Structuring Semantic Embeddings for Principle Evaluation: A Kernel-Guided Contrastive Learning Approach

Anonymous authors
Paper under double-blind review

Abstract

Reliable post-hoc principle evaluation—verifying whether generated text adheres to predefined human values such as safety, fairness, or helpfulness—is a critical bottleneck in AI alignment. While general-purpose text embeddings are widely deployed for this task, they inherently struggle with fine-grained principle distinctions due to severe feature entanglement. Texts sharing similar vocabulary but representing diametrically opposed principles often collapse into the same representation space, blurring critical decision boundaries. To overcome this limitation without the prohibitive costs of full-parameter fine-tuning, we introduce Kernel-Guided Contrastive Learning (KGCL), a framework that transitions the evaluation paradigm from generic semantic approximation to explicit decision boundary sculpting. Operating as a lightweight module atop frozen generalist encoders, KGCL projects entangled embeddings into a structured, principle-aligned subspace. We mathematically prove that our composite objective establishes a defined geometric margin and establishes strict bounds on geometric clustering metrics. Extensive experiments validate these theoretical guarantees, demonstrating that KGCL dramatically enhances the linear separability of highly confusable classes and provides a geometric shield against majority collapse. Remarkably, our explicitly optimized embeddings not only achieve absolute F1 improvements of up to 19.4% over task-agnostic contrastive baselines but also consistently outperform the implicit in-context reasoning of massive generative Large Language Models. Ultimately, KGCL establishes that targeted geometric sculpting provides a highly discriminative, computationally efficient paradigm for robust principle alignment.

1 Introduction

Ensuring that text representations can reliably capture complex, human-defined principles—such as fairness, honesty, and safety—remains a fundamental challenge for AI alignment and robust natural language understanding (Weidinger et al., 2021; Bommasani et al., 2021; Hendrycks et al., 2023). While much of the alignment literature focuses on steering model behavior *during* text generation (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022), an equally important yet comparatively understudied problem is how to determine whether a generated text adheres to such principles *after* it has been produced. This problem, commonly referred to as *post-hoc evaluation*, is central to applications such as content moderation, safety auditing, and large-scale monitoring of model outputs (Gehman et al., 2020; Rae et al., 2021).

At scale, post-hoc evaluation is typically built on high-dimensional text embeddings (Reimers & Gurevych, 2019; Neelakantan et al., 2022). Recent advances have produced large general-purpose embedding models, such as NV-Embed and GTE-Qwen2, with billions of parameters (Li et al., 2023; Lee et al., 2024). Trained on diverse corpora, these models aim to learn universal semantic representations and achieve strong performance on broad benchmarks such as MTEB (Muennighoff et al., 2023). Consequently, they are widely used as frozen feature extractors for downstream principle evaluation.

Limitations of Existing Methods Despite their strong general-purpose performance, these embedding models are not designed to resolve the fine-grained distinctions required for principle evaluation. Their objective of capturing broad semantic similarity often leads to representations that smooth over subtle but

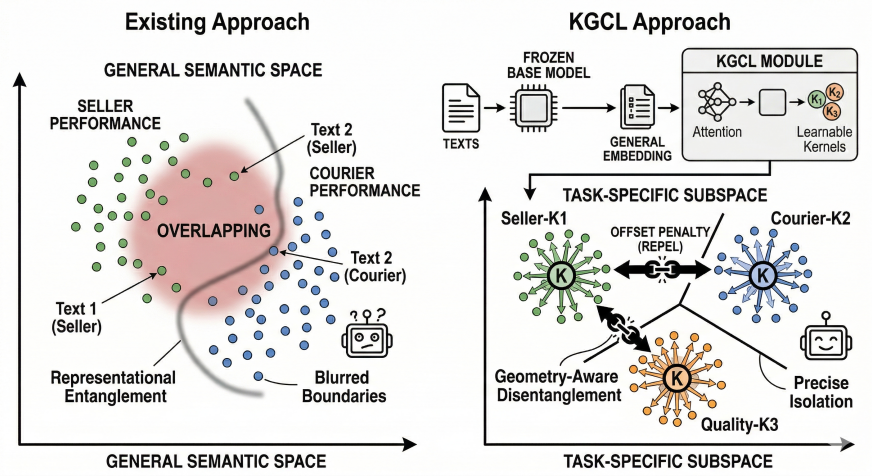


Figure 1: Conceptual comparison between standard general-purpose embeddings and the proposed KGCL framework. **Left:** In the general semantic space, texts sharing similar lexicon but representing distinct principles (e.g., negative reviews about seller performance versus courier performance) suffer from severe representational entanglement, leading to blurred decision boundaries. **Right:** Operating atop a frozen base model, the KGCL module utilizes an attention mechanism to project these general embeddings into a task-specific subspace. Guided by learnable prototype kernels and an explicit offset penalty that actively repels confusable samples, KGCL achieves geometry-aware disentanglement and precise topological isolation.

critical differences between texts. In many practical scenarios, texts can be highly similar at the lexical and semantic levels while reflecting fundamentally different evaluation principles. For example, in e-commerce moderation, the reviews "*The package never arrived, terrible seller*" and "*The package never arrived, terrible courier*" share nearly identical surface semantics but correspond to distinct evaluation targets. Similarly, in content safety, a non-toxic expression such as "*I absolutely hate this garbage situation*" can be lexically similar to a toxic attack like "*I absolutely hate you, you are garbage*" (Gehman et al., 2020; Welbl et al., 2021). Standard embedding models—including those refined via unsupervised contrastive learning such as SimCSE (Gao et al., 2021)—often map such examples to overlapping regions of the representation space. We refer to this phenomenon as *representational entanglement*: semantically similar texts that correspond to different principles are not well separated geometrically. This entanglement blurs decision boundaries and makes it difficult for lightweight downstream classifiers to reliably capture the subtle, context-dependent signals required for accurate principle evaluation (Devlin et al., 2019; Zhou & Srikumar, 2022).

Challenges for Tackling the Limitations Addressing representational entanglement is challenging. Two common approaches are (i) adapting the embedding model via fine-tuning, or (ii) bypassing embeddings altogether by leveraging large language models (LLMs) through few-shot prompting (Qiu et al., 2020; Brown et al., 2020). However, both approaches have notable limitations in this setting. Fine-tuning-based methods, including Parameter-Efficient Fine-Tuning (PEFT) techniques such as LoRA (Hu et al., 2022), still require substantial labeled data and computational resources. More importantly, they optimize model parameters for overall task performance without explicitly enforcing the geometric structure needed to separate entangled representations, often resulting in weak or unstable decision boundaries. On the other hand, leveraging in-context learning (ICL) with large LLMs for evaluation introduces significant computational overhead (Dong et al., 2024). Using generative models for inherently discriminative tasks leads to high latency and cost, making such approaches difficult to deploy at scale (Zheng et al., 2023; Wang et al., 2023). Taken together, these limitations highlight a key gap: the lack of lightweight, data-efficient methods that can directly reshape the geometry of representation spaces to disentangle semantically similar but logically distinct texts—without modifying the underlying large models or relying on expensive generative inference.

Proposed Method To address this gap, we propose *Kernel-Guided Contrastive Learning (KGCL)*, a lightweight and plug-and-play framework for explicitly reshaping the geometry of text representations for principle evaluation. KGCL operates on top of frozen base embeddings and projects them into a low-dimensional, task-specific subspace through an attention-based transformation. In this subspace, we introduce a set of learnable *prototype kernels* that serve as geometric anchors for different evaluation principles. To separate entangled representations, we further design a geometry-aware contrastive objective augmented with an *offset penalty*. This objective encourages texts associated with the same principle to cluster around their corresponding prototypes, while pushing apart samples that are semantically similar but belong to different principles. By explicitly imposing task-relevant geometric structure on the representation space, KGCL transforms general-purpose embeddings into principle-aware representations, enabling more accurate and efficient post-hoc evaluation.

Contributions Our contributions are threefold. First, we propose a lightweight and modular architecture that maps general-purpose text embeddings into a structured subspace tailored for principle evaluation, without requiring any updates to the underlying large-scale encoder. Second, we introduce a geometry-aware training objective that explicitly separates semantically similar but principle-distinct texts, thereby mitigating representational entanglement and yielding more discriminative representations. Third, extensive experiments across diverse datasets show that KGCL consistently outperforms standard embedding-based baselines and achieves performance competitive with LLM-based evaluators, while preserving the efficiency of lightweight downstream classifiers.

2 Related Work

General-Purpose Text Embeddings and Subspace Learning. Massive general-purpose embedding models have become the foundational paradigm for text representation. However, while these models excel at capturing broad semantic context, they frequently conflate subtle, principle-specific distinctions (Devlin et al., 2019; Zhou & Srikumar, 2022). To extract task-relevant features from these entangled spaces, prior works have explored subspace learning and task-specific projections (Fukumizu et al., 2003; Edelman & Intrator, 1997; Peng et al., 2019). Yet, general dimensionality reduction techniques, such as UMAP (McInnes et al., 2018) or Spectral Embedding (Von Luxburg, 2007), are primarily optimized for structural visualization rather than explicitly disentangling predefined, principle-specific features. Similarly, while geometric embeddings successfully structure representation spaces for tasks like knowledge graph querying (Ren et al., 2020), they are not tailored for the nuanced semantic isolation required in natural language. Our framework addresses this by learning a structured, low-dimensional subspace specifically optimized for principle evaluation, actively separating task-relevant signals from the general semantic basis.

