UNDERSTANDING GENERATIVE CONTRASTIVE LEARNING VIA THE CONDITIONAL GENERATION KERNEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in unsupervised text embedding leverage the stochasticity of autoregressive generation by LLMs to generate semantically related positive pairs for contrastive learning. The resulting paradigm, which we term *generative contrastive learning*, has the unique property that the data-generating distribution is explicitly parametrized by the LLM, thus opening the door to a formal analysis of the contrastive objective. Building on this, we introduce the *Conditional Generation Kernel* (CGK), a positive semi-definite kernel that measures textual similarity through expected co-generation probabilities across a set of anchor documents. We show that the standard InfoNCE loss optimizes dot-product similarities toward the *logarithm* of the CGK, which exposes a fundamental representational bottleneck: because the logarithm of a kernel matrix is generally indefinite, the optimal target cannot be exactly realized as an inner product in any Euclidean space. Motivated by this bottleneck, we derive a contrastive objective whose global optimum is exactly the CGK and find that it aligns with the known spectral contrastive loss. We empirically confirm that the two objectives actually approach their theoretically derived optimal states, resulting in clearly differing geometries of the resulting embedding spaces. Finally, we evaluate the resulting models on downstream tasks, demonstrating that training with the spectral loss yields favorable properties for clustering and classification.

1 INTRODUCTION

Contrastive learning is a dominant paradigm within representation learning, being used to train both state-of-the-art text embedding models (Wang et al., 2024; Xiao et al., 2024; Lee et al., 2025a;b; Zhang et al., 2025) and visual and multimodal encoders (Chen et al., 2020; He et al., 2020; Radford et al., 2021; Tschannen et al., 2025; Li et al., 2026). Central to this training paradigm is a distribution over *positive pairs*, i.e., inputs that should be mapped to nearby representations, as they largely determine the semantic structure of the resulting embedding space. Existing theory makes this dependence precise by characterizing the optima of contrastive objectives in terms of the positive-pair distribution, finding that the commonly used InfoNCE loss drives similarity scores toward the pointwise mutual information of positive pairs (van den Oord et al., 2019; Poole et al., 2019). In practice, however, this underlying distribution is usually just implicitly defined by the construction of the training dataset, so that the optima predicted by theory cannot be estimated reliably.

A recent line of work, which we term *generative contrastive learning*, instead obtains positive pairs by sampling multiple outputs from a generative model under the same conditioning input, such as images generated from the same caption (Tian et al., 2023; 2024) or LLM-generated summaries of the same scientific abstract (Brinner & Zarrieß, 2025; Brinner & Zarrieß, 2026). A key differentiating property of the text-based variant in particular is that its positive-pair distribution is explicitly parametrized by the LLM itself. Theoretical analyses can therefore be grounded in a concretely defined data-generating distribution, and the optima they predict can be approximated directly using LLM-derived probabilities, thus offering a unique setup for empirical validation.

In this work, we exploit this property to study generative contrastive learning both theoretically and empirically. To this end, we first introduce the *Conditional Generation Kernel* (CGK; Section 2),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

which measures the similarity of two texts through co-generation probabilities and closely resembles kernels that prior work has identified as targets within contrastive learning (Johnson et al., 2022; Marjieh et al., 2024). Unlike for these kernels, however, we can estimate the CGK directly on natural text, which allows us to empirically validate it as a meaningful optimization target whose quality improves predictably with the number of anchors.

Building on this kernel, our theoretical analysis of SemCSE (Section 3) shows that, at its optimum, the InfoNCE objective drives similarity scores toward the logarithm of the CGK. Estimating this target empirically reveals that it is highly indefinite for real text and thus cannot be realized exactly by dot products. This motivates us to derive a loss whose optimum is the CGK itself (Section 4), recovering the known spectral contrastive loss of Chen et al. (2021) within this novel context. While the theoretical targets of both the InfoNCE and spectral objectives have been explored for different setups, the explicitly computable CGK uniquely allows us to validate whether theoretically derived optimal states are actually recovered in practice, which we do in Section 5 by evaluating both objectives in increasingly realistic training scenarios. Finally, we evaluate both objectives on downstream tasks, demonstrating that our specifically designed architecture that decouples directional from norm prediction is crucial for making the spectral loss competitive.

