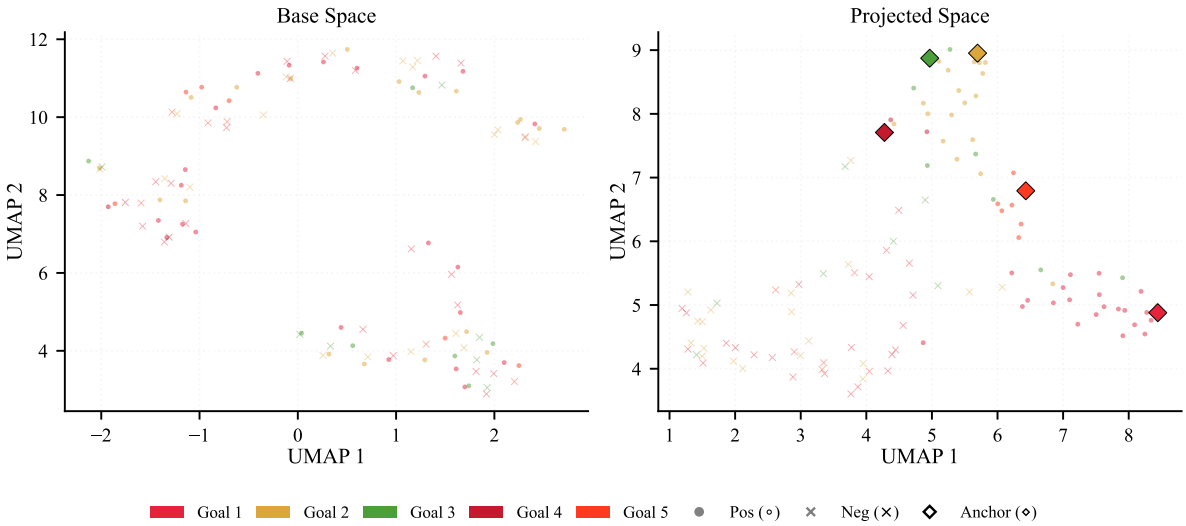

Figure 5: Two-dimensional UMAP visualizations of the FOMC embedding space under the frozen base space (left) and projected space (right). Circle markers (∘) denote positive statements, cross markers (×) denote counterfactual statements, and diamonds (⋄) denote learned anchor coordinates. Unlike the SDG domain, the projected FOMC space retains substantial overlap between macroeconomic objective regions, particularly among inflation-, employment-, and growth-oriented statements.

$\rho \geq .97$ for $k \geq 8$ relative to the primary setting), while the magnitude of directional asymmetry remained nearly constant (.040–.049). These results suggest that the observed asymmetric overlap patterns are not artifacts of a specific low-rank decomposition choice.

B Extended Dataset Characterization

Table 3 provides the fully granulated, category-by-category sample distribution for the generated counterfactual text pairs across the training, validation, and testing splits.

Compute. All experiments were conducted using PyTorch and CUDA on a Nvidia RTX 5090.

Domain / Target Policy Classification Objective	Train	Val	Test	Total Pairs
Domain: UN Sustainable Development Goals (SDG)				
G_1: No Poverty	104	13	14	131
G_2: Zero Hunger	85	10	12	107
G_3: Good Health and Well-being	118	14	16	148
G_4: Quality Education	172	21	23	216
G_5: Gender Equality	324	40	41	405
G_6: Clean Water and Sanitation	92	11	13	116
G_7: Affordable and Clean Energy	123	15	16	154
G_8: Decent Work and Economic Growth	68	8	9	85
G_9: Industry, Innovation and Infrastructure	145	18	19	182
G_{10}: Reduced Inequalities	91	11	12	114
G_{11}: Sustainable Cities and Communities	124	15	17	156
G_{12}: Responsible Consumption and Production	70	8	10	88
G_{13}: Climate Action	77	9	11	97
G_{14}: Life Below Water	44	5	7	56
G_{15}: Life on Land	93	11	13	117
G_{16}: Peace, Justice and Strong Institutions	262	32	34	328
Domain: FOMC Central Bank Mandates				
M_1: Price Stability (Inflation Management)	165	21	21	207
M_2: Maximum Sustainable Employment	153	19	19	191
M_3: Interest Rate Policy	58	7	7	72
M_4: Financial System Regulation	17	2	2	21
M_5: Macroeconomic Forecasting	46	6	6	58
Total Comprehensive Evaluation Corpus	**2,431**	**296**	**303**	**3,030**

Table 3: Granular structural distribution of generated text pairs across training, validation, and testing partitions. Each integer represents an explicit structural pair containing an original policy passage and its corresponding normatively inverted counterfactual twin.

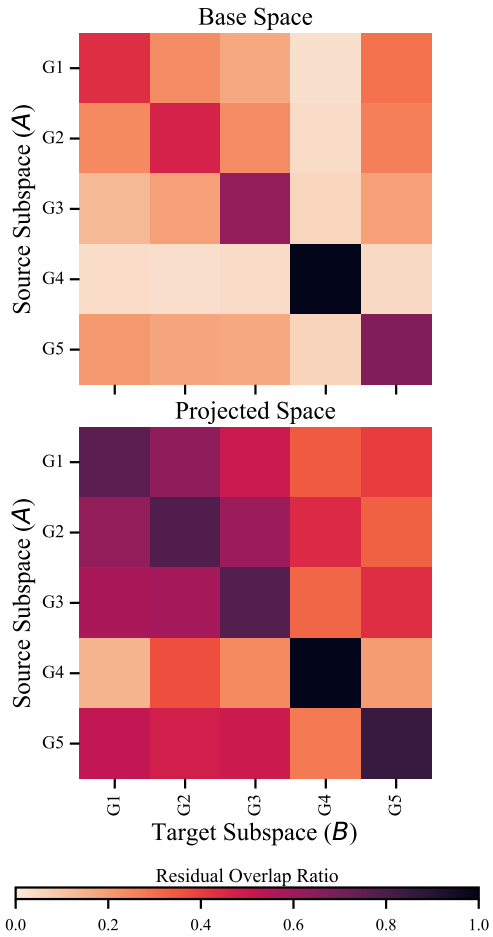

Figure 6: Directed subspace overlap matrices computed between localized FOMC objective regions for the base space (top) and projected space (bottom). Inflation, employment, and growth retain substantial shared structure following projection, whereas balance sheet operations remain comparatively isolated from the broader macroeconomic objective manifold.