Parameter-Efficient Fine-Tuning (PEFT). To adapt massive pre-trained models for specific downstream tasks, Parameter-Efficient Fine-Tuning (PEFT) methods, such as Low-Rank Adaptation (LoRA) (Hu et al., 2022), have been widely adopted to mitigate the prohibitive costs of full fine-tuning. While these techniques successfully reduce the number of trainable parameters, they inherently suffer from deployment bottlenecks in high-throughput post-hoc evaluation scenarios. Because PEFT integrates task-specific weights directly into the base architecture, evaluating a single text against multiple distinct principles requires repeated forward passes of the massive model (Ding et al., 2022; Sheng et al., 2023). Furthermore, these methods rely on implicit parameter updates rather than explicitly remodeling the geometric topology of the representation space to isolate confusable principles. In contrast, our approach operates entirely independently of the frozen base model's internal weights, achieving geometric disentanglement at a fraction of the computational and inference overhead.

Principle Alignment and LLM-Based Evaluation. Prior efforts in principle alignment predominantly focus on constraining language models *during* the text generation process, utilizing techniques such as Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) and Constitutional AI (Bai et al., 2022). However, reliably evaluating the principle alignment of already generated text—known as post-hoc evaluation—remains a distinct and critical challenge. Recently, a prevailing trend has been to leverage the In-Context Learning capabilities of massive LLMs to act as evaluators (Zheng

et al., 2023). While effective, invoking expensive generative models for purely discriminative evaluation tasks incurs prohibitive inference costs and latency, underutilizing the LLMs' generative strengths (Wang et al., 2023). Our framework directly addresses this evaluation gap by explicitly structuring a discernible representation space, offering a highly efficient, discriminative alternative to resource-intensive LLM-as-a-judge paradigms.

3 Methodology

3.1 Problem Formulation

Let us define the pre-trained embedding space as a metric space $(\mathbb{R}^D, d_{sem})$, where a frozen general-purpose encoder $E(\cdot)$ maps textual inputs from a dataset $\mathcal{X} = \{x_i\}_{i=1}^N$ into high-dimensional vectors $\mathbf{z}_i \in \mathbb{R}^D$. Given a discrete label space $\mathcal{Y} = \{y_1, \ldots, y_C\}$ representing C specific evaluation principles, the fundamental bottleneck of post-hoc evaluation is the severe *feature entanglement*. Because the distance metric d_{sem} is optimized for universal semantic similarity, it remains agnostic to the task-specific logical boundaries of $\mathcal{Y}$. Consequently, for two text instances x_i and x_j sharing identical lexical structures but representing diametrically opposed principles ($y_i \neq y_j$), their embeddings collapse within the general manifold, yielding a vanishing distance $d_{sem}(\mathbf{z}_i, \mathbf{z}_j) \to 0$. This topological entanglement obliterates the linear separability required by downstream evaluators.

To resolve this entanglement without the prohibitive costs of full-parameter fine-tuning, we formulate the task as a constrained geometric optimization problem. Our objective is to learn a nonlinear mapping function $f_\theta : \mathbb{R}^D \to \mathcal{S}^{d-1}$ (where $d \ll D$ and $\mathcal{S}^{d-1}$ denotes the hypersphere manifold) alongside a set of learnable prototype kernels $\mathcal{K} = \{\mathbf{c}_1, \ldots, \mathbf{c}_C\} \in \mathcal{S}^{d-1}$. Rather than relying on implicit parameter updates, we explicitly enforce a margin-based topological constraint in the projected subspace:

$$\forall i, j: \quad ||f_\theta(\mathbf{z}_i) - \mathbf{c}_{y_i}||_2 \leq \delta_{intra}, \quad ||\mathbf{c}_{y_i} - \mathbf{c}_{y_j}||_2 \geq \delta_{inter} \tag{1}$$

where δ_{intra} bounds the intra-class compactness and δ_{inter} dictates a strict inter-class separation margin. For ordinal regression tasks, this categorical separation is seamlessly extended by enforcing a monotonic progression constraint along the kernel magnitudes. By optimizing f_θ to satisfy these constraints, we actively sculpt a discriminative metric space $(\mathbb{R}^d, d_{task})$ where principle-specific features are rigorously disentangled from general semantic noise.

3.2 Kernel-Guided Principle Extractor

The primary objective of the neural principle extractor, f_θ, is to project the highly entangled semantic vectors $\mathbf{z}_i \in \mathbb{R}^D$ onto a discriminative low-dimensional manifold $\mathcal{S}^{d-1}$. However, directly mapping inputs to principle-specific coordinates risks *representation collapse*, where text instances lose their inherent lexical diversity and merge into indistinguishable points.

To circumvent this, we design a *Dual-Stream Architecture* governed by specific inductive biases. We hypothesize that a robust task-specific representation must balance two orthogonal information flows: (1) a structural context that preserves the necessary semantic background, and (2) a dynamic projection that isolates the specific evaluation principle.

Semantic Basis Stream (Contextual Regularization). To prevent representation collapse, the first stream establishes a semantic baseline. We project the original embedding $\mathbf{z}_i$ through a shared Multi-Layer Perceptron (MLP):

$$\mathbf{s}_i = \text{MLP}_{\text{sem}}(\mathbf{z}_i) \in \mathbb{R}^d$$

where $\mathbf{s}_i$ denotes the semantic basis. This pathway acts as an information bottleneck that preserves lexical nuances (e.g., the specific object being reviewed) independent of the targeted principle, ensuring the resulting metric space remains continuous. Detailed specifications for this MLP are provided in Appendix A.1.

Prototype Mapping Stream (Dynamic Manifold Projection). The second stream is the core mechanism for resolving feature entanglement. Instead of relying on hyperplanes for linear classification, we introduce a matrix of learnable prototype kernels, $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_C]^\top \in \mathbb{R}^{C \times d}$. These kernels serve as explicit, optimizable geometric anchors for the C evaluation principles, providing the necessary coordinate centers for margin-based distance calculations.

To determine the input's alignment with each principle, we employ an attention mechanism as a soft projection operator. We project both the input $\mathbf{z}_i$ and the prototype matrix $\mathbf{C}$ into a joint space using linear weights $\mathbf{W}_q \in \mathbb{R}^{D \times d}$ and $\mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$:

$$\mathbf{q}_i = \mathbf{z}_i \mathbf{W}_q, \quad \mathbf{K} = \mathbf{C}\mathbf{W}_k, \quad \mathbf{V} = \mathbf{C}\mathbf{W}_v$$

The attention distribution $\mathbf{a}_i \in \mathbb{R}^C$ computes the coordinate coefficients of the input query against each principle's basis:

$$\mathbf{a}_i = \text{Softmax}\left(\frac{\mathbf{q}_i \mathbf{K}^\top}{\sqrt{d}}\right)$$

The principle-specific mapping $\mathbf{m}_i$ is then dynamically constructed as a convex combination of the projected prototype values:

$$\mathbf{m}_i = \mathbf{a}_i \mathbf{V} \in \mathbb{R}^d$$

Task-specific initialization strategies for these prototype kernels are detailed in Appendix A.2.

Feature Fusion and Manifold Projection. Finally, the extractor fuses the contextual semantics ($\mathbf{s}_i$) with the principle-specific alignment ($\mathbf{m}_i$). To satisfy the geometric constraints formulated in Eq. 1, the fused vector must be projected onto the hypersphere manifold

$$\tilde{\mathbf{e}}_i = \alpha \mathbf{s}_i + (1 - \alpha)\mathbf{m}_i, \qquad \mathbf{e}_i = \frac{\tilde{\mathbf{e}}_i}{||\tilde{\mathbf{e}}_i||_2} \in \mathcal{S}^{d-1}$$

where $\alpha \in [0, 1]$ is a learnable gating parameter initialized to a small scalar. This ensures that the final representation $\mathbf{e}_i$ is rigorously bounded within the target metric space $(\mathbb{R}^d, d_{task})$, fully prepared for the subsequent geometry-aware contrastive optimization.

3.3 Geometry-Aware Contrastive Objective

While the dual-stream extractor f_θ (Section 3.2) provides the structural capacity to project representations onto the hypersphere $\mathcal{S}^{d-1}$, this architecture alone does not guarantee feature disentanglement. Without explicit topological constraints, the prototype kernels $\mathbf{C}$ remain arbitrary vectors, the attention mechanism risks uniform collapse, and the semantic stream $\mathbf{s}_i$ may redundantly encode task-specific signals. To activate the inductive biases designed in f_θ and physically sculpt the principle-aligned subspace described in Section 3.1, we must subject the network to a set of strict geometric optimization forces.

Relying solely on standard contrastive learning is insufficient for this purpose. Traditional InfoNCE objectives optimize relative probabilities via softmax, which encourages general clustering but fails to guarantee strict geometric margins. Consequently, semantically overlapping classes can still suffer from boundary entanglement. To enforce rigorous mathematical boundaries and drive the parameters of the extractor, we construct a composite objective $\mathcal{L}_{\text{total}}$ where each component serves a distinct topological purpose tailored to the dual-stream architecture. Specifically, this composite objective integrates four complementary mechanisms: a *Supervised Contrastive Loss* to establish macroscopic class clustering, an *Offset Loss* to enforce geometric separation margins, an *Orthogonality Loss* to decouple semantic context from principle features, and an optional *Magnitude Loss* to preserve ordinal intensity progressions.

Global Clustering via Supervised Contrastive Loss ($\mathcal{L}_{\text{contrastive}}$). Before carving fine-grained boundaries, the model must establish a macro-level topology. We utilize a Supervised InfoNCE loss to form the initial task-specific clusters. By utilizing target labels, it pulls the fused representation $\mathbf{e}_i$ towards positive samples $\mathbf{e}_{p_i}$ sharing the same principle y_i, while broadly repelling samples from differing principles. This establishes the foundational macroscopic structure of the metric space.