2 The Conditional Generation Kernel

In this section, we introduce the *Conditional Generation Kernel* (CGK), a positive semi-definite kernel that measures textual similarity through LLM-based conditional probabilities computed relative to a set of anchor documents and that can be used in practice for similarity-based retrieval.

2.1 Definition

Let $\mathcal{V}$ be a finite vocabulary, and let $\mathcal{X}_A = \bigcup_{\ell=1}^{L_A} \mathcal{V}^\ell$ and $\mathcal{X}_C = \bigcup_{\ell=1}^{L_C} \mathcal{V}^\ell$ denote the sets of token sequences of length at most L_A and L_C, representing the spaces of *anchor* and *continuation* texts, respectively. The CGK measures similarity $k(x, x')$ between continuations $x, x' \in \mathcal{X}_C$ using a finite set of anchors $\mathcal{A} \subset \mathcal{X}_A$ as reference points.

Central to the definition of the CGK is the conditional probability of generating a continuation $x \in \mathcal{X}_C$ for a given anchor $a \in \mathcal{X}_A$. We model this generative process as first drawing a sequence length ℓ uniformly from $\{1, \dots, L_C\}$, followed by sampling ℓ tokens autoregressively from the LLM, yielding:

$$p(x \mid a) \;=\; \frac{1}{L_C} \prod_{t=1}^{|x|} p(x_t \mid a,\, x_{<t})\,. \tag{1}$$

For the finite set of anchors $\mathcal{A} = \{a_1, \dots, a_m\} \subset \mathcal{X}_A$ equipped with a probability measure μ, the *marginal generation probability* of a continuation $x \in \mathcal{X}_C$ under this model is

$$p(x) \;=\; \mathbb{E}_{a\sim\mu}\big[p(x \mid a)\big] \;=\; \sum_{j=1}^{m} \mu(a_j)\, p(x \mid a_j). \tag{2}$$

From these components, we define the *Conditional Generation Kernel* $k\colon \mathcal{X}_C \times \mathcal{X}_C \to \mathbb{R}$ as

$$k(x, x') \;=\; \frac{\mathbb{E}_{a\sim\mu}\big[p(x \mid a)\, p(x' \mid a)\big]}{p(x)\, p(x')}. \tag{3}$$

Thus, $k(x, x')$ is the exponentiated pointwise mutual information between x and x' under the joint co-generation distribution $p(x, x') = \mathbb{E}_{a\sim\mu}[p(x \mid a)\, p(x' \mid a)]$, whose marginals are exactly $p(x)$ and $p(x')$. Intuitively, two continuations are similar if they are likely generated from the same anchors, while the normalization by $p(x)$ and $p(x')$ discounts generic continuations that are likely under most anchors.

Importantly, the kernel can be explicitly realized as an inner product $k(x, x') = \langle \phi(x), \phi(x') \rangle$ via the following feature map:

$$\phi\colon \mathcal{X}_C \to \mathbb{R}^m, \qquad \phi(x)_j \;=\; \frac{\sqrt{\mu(a_j)}\; p(x \mid a_j)}{p(x)}, \qquad j = 1, \dots, m. \tag{4}$$

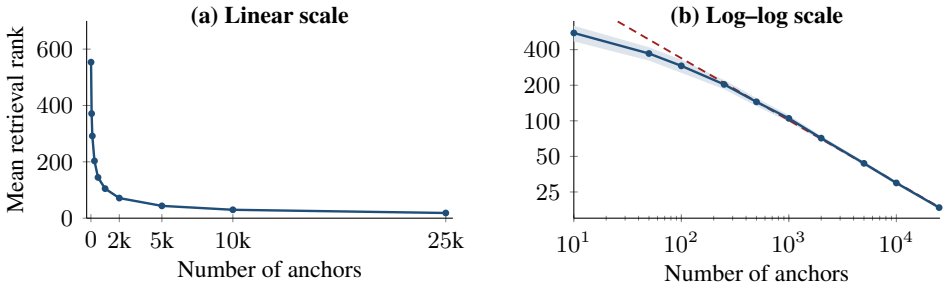

Figure 1: Retrieval performance of the CGK as a function of the number of anchors $|\mathcal{A}|$.

This finite-dimensional inner product representation demonstrates that the CGK is a positive semi-definite kernel. Note that the length prior $1/L_C$ from Eq. equation 1 scales both $p(x \mid a_j)$ and $p(x)$ equally, thus canceling in Eq. equation 3 and in this feature map.