Strict Margin Enforcement via Offset Loss ($\mathcal{L}_{\text{offset}}$). To upgrade the relative separation provided by contrastive learning into a physical margin, we introduce the Offset Penalty. This acts as our primary geometric regularizer. It directly controls the coordinates of the kernel mapping $\mathbf{m}_i$ relative to the prototype anchors, ensuring that the theoretical margins formulated in Eq. 1 are satisfied. It operates through two tandem constraints:

- **Intra-Class Penalty (Bounding Variance):** To prevent clusters from expanding indefinitely and to ensure semantic diversity remains controlled within a local neighborhood, we introduce a "safe radius" δ_{intra}. It penalizes samples only if they drift beyond this distance from their target prototype $\mathbf{c}_{y_i}$:
$$P_{\text{intra},i} = \max(0, ||\mathbf{m}_i - \mathbf{c}_{y_i}||_2 - \delta_{\text{intra}})^2$$
- **Inter-Class Penalty (Forcing Isolation):** To physically break feature entanglement, we must guarantee a clear geometric vacuum between confusable classes. This penalty ensures that a sample is closer to its true prototype than to any incorrect prototype $\mathbf{c}_k$ by a minimum distance δ_{inter}:
$$P_{\text{inter},i} = \max(0, ||\mathbf{m}_i - \mathbf{c}_{y_i}||_2 - \min_{k \neq y_i} ||\mathbf{m}_i - \mathbf{c}_k||_2 + \delta_{\text{inter}})^2$$

The batch-level offset loss is the dynamically weighted expectation of these penalties, where w_{y_i} serves to counteract class imbalance:

$$\mathcal{L}_{\text{offset}} = \frac{1}{B} \sum_{i=1}^{B} w_{y_i} (\lambda_{\text{inclass}} P_{\text{intra},i} + \lambda_{\text{crossclass}} P_{\text{inter},i})$$

Information Decoupling via Orthogonality Loss ($\mathcal{L}_{\text{orthogonality}}$). For the dual-stream architecture (Section 3.2) to function properly, the Semantic Basis ($\mathbf{s}_i$) and the Prototype Mapping ($\mathbf{m}_i$) must capture mutually exclusive information. If $\mathbf{s}_i$ leaks principle-specific signals, it bypasses the geometric constraints of the offset penalty. To mathematically enforce this information bottleneck, we apply a soft Orthogonality Loss that penalizes high cosine similarity between the two streams:

$$\mathcal{L}_{\text{orthogonality}} = \frac{1}{B} \sum_{i=1}^{B} w_{y_i} \max(0, |\cos(\mathbf{s}_i, \mathbf{m}_i)| - \delta_{\text{orthogonal}})$$

Structural Progression via Magnitude Loss ($\mathcal{L}_{\text{magnitude}}$). When the evaluation principles exhibit an inherent ordinal relationship (e.g., 1 to 5 star ratings), treating them as independent categorical clusters ignores the severity of misclassification (e.g., confusing 1-star with 5-star is worse than with 2-star). To capture this ordered progression, we introduce a Magnitude Loss that anchors the geometric norm of the representations to their numerical intensity $I(y_i)$:

$$\mathcal{L}_{\text{magnitude}} = \frac{1}{B} \sum_{i=1}^{B} w_{y_i} (||\mathbf{m}_i||_2 - \lambda_{\text{scale}} I(y_i) \cdot ||\mathbf{c}_{y_i}||_2)^2$$

Overall Objective. The final network is trained end-to-end by minimizing the weighted summation:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{offset}} \mathcal{L}_{\text{offset}} + \lambda_{\text{contrastive}} \mathcal{L}_{\text{contrastive}} + \lambda_{\text{orth}} \mathcal{L}_{\text{orthogonality}} + \lambda_{\text{mag}} \mathcal{L}_{\text{magnitude}}$$

The magnitude term, $\lambda_{\text{mag}} \mathcal{L}_{\text{magnitude}}$, is activated exclusively for ordinal tasks. The exhaustive training configurations, including optimizer setups and hyperparameter selection, are provided in Appendix A.3. A detailed computational complexity analysis demonstrating the extreme efficiency of this objective is presented in Appendix A.4.

3.4 Theoretical Bounds on Geometric Quality

A core motivation for explicitly designing the Offset Loss ($\mathcal{L}_{\text{offset}}$) is to mathematically guarantee the geometric properties of the projected subspace $(\mathbb{R}^d, d_{\text{task}})$. In this section, we formally state the topological margins established by our formulation. Detailed mathematical proofs and extended discussions on empirical error bounds and architecture justifications are provided in Appendix B.

Theorem 1 (Margin and Guaranteed Separation). Let $\mathcal{M}_A$ and $\mathcal{M}_B$ denote the sets of extracted representations belonging to two distinct evaluation principles A and B, with prototype kernels $\mathbf{c}_A$ and $\mathbf{c}_B$. Under the assumption of sufficient parameterization and given that the hyperparameter margin capacity is designed such that $\delta_{\text{inter}} > \delta_{\text{intra}}$, the Euclidean distance between any $\mathbf{m}_a \in \mathcal{M}_A$ and $\mathbf{m}_b \in \mathcal{M}_B$ is lower-bounded by a positive constant:

$$||\mathbf{m}_a - \mathbf{m}_b||_2 > 0 \tag{2}$$

Theorem 2 (Bounds on Geometric Clustering Metrics). Building upon the assumption of sufficient parameterization and given that the hyperparameters satisfy the margin condition $\delta_{\text{inter}} > 3\delta_{\text{intra}}$, the representation space guarantees a theoretical upper bound for the Within/Between distance ratio, and a positive theoretical lower bound for the Silhouette Score:

$$\text{Ratio}_{W/B} \leq \frac{2\delta_{\text{intra}}}{\delta_{\text{inter}} - \delta_{\text{intra}}} \qquad \text{and} \qquad S(\mathbf{m}_i) \geq 1 - \frac{2\delta_{\text{intra}}}{\delta_{\text{inter}} - \delta_{\text{intra}}} > 0 \tag{3}$$

Discussion. These theorems establish the mathematical requirements for structuring the representation space. *Theorem 1* shows that enforcing a separation between prototype kernels, alongside bounded intra-class deviation, induces a non-trivial lower bound on inter-class distances, providing a sufficient condition for separability. *Theorem 2* connects these constraints to standard clustering metrics, demonstrating that $\delta_{\text{inter}} > 3\delta_{\text{intra}}$ yields a positive Silhouette Score ($S > 0$). This condition arises naturally from comparing the maximum intra-class diameter ($2\delta_{\text{intra}}$) with the required inter-class separation. To prevent boundary overlap, the center-to-center margin (δ_{inter}) must accommodate its own radius (δ_{intra}) plus the adjacent cluster's diameter ($2\delta_{\text{intra}}$). While empirical optimization errors may relax these guarantees in practice, the derived bounds provide a useful characterization of the geometric bias induced by the objective.

Implications. The practical value of these bounds lies in reducing the burden on downstream models by promoting structured representations with bounded intra-class variation and controlled inter-class separation. The separability established in *Theorem 1* bypasses the need for non-linear architectures to resolve semantic ambiguities. Furthermore, the $S > 0$ guarantee from *Theorem 2* restricts the hypothesis space; because semantic noise is isolated from the inter-class gap, it does not erode the decision boundary. This structured topology reduces sample complexity, allowing lightweight linear probes to operate effectively. For ordinal settings, the framework accommodates monotonic intensity progressions, capturing graded relationships while preserving class separation.

4 Experiment

In this section, we empirically validate the efficacy, efficiency, and geometric properties of the Kernel-Guided Contrastive Learning (KGCL) framework. Our evaluation is designed to answer three core questions directly tied to our theoretical formulations: (1) Does the theoretical disentanglement guaranteed by *Theorem 1* empirically translate into enhanced linear separability? (2) Do the optimized representations satisfy the theoretical bounds on clustering metrics established in *Theorem 2*? (3) How does this explicitly sculpted geometry compare against massive generalist representation paradigms?

4.1 Experimental Setup

Datasets. To comprehensively test our framework's ability to isolate subtle principles within semantically overlapping texts, we select three challenging datasets representing distinct evaluation tasks: **(1) GoEmotions (Demszky et al., 2020):** A large-scale corpus of Reddit comments. Serving as a highly controlled

proxy for complex subjective alignment, we focus on a highly confusable subset of five emotion principles (Disappointment, Sadness, Disapproval, Gratitude, Approval). Because these emotions frequently share similar vocabulary (e.g., dense negative sentiment), this task rigorously tests the framework's ability to enforce *Theorem 1* in densely entangled spaces, a prerequisite for resolving nuanced human-centric evaluation criteria. **(2) Amazon Reviews (Ni et al., 2019):** Comprising user reviews and 1-5 star ratings, this dataset acts as a structural proxy for evaluating the severity or degree of principle violations. We utilize this to validate that our geometric separation seamlessly extends to ordinal intensities, modeling scenarios where alignment requires measuring monotonic progression rather than just binary classification. **(3) Toxic Comment Classification Challenge (cjadams et al., 2017):** A critical dataset for evaluating principle alignment in content safety. This presents a severe lexical conflation challenge, where non-toxic venting can closely resemble toxic attacks. We utilize an extremely unbalanced test set (approximately 1:25 toxic vs. non-toxic), mirroring real-world moderation scenarios. The training set is resampled to a 1:3 ratio to facilitate stable learning.

Base Encoder and Baselines. Across all primary experiments, we utilize `jina-embeddings-v3` (Sturua et al., 2024) ($D = 1024$) as our frozen base encoder, owing to its state-of-the-art performance in capturing broad semantic similarity. Our neural principle extractor projects these into a $d = 64$ dimensional subspace (dimension justification in Appendix A.5). To demonstrate the superiority of KGCL, we establish comparisons against three tiers of baselines: (1) **Raw Embeddings:** Direct evaluation on the frozen `jina-embeddings-v3` features using linear probes; (2) **Unsupervised Contrastive Learning:** General structure-enhancing methods like SimCSE (Gao et al., 2021); (3) **Generative LLMs:** Massive language models evaluated via few-shot prompting to benchmark explicitly sculpted geometry against implicit in-context reasoning.

Evaluation Metrics. Due to the severe class imbalance in datasets like Toxic Comments and Amazon Reviews, we report *Macro-F1* alongside standard *Accuracy*. All performance metrics are reported as Mean $\pm$ Standard Deviation over 10-fold cross-validation. Crucially, to directly validate the bounds established in *Theorem 2*, we compute standard geometric indices alongside our composite *Geometric Quality Index (GQI)*. The GQI explicitly quantifies the topological quality of the representation space by measuring the ratio of inter-class separation to intra-class compactness. A higher GQI indicates that the representations are successfully disentangled and rigorously bounded around their respective principle kernels.

4.2 Core Task Validation: Downstream Efficacy of Geometric Disentanglement

To rigorously evaluate our framework, we first isolate the pure representational gain achieved by the KGCL module. We compare the optimized embeddings against raw embeddings to evaluate whether the theoretical disentanglement guaranteed by *Theorem 1* effectively translates into enhanced linear separability for downstream classification.

Enhancing Linear Separability (GoEmotions). A critical question is whether the performance improvements stem from a superior embedding geometry or merely from adding a downstream classification head. To address this, we establish a strict "Simple Fine-Tuning Baseline" on the GoEmotions five-principle subset. We freeze the base embeddings and train an identical suite of classifiers (SVM, Random Forest, Logistic Regression, XGBoost, Transformer) on both the high-dimensional raw embeddings (1024-d) and our optimized representations (64-d).

As summarized in Table 1, training standard classifiers directly on raw embeddings plateaus at an Overall F1-score around 0.72-0.73. However, simply replacing the input with our optimized embeddings immediately yields consistent and statistically significant improvements across all metrics. For instance, Logistic Regression (LR) F1 jumps from 0.726 ± 0.031 to 0.776 ± 0.032. This improvement provides strong empirical support for the practical utility of *Theorem 1*: by mathematically enforcing a separation margin, KGCL transforms a highly entangled space into a discriminative topology, allowing even simple linear probes to find clear decision boundaries.

Table 1: Overall (Avg. Principle) Performance on GoEmotions Five-Principle Set (Mean ± Std. Dev.)

Metric	Emb. Type	SVM	RF	LR	XGBoost	Transformer
Precision	Raw Emb.	0.748 ± 0.049	0.733 ± 0.059	0.737 ± 0.031	0.747 ± 0.020	0.785 ± 0.031
	Opt. Emb.	**0.787** ± 0.035	**0.789** ± 0.036	**0.791** ± 0.033	**0.773** ± 0.028	**0.787** ± 0.029
Recall	Raw Emb.	0.721 ± 0.045	0.737 ± 0.031	0.722 ± 0.031	0.741 ± 0.014	0.763 ± 0.024
	Opt. Emb.	**0.764** ± 0.034	**0.769** ± 0.039	**0.772** ± 0.033	**0.765** ± 0.039	**0.769** ± 0.031
F1	Raw Emb.	0.729 ± 0.046	0.722 ± 0.035	0.726 ± 0.031	0.737 ± 0.018	0.764 ± 0.036
	Opt. Emb.	**0.770** ± 0.033	**0.767** ± 0.036	**0.776** ± 0.032	**0.764** ± 0.026	**0.770** ± 0.030

Extension to Ordinal Intensities (Amazon Reviews). As a natural extension of our geometric separation, we evaluate the framework on ordinal intensities (1-5 star ratings). As shown in Table 2, optimized embeddings consistently improve overall regression metrics (e.g., MSE, RMSE) compared to raw embeddings, confirming that the structured subspace seamlessly accommodates continuous ordinal constraints without disrupting the primary categorical disentanglement.

Table 2: Overall Ordinal Regression Performance on Amazon Reviews (Mean ± Std. Dev.)

Metric	Emb. Type	SVM	RF	LR	XGBoost	Transformer
MSE	Raw Emb.	0.668 ± 0.135	0.506 ± 0.120	0.635 ± 0.097	0.546 ± 0.163	0.602 ± 0.173
	Opt. Emb.	**0.365** ± 0.158	**0.392** ± 0.143	**0.394** ± 0.149	**0.377** ± 0.097	**0.359** ± 0.086
RMSE	Raw Emb.	0.813 ± 0.083	0.706 ± 0.083	0.795 ± 0.059	0.731 ± 0.111	0.768 ± 0.110
	Opt. Emb.	**0.593** ± 0.119	**0.617** ± 0.107	**0.618** ± 0.112	**0.609** ± 0.080	**0.595** ± 0.071
R^2	Raw Emb.	0.604 ± 0.086	0.700 ± 0.074	0.624 ± 0.060	0.677 ± 0.095	0.643 ± 0.103
	Opt. Emb.	**0.785** ± 0.089	**0.770** ± 0.080	**0.768** ± 0.083	**0.777** ± 0.055	**0.788** ± 0.047

Robustness Against Majority Collapse (Toxic Comments). For principle alignment evaluation in a sensitive domain, we assess our framework on the highly unbalanced Toxic Comment dataset. Table 3 shows that optimized embeddings yield statistically significant improvements in both Average F1 and Minority F1 across all classifiers. This demonstrates the downstream value of the inter-class margin (δ_{inter}) established in *Theorem 1*. By geometrically shielding the sparse toxic class, KGCL prevents minority features from being overwhelmed by the dominant non-toxic majority during semantic projection.

Table 3: Performance on Toxic Comment Classification Challenge (Mean ± Std. Dev.)

Metric	Emb. Type	SVM	RF	LR	XGBoost	Transformer
Avg. F1	Raw Emb.	0.932 ± 0.004	0.949 ± 0.003	0.897 ± 0.004	0.918 ± 0.004	0.956 ± 0.004
	Opt. Emb.	**0.938** ± 0.004	**0.949** ± 0.004	**0.936** ± 0.003	**0.943** ± 0.004	**0.959** ± 0.004
Minority F1	Raw Emb.	0.497 ± 0.025	0.405 ± 0.044	0.396 ± 0.018	0.433 ± 0.024	0.574 ± 0.027
	Opt. Emb.	**0.507** ± 0.024	**0.537** ± 0.035	**0.493** ± 0.023	**0.518** ± 0.028	**0.589** ± 0.023

4.3 Paradigm Comparisons: Contrastive Baselines and LLMs

Having established the superiority of KGCL over raw embeddings, we now benchmark our framework against alternative learning paradigms, specifically task-agnostic contrastive learning and massive generative language models.

Overcoming the Limitations of Task-Agnostic Contrastive Learning. We compare KGCL against structure-enhancing contrastive baselines to validate our theoretical motivation. As detailed in Table 4, maximizing general semantic similarity via Unsupervised SimCSE actually degrades downstream performance

compared to raw embeddings (Average F1 dropping from 0.620 to 0.531). This confirms that without the explicit inter-class margin (δ_{inter}) enforced by our objective, task-agnostic optimization conflates crucial task boundaries (e.g., confusing "logistics failure" with "product defect" due to shared negative sentiment). Conversely, our custom supervised KGCL achieves a substantial +19.4% absolute F1 improvement over SimCSE, significantly outperforming even standard supervised baselines.

Table 4: Downstream Classification Performance on Amazon Reviews: Contrastive Baselines Comparison (AUC / Overall F1-Score)

Embedding Source	SVM	Random Forest	Logistic Reg.	XGBoost	Transformer	**Avg. F1**
Unsupervised SimCSE	0.851 / 0.568	0.823 / 0.539	0.843 / 0.572	0.840 / 0.567	0.730 / 0.407	0.531
Standard Supervised	0.921 / 0.697	0.916 / 0.718	0.918 / 0.687	0.922 / 0.687	0.921 / 0.687	0.695
KGCL (Ours)	**0.934** / **0.709**	**0.933** / **0.749**	**0.928** / **0.719**	**0.930** / **0.724**	**0.927** / **0.726**	**0.725**

Comparison with Few-shot Large Language Models. To establish the performance ceiling and address the prevailing trend of utilizing LLM-as-a-judge, we compare our method's performance with that of few-shot prompted Large Language Models (LLMs) serving as generalist evaluators. We evaluated `grattafiori2024llama_70b_Q4_K` (Grattafiori et al., 2024), `deepseek-chat-v3-0324` (Liu et al., 2024), and `gemini-2.5-pro-exp-03-25` (Comanici et al., 2025) via direct API prompting.

Table 5 summarizes this comparison. As shown in Table 5, our explicitly optimized embeddings paired with a simple Transformer head consistently outperform the few-shot LLMs across all three principle alignment tasks. This underscores a critical paradigm advantage: explicitly sculpting a discriminative geometric topology for specific principles yields more robust decision boundaries than relying on the implicit in-context reasoning of general-purpose LLMs.

Table 5: Performance Comparison with Few-shot Large Language Models

Dataset	Metric	grattafiori2024llama	DeepSeek-chat-v3	Gemini-2.5-pro	Opt. Emb. + Transformer
GoEmotions	F1-score	0.67	0.70	0.70	**0.77**
Amazon Reviews	MSE	0.60	0.45	0.56	**0.36**
Toxic Comment	Avg. F1	0.91	0.89	0.91	**0.96**

4.4 Geometric Analysis, Ablation, and Efficiency

To understand the mechanics behind KGCL's empirical superiority, we analyze the geometric properties of the learned subspace, ablate its core loss components, and evaluate its practical deployment advantages.