Via Bayes' rule, the feature map can also be written as $\phi(x)_j = p(a_j \mid x)/\sqrt{\mu(a_j)}$, showing that the CGK compares texts through their posteriors over anchors and thus belongs to the family of kernels derived from generative models (Jaakkola & Haussler, 1998; Jebara et al., 2004). Specifically, it corresponds to both the positive-pair kernel of Johnson et al. (2022) and the generative similarity of Marjieh et al. (2024), but using an LLM as generative model. While generative similarity is generally intractable, the LLM provides the conditional probabilities $p(x \mid a)$ explicitly, so that the CGK can be computed exactly over a finite anchor set. This exact computability is the key feature that enables our subsequent analyses.

2.2 Empirical Evaluation

Given a set of anchor documents, the feature map ϕ is exactly computable, which allows us to evaluate the CGK empirically as a measure of textual similarity. We do so in a retrieval setting, testing whether its similarity ranking reliably identifies semantically related texts and how performance depends on the number of anchors underlying ϕ. Following SemCSE (Brinner & Zarrieß, 2025), the central subject of our subsequent analyses, we use 2,000 scientific abstracts from the SciRepEval benchmark (Singh et al., 2023) and generate two summarizing continuations for each, prompted by the phrase *"A comprehensive summary for our work would be that"*. As anchors for computing ϕ, we use 28,000 further abstracts from the same corpus, and compute $p(s \mid a)$ for each combination of summary s and anchor a using the same LLM as used for generation. The resulting conditional probabilities then allow for computing $\phi(s)$ and thus the empirical kernel Gram matrix $\widehat{K}$.

We then define the retrieval task as matching each summary to its counterpart via the feature vectors ϕ and report the average rank at which it is retrieved among 1,946 candidates (54 pairs were removed by filtering). Importantly, subsampling the anchor set allows us to measure how retrieval performance depends on the number of anchors $|\mathcal{A}|$ used to compute the representation ϕ.

As shown in Figure 1, retrieval performance improves consistently as the anchor set grows, with the mean rank following a clear power law (rank $\propto |\mathcal{A}|^{-0.53}$), dropping to 18.44 at 25,000 anchors, while the median rank reaches a near-perfect 2.0. The mean is therefore dominated by a small subset of poorly matched summaries, for which the feature vectors empirically show increased entropy, which indicates that the anchor set contains no specifically informative anchor for them. Since larger anchor sets are more likely to contain such anchors, and the power law shows no sign of saturation within the evaluated range, we expect retrieval performance to approach near-perfect levels with substantially larger anchor counts. Full experimental details and a more detailed analysis are provided in Appendix B.1.

3 Analyzing Generative Contrastive Learning

This section analyzes generative contrastive learning using the example of SemCSE (Brinner & Zarrieß, 2025), which creates positive pairs for contrastive learning by sampling multiple continuations of the same anchor document from an LLM-induced distribution.

3.1 SemCSE Definition

Recent work by Brinner & Zarrieß (2025) introduced SemCSE, an unsupervised framework for training text embedding models that leverages the stochasticity of autoregressive generation in LLMs. Given a set of anchor documents, the method samples multiple continuations for each anchor from the LLM's conditional distribution, and then trains the model to map continuations of the same anchor to nearby positions in the embedding space using a contrastive loss.

To formally analyze this training paradigm, we align it with the notation established in Section 2. Let $\mathcal{A}$ denote a set of anchor documents distributed according to probability measure μ. For a given sampled anchor $a \sim \mu$, we construct a positive pair by drawing two independent continuations $x, x' \sim p(\cdot \mid a)$ given by an LLM. To create negatives, we independently draw M anchors and sample one continuation from each, which is equivalent to drawing M continuations directly from the marginal generation distribution, $x_i^- \overset{\text{iid}}{\sim} p(\cdot)$.

Let $s_{xx'}$ denote the similarity score (e.g., dot product or cosine similarity) between the embeddings of x and x' as predicted by the embedding model, and let $\tau > 0$ be a temperature hyperparameter. Under this sampling scheme, the SemCSE objective using the InfoNCE loss (van den Oord et al., 2019) with M negative samples is formulated as:

$$\mathcal{L}_M \;\coloneqq\; \mathbb{E}_{\substack{a\sim\mu \\ x,x'\sim p(\cdot|a)}} \left[\mathbb{E}_{x_1^-,\ldots,x_M^- \overset{\text{iid}}{\sim} p(\cdot)} \left[-\log \frac{e^{s_{xx'}/\tau}}{e^{s_{xx'}/\tau} + \sum_{i=1}^{M} e^{s_{xx_i^-}/\tau}} \right] \right]. \tag{5}$$

3.2 Characterizing the Optimal State

To understand the optimal state this loss targets, we analyze the behavior of the log-normalized loss $\mathcal{L}_M - \log M$ in the limit of infinite negative samples, following Wang & Isola (2020). Computing the limit and solving for the stationary point yields the following result:

Theorem 1 *Let M denote the number of negative samples. In the limit $M \to \infty$, the log-normalized InfoNCE loss $\mathcal{L}_M - \log M$ is minimized over all symmetric similarity functions by*

$$s^*_{xx'} \;=\; \tau \log k(x, x') \;+\; C, \tag{6}$$

where C is a constant independent of x and x'.

The proof is given in Appendix A.1. Since the logarithm is strictly monotonic, it follows that the CGK plays a central role in textual generative contrastive learning, as optimal similarity scores exactly preserve the CGK's ranking. We note that a similar result has been obtained by Johnson et al. (2022), who show that a positive-pair kernel arises as the central object of contrastive learning with random image augmentations under exponentiated similarities.

While this result establishes that SemCSE targets a semantically meaningful similarity structure in the form of the CGK, it also reveals a fundamental limitation. In practice, the similarity scores $s_{xx'}$ are not free parameters but inner products of model-produced embeddings, $s_{xx'} = \langle f(x), f(x') \rangle$. The matrix of all such similarities over a finite set of inputs is therefore a Gram matrix and must be positive semi-definite. The optimal target identified by Theorem 1, however, is $\log K$, which is itself generally *not* positive semi-definite. The optimum is thus unreachable by any dot-product-based embedding model, leaving an irreducible approximation error between the model's similarities and the targeted state. Importantly, this approximation error can be empirically quantified due to the CGK being estimable in practice, which contrasts with related settings in which the underlying kernel is either intractable or only computable for synthetic distributions (Johnson et al., 2022; Marjieh et al., 2024).

3.3 Reconstruction Analysis

To empirically quantify the irreducible error of reconstructing $\log K$, we use the data from Section 2.2 to compute an empirical Gram matrix $\widehat{K} \in \mathbb{R}^{1946\times 1946}$ by using ϕ for just the first summary of each of the $N = 1{,}946$ available pairs. We then compute the best possible dot-product reconstructions of both $\widehat{K}$ and $\log \widehat{K}$ using an embedding dimension $d = N$, so that capacity imposes no

constraint, leaving positive semi-definiteness as the only remaining obstruction. The optimal reconstruction in Frobenius norm is then simply obtained by discarding all negative eigenvalues. Since $\widehat{K}$ is a full-rank Gram matrix ($\lambda_{\min} = 1.7 \times 10^{-4}$), it can be reconstructed exactly. In contrast, 904 of the $1{,}946$ eigenvalues of $\log \widehat{K}$ (46.5%) are negative, carrying 16.6% of its spectral mass and thus leading to a substantial loss in information if reconstructed via a dot product.

To measure how this reconstruction error affects the similarity structure, we evaluate how well the reconstruction preserves each summary's n nearest neighbors under $\widehat{K}$ via NDCG@n. The reconstruction of $\log \widehat{K}$ yields an NDCG@1 of just 0.375, meaning that the nearest neighbor is preserved for only 37.5% of samples, and extended rankings show similar losses (NDCG@10 $= 0.658$, NDCG@50 $= 0.784$). Although in practice InfoNCE does not minimize the Frobenius distance to its target and may thus preserve different properties of $\log K$, these values nevertheless illustrate the fundamental bottleneck inherent to this optimization target.

4 Optimizing Toward K

Motivated by the irreducible approximation error inherent to targeting $\log K$, we derive a loss that, like InfoNCE, relies only on the similarity scores of sampled pairs, but whose optimum is the CGK itself, thus making this optimum exactly realizable given sufficient embedding dimensionality.