Quantitative Geometric Analysis and the GQI Metric. To explicitly validate the theoretical bounds established in *Theorem 2*, we evaluate the topological quality of the embeddings using standard geometric metrics: the ratio of Within-class to Between-class Variance (Fisher, 1936), Silhouette Score (Rousseeuw, 1987), and Class Overlap (Dom, 2012). Furthermore, to provide a holistic measure of the subspace's suitability for linear classification, we formulate the *Geometric Quality Index (GQI)*:

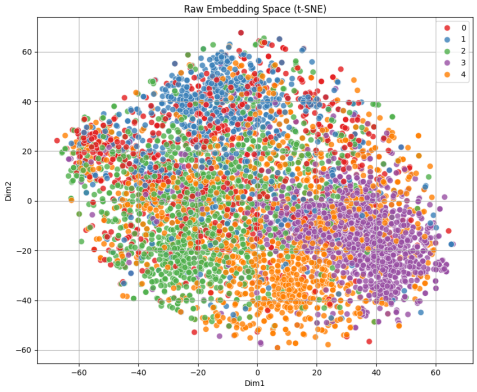
$$GQI = \left(1 - \frac{\text{Within Variance}}{\text{Between Variance}}\right) \times \text{Silhouette Score} \times (1 - \text{Class Overlap}) \tag{4}$$

A higher GQI indicates a space where clusters are internally compact and externally well-separated.

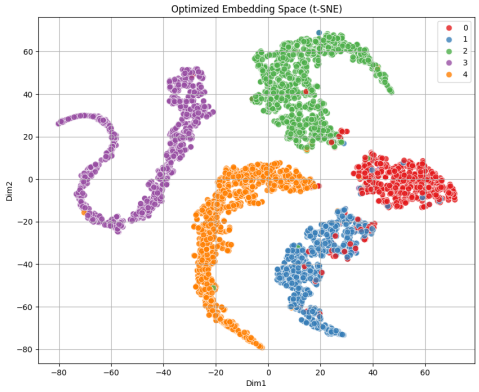
As detailed in Table 6 for the Amazon Reviews dataset, raw embeddings and task-agnostic methods like SimCSE fail to create a discriminative geometry. Notably, SimCSE exhibits a Within/Between Ratio greater than 1.0 (indicating intra-class variance exceeds inter-class distance), resulting in a negative GQI (-0.0005). In contrast, KGCL drastically compresses the clusters and pushes them apart, achieving a Within/Between Ratio of 0.358 and a GQI of 0.0975. This structural breakthrough empirically validates the bounds derived in *Theorem 2* and perfectly aligns with the downstream performance leaps observed in Section 4.3.

Table 6: Quantitative Geometric Quality and GQI Comparison on Amazon Reviews

Embedding Paradigm	Within/Between Ratio (↓)	Silhouette Score (↑)	Class Overlap (↓)	**GQI (↑)**
Raw Embeddings	8.760	0.018	0.563	< 0.000
Unsupervised SimCSE	1.010	0.010	0.491	-0.0005
Standard Supervised	0.458	0.158	0.286	0.0614
KGCL (Ours)	**0.358**	**0.203**	**0.253**	**0.0975**

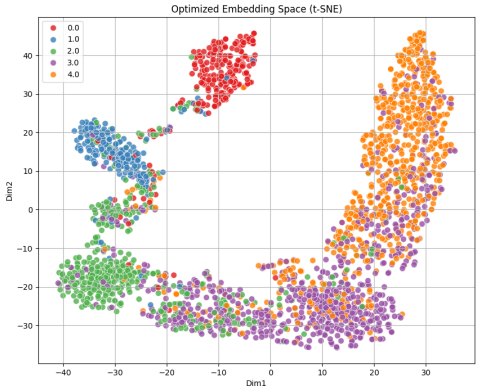


(a) GoEmotions - Raw



(b) GoEmotions - Opt.

(c) Amazon Reviews - Raw



(d) Amazon Reviews - Opt.

Figure 2: t-SNE comparison of embedding spaces. Raw embeddings (a, c) suffer from severe semantic conflation. KGCL optimized embeddings (b, d) demonstrate physical boundaries and ordinal structure.

Qualitative Visualization. This quantitative structural improvement is vividly reflected in the t-SNE (Van der Maaten & Hinton, 2008) visualizations (Figure 2). The raw embeddings (Figures 2a and 2c) show severe representational entanglement, with semantically overlapping principles completely conflated. After applying KGCL (Figures 2b and 2d), the embeddings are distinctly pulled towards their respective prototype kernels, forming highly separable clusters. For ordinal tasks (Amazon), the clusters seamlessly exhibit an ordered geometric progression, visually confirming the structural flexibility of the subspace.

Ablation Study: The Role of Offset Penalty. To isolate the contribution of each training objective, we conducted an ablation study on the GoEmotions dataset (Table 7). The integration of the *Offset Loss* is particularly crucial; removing it drops the performance on highly confusable principles like 'Disappointment'

(from 0.49 to 0.44). This confirms that the inter-class margin (δ_{inter}) mathematically guaranteed by the offset penalty (*Theorem 1*) is mandatory for disentangling nuanced semantic overlap, proving that the losses act synergistically.

Table 7: Ablation study on GoEmotions (F1 score Mean). Disappt.-Disappointment, Sad.-Sadness, Disapprv.-Disapproval, Grat.-Gratitude, Apprv.-Approval.

Configuration	Disappt.	Sad.	Disapprv.	Grat.	Apprv.	**Average**
Only Contrastive Loss	0.37	0.75	0.72	0.92	0.74	0.75
Only Offset Loss	0.40	0.75	0.72	0.94	0.77	0.77
Without Contrastive Loss	0.42	0.77	0.72	0.94	0.77	0.77
Without Offset Loss	0.44	0.71	0.73	0.93	0.76	0.77
Raw Embeddings	0.36	0.64	0.66	0.93	0.72	0.72
KGCL (Full Model)	**0.49**	**0.77**	**0.72**	**0.94**	**0.76**	**0.78**

Beyond Accuracy: Deployment Efficiency. Ultimately, the value of transforming the embedding geometry extends beyond metric gains. By restructuring the massive 1024-d raw vectors into a 64-d principle-aligned subspace, KGCL provides a highly reusable intermediate representation. This extreme dimensionality reduction directly translates to enhanced computational efficiency during inference. For instance, the training and inference times for downstream models like XGBoost were reduced by up to 96.5% compared to using raw embeddings. Consequently, our framework empowers simple, low-latency linear classifiers to match or exceed the performance of generative LLMs, rendering high-precision, high-throughput principle evaluation practically viable for edge deployments.

5 Conclusion and Future Work

In this paper, we address the critical bottleneck of feature entanglement in standard text embeddings, which severely hinders precise principle evaluation. Rather than relying on the computationally prohibitive fine-tuning of massive models, we introduce Kernel-Guided Contrastive Learning (KGCL), a highly efficient framework designed to explicitly sculpt the geometric topology of the representation space. By employing learnable prototype kernels as structural anchors and introducing a novel offset penalty, our method enforces strict, physically verifiable margins between semantically confusable principles. We formally prove that this formulation guarantees topological separation and dictates the theoretical limits of clustering quality. Extensive evaluations confirm that this explicit geometric remodeling provides a pure representational gain, successfully translating mathematical bounds into empirically robust linear separability. Consequently, our framework empowers simple, low-latency linear classifiers to consistently match or exceed the performance of massive generative LLMs, indicating that explicitly sculpted geometry offers a more robust alternative to implicit semantic matching for targeted tasks. Ultimately, KGCL validates our central thesis: whether identifying toxic attacks or parsing highly entangled subjective criteria, transitioning from generic semantic approximation to targeted decision boundary sculpting is a critical step toward building robust, high-resolution principle evaluation systems.

While KGCL establishes a robust paradigm for task-oriented embedding alignment, current limitations include its reliance on supervised data for predefined principles and the need for further verification across highly diverse linguistic domains. Furthermore, as with any alignment technology, risks concerning potential misuse or the amplification of inherent biases warrant careful consideration. Future work will explore extending this geometric framework to semi-supervised or few-shot scenarios to dynamically adapt to unseen principles. Additionally, integrating our highly discriminative, low-latency principle extractors as automated evaluators within Reinforcement Learning from AI Feedback (RLAIF) pipelines (Bai et al., 2022; Lee et al., 2023) presents a highly promising direction for generating robust, high-throughput feedback signals.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. Toxic comment classification challenge. `https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge`, 2017. Kaggle.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4040–4054, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, pp. 4171–4186, 2019.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*, 2022.

Byron E Dom. An information-theoretic external cluster-validity measure. *arXiv preprint arXiv:1301.0565*, 2012.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Shimon Edelman and Nathan Intrator. Learning as extraction of low-dimensional representations. In *Psychology of learning and motivation*, volume 36, pp. 353–380. Elsevier, 1997.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2): 179–188, 1936.

Kenji Fukumizu, Francis Bach, and Michael Jordan. Kernel dimensionality reduction for supervised learning. *Advances in neural information processing systems*, 16, 2003.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6894–6910, 2021.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 3356–3369, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Minlong Peng, Qi Zhang, Xiaoyu Xing, Tao Gui, Jinlan Fu, and Xuanjing Huang. Learning task-specific representation for novel words in sequence labeling. *arXiv preprint arXiv:1905.12277*, 2019.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China technological sciences*, 63(10):1872–1897, 2020.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. *arXiv preprint arXiv:2002.05969*, 2020.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. Flexgen: High-throughput generative inference of large language models with a single gpu. In *International Conference on Machine Learning*, pp. 31094–31116. PMLR, 2023.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*, 2024.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pp. 1–11, 2023.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. In *Findings of the association for computational linguistics: EMNLP 2021*, pp. 2447–2469, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yichu Zhou and Vivek Srikumar. A closer look at how fine-tuning changes bert. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1046–1061, 2022.

A Implementation Details

This appendix provides further details regarding the neural principle extractor's architecture, prototype kernel initialization strategies, specific training hyperparameters and loss function configurations, computational complexity, and justification for the principle subspace dimension, as referenced in the main paper.

A.1 Neural Network Architecture Details

The neural principle extractor f_θ is implemented as a neural network that maps the input text embedding $\mathbf{X}_i \in \mathbb{R}^{1024}$ to a $d = 64$ dimensional principle-aware representation $\mathbf{e}_i$. The architecture is composed of a shared Multi-Layer Perceptron (MLP) and an attention mechanism.

The shared MLP used to compute the semantic basis $\mathbf{s}_i$ consists of two fully connected layers with LeakyReLU activation functions and Batch Normalization. Dropout is applied after each hidden layer for regularization. The layer dimensions are as follows:

- Input layer: $\mathbb{R}^{1024} \to \mathbb{R}^{512}$
- Hidden layer 1: $\mathbb{R}^{512} \to \mathbb{R}^{256}$ (followed by LeakyReLU, Batch Norm, Dropout)
- Hidden layer 2: $\mathbb{R}^{256} \to \mathbb{R}^{d}$ (followed by LeakyReLU, Batch Norm, Dropout), where $d = 64$. The output of this layer is the semantic basis $\mathbf{s}_i$.

The Dropout rate used throughout the MLP is 0.2.

The attention mechanism involves linear transformations of the input embedding and the prototype kernels to compute queries, keys, and values:

- Query projection: query_fc : $\mathbb{R}^{1024} \to \mathbb{R}^{d}$
- Key projection: key_fc : $\mathbb{R}^{d} \to \mathbb{R}^{d}$
- Value projection: value_fc : $\mathbb{R}^{d} \to \mathbb{R}^{d}$

These projected vectors are used in the scaled dot-product attention calculation.

The learnable parameter α that weights the semantic basis and kernel mapping in the final fusion layer is a scalar variable initialized to 0.05.

A.2 Prototype Kernel Initialization Details

The K learnable prototype kernels $\mathbf{c}_k \in \mathbb{R}^d$ are initialized based on the task type to encourage structured learning.

For Classification Tasks: For classification tasks (GoEmotions 5-principle set, Amazon Reviews classification), the K prototype kernels are initialized randomly on the unit hypersphere in $\mathbb{R}^d$. To ensure distinct starting points and facilitate separation during training, we apply a procedure to guarantee a minimum pairwise Euclidean distance between any two initialized kernels. While not strictly enforcing orthogonality, this initial separation prevents kernels from collapsing onto the same point early in training. A target minimum distance of $\sqrt{2}$ (the Euclidean distance between orthogonal unit vectors) is aimed for during this initialization step.

For Ordinal Regression Tasks: To strictly adhere to the topological constraint defined in Section 3.1 ($\mathcal{K} \in \mathcal{S}^{d-1}$), the prototype kernels for ordinal ratings (e.g., 1 to 5 stars) are exclusively initialized on the unit hypersphere. Rather than scaling their magnitudes, we initialize them to form a sequential angular progression along a defined geodesic path. This ensures that the distance between c_1 and c_5 is geometrically maximized, while adjacent ratings (e.g., c_2 and c_3) remain topologically adjacent, thereby preserving ordinal intensity without violating the strict unit-norm constraint of the projected metric space.

A.3 Training and Loss Function Details

This appendix provides the full mathematical formulations for each component of the composite objective $\mathcal{L}_{\text{total}}$ referenced in Section 3.3, along with specific training configurations.

The neural principle extractor is trained end-to-end using the AdamW optimizer with an initial learning rate of 1e-4 and a weight decay of 1e-5. A learning rate scheduler (Cosine Annealing or ReduceLROnPlateau) dynamically adjusts the learning rate based on validation performance. Training is capped at 100 epochs with early stopping to prevent overfitting, using a consistent batch size of 128. To counteract class imbalance, dynamic class weights w_{y_i} are applied across all loss calculations, computed as the inverse frequency of each true class within the current training batch.

Contrastive Loss ($\mathcal{L}_{\text{contrastive}}$). To establish the macroscopic class clustering, we employ the standard Supervised Contrastive Loss (SupCon), which effectively leverages label information to pull together samples from the same principle while repelling negatives:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{B}\sum_{i=1}^{B}\frac{-w_{y_i}}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\text{sim}(e_i,e_p)/\tau)}{\sum_{a\in A(i)}\exp(\text{sim}(e_i,e_a)/\tau)} \tag{5}$$

where $A(i)$ is the set of all other samples in the batch excluding i, $P(i)$ is the set of positive samples sharing the same label y_i as anchor i, $|P(i)|$ is its cardinality, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and $\tau = 0.1$ is the temperature scaling factor.

Offset Loss ($\mathcal{L}_{\text{offset}}$). The total offset loss combines the intra-class and inter-class penalties:

$$\mathcal{L}_{\text{offset}} = \frac{1}{B}\sum_{i=1}^{B} w_{y_i}\left(\lambda_{\text{inclass}}P_{\text{intra},i} + \lambda_{\text{crossclass}}P_{\text{inter},i}\right)$$

where λ_{inclass} and $\lambda_{\text{crossclass}}$ dictate the relative strengths of the penalties, and the margins δ_{intra} and δ_{inter} typically range within $[0.1, 0.5]$.

Orthogonality Loss ($\mathcal{L}_{\text{orthogonality}}$). Promotes "soft" orthogonality between the semantic basis $\mathbf{s}_i$ and kernel mapping $\mathbf{m}_i$:

$$\mathcal{L}_{\text{orthogonality}} = \frac{1}{B}\sum_{i=1}^{B} w_{y_i}\cdot\max(0, |\cos(\mathbf{s}_i,\mathbf{m}_i)| - \delta_{\text{orthogonal}})$$

where the dynamic margin $\delta_{\text{orthogonal}}$ is linearly annealed from 0.5 down to 0.05 during training.

Magnitude Loss ($\mathcal{L}_{\text{magnitude}}$). Applied exclusively for ordinal regression tasks to enforce natural ordering:

$$\mathcal{L}_{\text{magnitude}} = \frac{1}{B}\sum_{i=1}^{B} w_{y_i}\left(||\mathbf{m}_i||_2 - \lambda_{\text{scale}}I(y_i)\cdot||\mathbf{c}_{y_i}||_2\right)^2$$

where λ_{scale} is a learnable scaling factor and $I(y_i)$ maps the label to a numerical intensity (e.g., $I(y_i) = y_i$).

The final total objective ($\mathcal{L}_{\text{total}}$) is optimized by balancing these components.

A.4 Computational Complexity Analysis

We analyze the computational complexity of our framework during training and inference.

Training Complexity: The primary computational cost during training arises from the forward and backward passes through the neural principle extractor and the calculation of the loss components over a batch of size B. The extractor involves:

- Shared MLP: A sequence of matrix multiplications. Given input dimension $D = 1024$, output dimension $d = 64$, and hidden dimensions $h_1 = 512, h_2 = 256$, the complexity is $O(D\cdot h_1 + h_1\cdot h_2 + h_2\cdot d)$ per sample.

- Attention Mechanism: Involves linear projections ($O(D \cdot d + d^2 \cdot K)$ for a batch of size B, where K is the number of principles), computing attention scores ($O(B \cdot K \cdot d)$), and weighted summation ($O(B \cdot K \cdot d)$).

The dominant part of the forward pass per batch is approximately $O(B \cdot (D \cdot h_1 + h_1 \cdot h_2 + h_2 \cdot d + K \cdot d))$. Loss calculations involve vector operations and distance calculations on the d-dimensional embeddings and K kernels:

- Contrastive Loss: $O(B^2 \cdot d)$ in the standard form, often optimized to $O(B^2)$ or $O(B \cdot P \cdot d)$ with P positives per sample.
- Offset Loss: Involves distances to K kernels, $O(B \cdot K \cdot d)$.
- Orthogonality, Classification, Magnitude Losses: $O(B \cdot d)$ or $O(B)$.

The overall training complexity per batch is dominated by the forward/backward passes and loss calculations, roughly $O(B \cdot (D \cdot h_{max} + K \cdot d) + B^2 \cdot d)$ in the worst case (for contrastive) or $O(B \cdot (D \cdot h_{max} + K \cdot d))$ with typical batch sizes and optimizations. This is comparable to other deep metric learning or contrastive learning frameworks.

Inference Complexity: Inference requires a single forward pass through the extractor. The complexity per sample is $O(D \cdot h_1 + h_1 \cdot h_2 + h_2 \cdot d + K \cdot d)$, which is linear with respect to D and K. This makes obtaining the optimized embedding efficient.

Downstream Efficiency Gains: A practical benefit is the reduced computational cost for downstream tasks operating on the $d = 64$ dimensional embeddings compared to $D = 1024$ raw embeddings. This reduction is significant for many standard classifiers and contributes to faster downstream training and inference times. This led to substantial training time reductions for downstream models.

A.5 Justification for Principle Subspace Dimension ($d = 64$)

The choice of the principle subspace dimension $d = 64$ for the output embeddings was guided by preliminary experiments. We evaluated model performance on a validation set using various output dimensions (e.g., 32, 128, 256). $d = 64$ was found to provide a robust balance, offering significant dimensionality reduction from the input (1024 dimensions) while preserving sufficient information for effective principle discrimination in downstream tasks, yielding performance comparable to or better than higher dimensions with reduced computational cost for both model training and subsequent downstream task training/inference.

A.6 Compute Resources

All experiments, including the training of the Neural Principle Extractor and evaluation of downstream models, were conducted on a machine equipped with four NVIDIA RTX 4090 GPUs (24GB VRAM each) and 128GB of system RAM. The CPU used was an Intel(R) Xeon(R) Platinum 8336C CPU @ 2.30GHz, running on Ubuntu 24.04 LTS.

Training of the Neural Principle Extractor is computationally efficient. A full training run typically completed within 3 to 15 minutes on a single NVIDIA RTX 4090, depending on the dataset size and complexity. Using multiple GPUs can further reduce this time. Inference using the trained extractor to produce optimized embeddings is significantly faster, requiring only a single forward pass per sample. Evaluating downstream models on the optimized embeddings is also substantially more efficient than using raw embeddings, as discussed in Appendix A.4.

B Proof of Theorem 1

Proof. Under the assumption of sufficient parameterization, the network converges to a zero-residual state (Zhang et al., 2016; Allen-Zhu et al., 2019). Given $P_{\text{intra}} = 0$, the maximum topological radius for each

cluster is tightly bounded by the hyperparameter:

$$||\mathbf{m}_a - \mathbf{c}_A||_2 \leq \delta_{\text{intra}} \quad \text{and} \quad ||\mathbf{m}_b - \mathbf{c}_B||_2 \leq \delta_{\text{intra}}$$

Given $P_{\text{inter}} = 0$, the minimum distance to the negative prototype satisfies the exact margin:

$$||\mathbf{m}_a - \mathbf{c}_B||_2 \geq ||\mathbf{m}_a - \mathbf{c}_A||_2 + \delta_{\text{inter}}$$

Applying the triangle inequality, the inter-sample distance is bounded by routing through $\mathbf{m}_b$:

$$||\mathbf{m}_a - \mathbf{m}_b||_2 + ||\mathbf{m}_b - \mathbf{c}_B||_2 \geq ||\mathbf{m}_a - \mathbf{c}_B||_2 \tag{6}$$

Substituting the intra-class and inter-class bounds into Equation 6 yields:

$$||\mathbf{m}_a - \mathbf{m}_b||_2 \geq \big(||\mathbf{m}_a - \mathbf{c}_A||_2 + \delta_{\text{inter}}\big) - \delta_{\text{intra}} \tag{7}$$

Since $||\mathbf{m}_a - \mathbf{c}_A||_2 \geq 0$ by definition, the minimum inter-sample distance is:

$$||\mathbf{m}_a - \mathbf{m}_b||_2 \geq \delta_{\text{inter}} - \delta_{\text{intra}} \tag{8}$$

■

Transition to Final Embeddings. While Theorem 1 formally bounds the separation within the principle-specific mapping space ($\mathbf{m}_i$), the final projected representation $\mathbf{e}_i$ retains these geometric properties. Because the gating parameter α (Section 3.2) dynamically prioritizes $\mathbf{m}_i$ over the semantic basis $\mathbf{s}_i$, and the projection onto $\mathcal{S}^{d-1}$ is a topology-preserving L_2 normalization, the margins established between the unnormalized mappings ($\mathbf{m}_a, \mathbf{m}_b$) mathematically dictate the linear separability of the final extracted embeddings ($\mathbf{e}_a, \mathbf{e}_b$).

B.1 Proof of Theorem 2

Proof. Let $a(\mathbf{m}_i)$ be the mean distance between a sample $\mathbf{m}_i \in \mathcal{M}_A$ and all other samples in the same cluster $\mathcal{M}_A$. As established in the zero-residual condition of Theorem 1, the maximum topological radius is δ_{intra}. The maximum possible distance between any two points within this hypersphere is its diameter. Thus, the intra-cluster distance is bounded by:

$$\max a(\mathbf{m}_i) \leq 2\delta_{\text{intra}}$$

Let $b(\mathbf{m}_i)$ be the mean distance from $\mathbf{m}_i$ to all samples in the nearest differing cluster $\mathcal{M}_B$. Directly from the lower bound derived in Theorem 1 (Equation 8), the minimum distance between these clusters is:

$$\min b(\mathbf{m}_i) \geq \delta_{\text{inter}} - \delta_{\text{intra}}$$

Consequently, the *Within/Between Ratio* is mathematically upper-bounded by dividing the maximum intra-cluster distance by the minimum inter-cluster distance:

$$\text{Ratio}_{W/B} = \frac{a(\mathbf{m}_i)}{b(\mathbf{m}_i)} \leq \frac{2\delta_{\text{intra}}}{\delta_{\text{inter}} - \delta_{\text{intra}}} \tag{9}$$

Furthermore, the *Silhouette Score* is defined as $S(\mathbf{m}_i) = \frac{b(\mathbf{m}_i) - a(\mathbf{m}_i)}{\max(a(\mathbf{m}_i), b(\mathbf{m}_i))}$. Given our margin assumption $\delta_{\text{inter}} > 3\delta_{\text{intra}}$, it guarantees that the minimum inter-cluster distance remains greater than the maximum intra-cluster diameter (i.e., $b(\mathbf{m}_i) > a(\mathbf{m}_i)$). Therefore, the Silhouette Score is lower-bounded:

$$S(\mathbf{m}_i) \geq \frac{(\delta_{\text{inter}} - \delta_{\text{intra}}) - 2\delta_{\text{intra}}}{\delta_{\text{inter}} - \delta_{\text{intra}}} = 1 - \frac{2\delta_{\text{intra}}}{\delta_{\text{inter}} - \delta_{\text{intra}}} > 0 \tag{10}$$

■

Justification of Clustering Bounds and Hyperparameter Design. The condition $\delta_{\text{inter}} > 3\delta_{\text{intra}}$ in Theorem 2 is a principled design constraint. From a geometric perspective, this "3x Rule" ensures that the inter-cluster gap is sufficiently large to exceed the combined diameters of the individual clusters, which is the prerequisite for a positive Silhouette Score and a low Within/Between ratio. In practice, achieving this theoretical bound relies on the representational capacity of the dual-stream architecture (Section 3.2). By offloading contextual variance to the Semantic Basis stream ($\mathbf{s}_i$), the framework prevents the expansion of the principle-specific clusters ($\mathbf{m}_i$). This structural optimization enables the optimizer to satisfy the designed margin capacity (δ_{inter}) without gradient conflicts, optimizing the subspace to minimize the hypothesis space overlap for downstream classifiers.

C Detailed Experimental Results

This appendix provides supplementary detailed results for the experiments presented in Section 4.

C.1 GoEmotions Per-Principle Performance

This appendix provides detailed per-principle F1 performance for the GoEmotions dataset, complementing the overall results presented in Section 4.2. Table 8 shows the Mean ± Standard Deviation F1 scores for each of the five selected emotion principles.

Table 8: Per-Principle F1 Performance on GoEmotions Five-Principle Set (Mean ± Std. Dev.). Principles are abbreviated as Disappt., Sad., Disapprv., Grat., Apprv.

Principle	Emb. Type	SVM	RF	LR	XGBoost	Transformer
Disappt.	Raw Emb.	0.387 ± 0.090	0.237 ± 0.202	0.375 ± 0.054	0.359 ± 0.144	0.315 ± 0.073
	Opt. Emb.	**0.482** ± 0.102	**0.410** ± 0.087	**0.479** ± 0.106	**0.439** ± 0.099	**0.386** ± 0.117
Sad.	Raw Emb.	0.643 ± 0.059	0.728 ± 0.069	0.672 ± 0.081	0.711 ± 0.082	0.711 ± 0.087
	Opt. Emb.	**0.687** ± 0.048	**0.734** ± 0.031	**0.721** ± 0.054	**0.698** ± 0.031	**0.714** ± 0.034
Disapprv.	Raw Emb.	0.663 ± 0.074	0.691 ± 0.050	0.652 ± 0.070	0.703 ± 0.032	0.677 ± 0.055
	Opt. Emb.	**0.720** ± 0.074	**0.740** ± 0.064	**0.728** ± 0.063	**0.724** ± 0.082	**0.733** ± 0.059
Grat.	Raw Emb.	0.925 ± 0.032	0.920 ± 0.024	0.921 ± 0.020	0.905 ± 0.031	0.915 ± 0.026
	Opt. Emb.	**0.939** ± 0.021	**0.940** ± 0.028	**0.941** ± 0.024	**0.934** ± 0.032	**0.938** ± 0.029
Apprv.	Raw Emb.	0.732 ± 0.090	0.707 ± 0.031	0.727 ± 0.061	0.730 ± 0.060	0.716 ± 0.075
	Opt. Emb.	**0.769** ± 0.053	**0.747** ± 0.078	**0.771** ± 0.055	**0.762** ± 0.067	**0.758** ± 0.053

The improvements are most pronounced for semantically similar and initially challenging principles with lower initial F1 scores, such as Disappointment and Sadness. Conversely, for principles like Gratitude, which already achieved high F1 scores with raw embeddings, the relative improvement is more modest across most classifiers. These results demonstrate that our method is particularly effective at refining distinctions for principles that are difficult to classify using standard embedding techniques, raising the performance ceiling for challenging cases while maintaining strong performance on easier ones.

C.2 Amazon Reviews Per-Rating Performance

This appendix provides detailed per-rating performance for the Amazon Reviews dataset, supplementing the summarized classification and ordinal regression results presented in Section 4.2.

Table 12 shows the F1 performance for each star rating (1-5 S) on the Amazon Reviews dataset using raw and optimized embeddings.

Table 10 provides the per-rating Mean Squared Error (MSE) for the Amazon Reviews ordinal regression task.

Table 9: Classification F1 Performance per Rating on Amazon Reviews (Mean ± Std. Dev.)

Ratings	Emb. Type	SVM	RF	LR	XGBoost	Transformer
1 - S	Raw Emb.	0.712 ± 0.219	0.772 ± 0.166	0.713 ± 0.208	0.744 ± 0.235	0.731 ± 0.209
	Opt. Emb.	**0.869** ± 0.112	**0.874** ± 0.085	**0.875** ± 0.084	**0.894** ± 0.093	**0.888** ± 0.090
2 - S	Raw Emb.	0.277 ± 0.118	0.204 ± 0.213	0.297 ± 0.168	0.288 ± 0.221	0.432 ± 0.163
	Opt. Emb.	**0.691** ± 0.187	**0.708** ± 0.315	**0.667** ± 0.176	**0.760** ± 0.170	**0.711** ± 0.141
3 - S	Raw Emb.	0.433 ± 0.158	0.556 ± 0.176	0.503 ± 0.081	0.520 ± 0.141	0.584 ± 0.085
	Opt. Emb.	**0.669** ± 0.106	**0.697** ± 0.073	**0.662** ± 0.112	**0.657** ± 0.076	**0.696** ± 0.143
4 - S	Raw Emb.	0.478 ± 0.071	0.598 ± 0.114	0.565 ± 0.096	0.613 ± 0.078	0.558 ± 0.123
	Opt. Emb.	**0.650** ± 0.094	**0.639** ± 0.120	**0.637** ± 0.129	**0.622** ± 0.113	**0.620** ± 0.105
5 - S	Raw Emb.	0.614 ± 0.074	0.710 ± 0.099	0.676 ± 0.087	0.736 ± 0.069	0.724 ± 0.103
	Opt. Emb.	**0.764** ± 0.112	**0.766** ± 0.093	**0.764** ± 0.115	**0.741** ± 0.101	**0.768** ± 0.082

Table 10: Ordinal Regression MSE Performance per Rating on Amazon Reviews (Mean ± Std. Dev.)

Metric	Emb. Type	SVM	RF	LR	XGBoost	Transformer
1-S MSE	Raw Emb.	0.483 ± 0.531	0.750 ± 1.207	0.567 ± 0.533	0.957 ± 1.145	0.350 ± 0.449
	Opt. Emb.	**0.177** ± 0.260	**0.360** ± 0.543	**0.140** ± 0.254	**0.420** ± 0.555	**0.157** ± 0.250
2-S MSE	Raw Emb.	1.080 ± 0.867	1.415 ± 0.880	1.250 ± 0.972	1.465 ± 0.912	1.025 ± 1.072
	Opt. Emb.	**0.290** ± 0.386	**0.435** ± 0.547	**0.405** ± 0.334	**0.335** ± 0.338	**0.265** ± 0.363
3-S MSE	Raw Emb.	0.726 ± 0.431	0.623 ± 0.235	0.804 ± 0.378	0.712 ± 0.269	0.539 ± 0.236
	Opt. Emb.	**0.386** ± 0.260	**0.442** ± 0.312	**0.442** ± 0.367	**0.509** ± 0.361	**0.376** ± 0.261
4-S MSE	Raw Emb.	0.653 ± 0.221	0.320 ± 0.154	0.567 ± 0.211	0.340 ± 0.187	0.653 ± 0.260
	Opt. Emb.	**0.387** ± 0.171	**0.433** ± 0.196	**0.453** ± 0.195	**0.380** ± 0.161	**0.500** ± 0.189
5-S MSE	Raw Emb.	0.607 ± 0.348	0.287 ± 0.149	0.467 ± 0.163	0.247 ± 0.095	0.567 ± 0.438
	Opt. Emb.	**0.427** ± 0.389	**0.333** ± 0.365	**0.400** ± 0.394	**0.307** ± 0.200	**0.313** ± 0.193

C.3 Amazon Reviews Classification Results

This appendix section provides detailed classification performance results on the Amazon Reviews dataset, supplementing the main text discussion which focuses on ordinal regression. For this task, the 1-5 star ratings are treated as distinct discrete categories.

Table 11 summarizes the overall (average per rating) classification performance across different classifiers using both raw and optimized embeddings.

Table 11: Overall (Avg. Rating) Classification Performance on Amazon Reviews (Mean ± Std. Dev.)

Metric	Emb. Type	SVM	RF	LR	XGBoost	Transformer
Precision	Raw Emb.	0.541 ± 0.038	0.627 ± 0.093	0.595 ± 0.058	0.630 ± 0.073	0.639 ± 0.067
	Opt. Emb.	**0.728** ± 0.077	**0.726** ± 0.094	**0.716** ± 0.084	**0.718** ± 0.077	**0.725** ± 0.062
Recall	Raw Emb.	0.522 ± 0.043	0.628 ± 0.083	0.583 ± 0.055	0.634 ± 0.060	0.628 ± 0.064
	Opt. Emb.	**0.721** ± 0.073	**0.729** ± 0.080	**0.715** ± 0.078	**0.713** ± 0.069	**0.723** ± 0.056
Avg. F1	Raw Emb.	0.521 ± 0.041	0.609 ± 0.082	0.582 ± 0.052	0.619 ± 0.059	0.622 ± 0.061
	Opt. Emb.	**0.717** ± 0.074	**0.721** ± 0.085	**0.710** ± 0.081	**0.708** ± 0.069	**0.718** ± 0.058

Table 12 presents the F1 performance for each individual star rating (1-5) using both raw and optimized embeddings. Optimized embeddings generally show improved performance across most individual ratings, particularly for the intermediate ratings (2, 3, 4 stars) which are often more challenging to distinguish.

Table 12: Classification F1 Performance per Rating on Amazon Reviews (Mean ± Std. Dev.)

Ratings	Emb. Type	SVM	RF	LR	XGBoost	Transformer
1 - S	Raw Emb.	0.712 ± 0.219	0.772 ± 0.166	0.713 ± 0.208	0.744 ± 0.235	0.731 ± 0.209
	Opt. Emb.	**0.869** ± 0.112	**0.874** ± 0.085	**0.875** ± 0.084	**0.894** ± 0.093	**0.888** ± 0.090
2 - S	Raw Emb.	0.277 ± 0.118	0.204 ± 0.213	0.297 ± 0.168	0.288 ± 0.221	0.432 ± 0.163
	Opt. Emb.	**0.691** ± 0.187	**0.708** ± 0.315	**0.667** ± 0.176	**0.760** ± 0.170	**0.711** ± 0.141
3 - S	Raw Emb.	0.433 ± 0.158	0.556 ± 0.176	0.503 ± 0.081	0.520 ± 0.141	0.584 ± 0.085
	Opt. Emb.	**0.669** ± 0.106	**0.697** ± 0.073	**0.662** ± 0.112	**0.657** ± 0.076	**0.696** ± 0.143
4 - S	Raw Emb.	0.478 ± 0.071	0.598 ± 0.114	0.565 ± 0.096	0.613 ± 0.078	0.558 ± 0.123
	Opt. Emb.	**0.650** ± 0.094	**0.639** ± 0.120	**0.637** ± 0.129	**0.622** ± 0.113	**0.620** ± 0.105
5 - S	Raw Emb.	0.614 ± 0.074	0.710 ± 0.099	0.676 ± 0.087	0.736 ± 0.069	0.724 ± 0.103
	Opt. Emb.	**0.764** ± 0.112	**0.766** ± 0.093	**0.764** ± 0.115	**0.741** ± 0.101	**0.768** ± 0.082

D Additional Experimental Details

D.1 Details on Used Assets and Licenses

This appendix provides details on the licenses and terms of use for the external datasets, embedding models, and language models used in this research, as referenced from the main paper. Our use of these assets adheres to their respective licenses and terms.

Datasets.

- **GoEmotions Dataset** (Demszky et al., 2020): This dataset is released under the **Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0)**. Available at `https://github.com/google-research/goemotions`.
- **Amazon Reviews Dataset** (Ni et al., 2019): This dataset is provided for research purposes. Its use is subject to the terms specified by the data providers (e.g., Stanford/UCSD). Researchers should refer to the original source for specific usage guidelines. Available via the cited research project website.
- **Toxic Comment Classification Challenge**: This dataset, originally hosted on Kaggle (cjadams et al., 2017), is made available under the **CC0 1.0 Universal Public Domain Dedication**. Available at `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`.

Embedding Model.

- **Jina Embeddings v3** (Sturua et al., 2024): The embeddings used were generated by the `jina-embeddings-v3` model. Jina AI models are typically licensed under the **Apache 2.0 License**. Researchers should consult the official Jina AI model documentation or Hugging Face model card for the most precise license information and terms of use.

Large Language Models (for Comparison).

- **LLama 3.3** (Grattafiori et al., 2024): The Llama 3 family of models is available under the **Llama 3 Community License**. Use of the quantized version (`grattafiori2024llama_70b_Q4_K`) adheres to the terms of this license.
- **DeepSeek-Chat-v3** (Liu et al., 2024): Used via API. Use is subject to **DeepSeek AI's API Terms of Service**.
- **Gemini 2.5 Pro** (Comanici et al., 2025): Used via API. Use is subject to **Google's API Terms of Service** (e.g., Google AI or Google Cloud terms).

D.2 Data Distribution

In strict adherence to the 10-fold cross-validation protocol outlined in Section 4.1, the datasets were dynamically partitioned during training and evaluation. The total pool of valid samples utilized across all folds for each dataset is as follows: the GoEmotions five-principle subset comprises 6,857 samples; the Amazon Reviews sampled subset comprises 3,028 samples; and the resampled Toxic Comment Classification dataset comprises 81,948 samples. Class imbalance ratios were preserved across all dynamic splits to rigorously test the framework's robustness.