To construct this loss, we start with the fundamental optimality condition that the gradient of the loss must vanish exactly at $s_{xx'} = k(x, x')$ (Moustakides & Basioti, 2019). This condition holds precisely for gradients of the form $\partial\mathcal{L}/\partial s_{xx'} = c(x, x')\,(s_{xx'} - k(x, x'))$ with a positive weighting function c. Choosing $c(x, x') = p(x)p(x')$ cancels the denominator of the kernel k and therefore yields $\partial\mathcal{L}/\partial s_{xx'} = p(x)p(x')\,s_{xx'} - p(x, x')$. Since this form constitutes the derivative of the intended loss, we integrate with respect to $s_{xx'}$ and sum over all pairs to get the following objective:

$$\mathcal{L}_{\text{spectral}} \;=\; -\mathbb{E}_{\substack{a\sim\mu \\ x,x'\sim p(\cdot|a)}}\big[s_{xx'}\big] \;+\; \frac{1}{2}\,\mathbb{E}_{x,x'\sim p(\cdot)}\big[s_{xx'}^2\big], \tag{7}$$

which is strictly convex in the scores and uniquely minimized by $s^*_{xx'} = k(x, x')$ (Appendix A.2). Also, up to an additive constant, the loss is equivalent to a weighted least-squares regression of the similarity scores onto k (Appendix A.3).

This resulting loss derived under the motivation of realizability within generative contrastive learning ultimately coincides with the spectral contrastive loss of Chen et al. (2021), who instead derive it from a spectral decomposition of an augmentation graph, and whose optimum was later identified as the positive-pair kernel by Johnson et al. (2022). Following the existing literature, we also refer to this loss as the *spectral loss* throughout this work.

5 Do Contrastive Losses Realize Their Target Geometries?

While previous work has analyzed theoretical properties of both the spectral and InfoNCE losses (Chen et al., 2021; Johnson et al., 2022), the kernels involved are generally intractable, thus making an empirical verification of the identified properties infeasible. In our setup, by contrast, the CGK estimated from a finite set of anchors provides an empirical approximation of the optimization targets of both losses. This allows us to test whether the two losses actually lead to different *target geometries* in practice, i.e., embedding spaces whose similarity scores follow the linear structure of K for the spectral loss or the logarithmic structure of $\log K$ for InfoNCE. To identify factors that limit the realization of the specific target geometries, we start with a simplified setting in which vectors are optimized freely and progressively move toward full contrastive training of text encoders.

5.1 Evaluating Embedding Geometry

To determine whether a model realizes a linear or logarithmic embedding geometry, we use the empirical Gram matrix $\widehat{K} \in \mathbb{R}^{1946\times 1946}$ of Section 3.3 and measure whether the model's predicted similarities C align more closely with the linear target $\widehat{K}$ or the logarithmic target $\log \widehat{K}$.

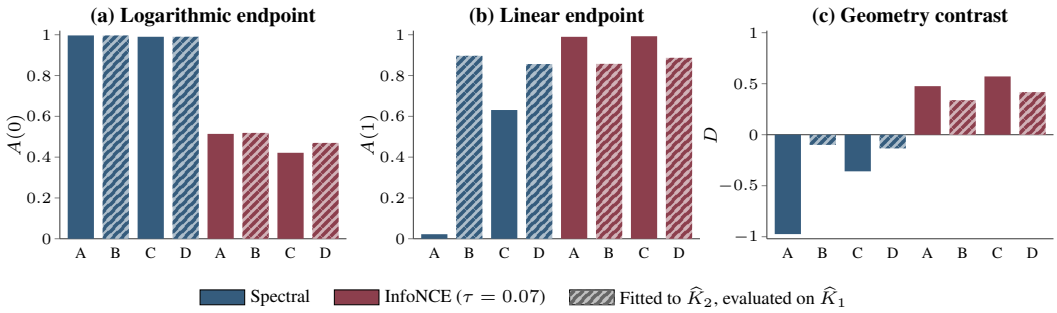

Figure 2: Alignment of free-vector and transformer fits trained directly on kernel targets $\widehat{K}$ at dimension $d = 768$, for both the Spectral objective and InfoNCE ($\tau = 0.07$). Each objective is evaluated under four different setups: free vector fits without noise (A) and with noise (B), and transformer fitted without noise (C) and with noise (D). Lower values of $A(0)$ and $A(1)$ indicate a closer fit to the logarithmic and linear target geometries, respectively; negative values of $D = A(1) - A(0)$ indicate a preference for the linear geometry, whereas positive values indicate logarithmic preference.

The core idea of our proposed metric is to measure how much of the variance in C is explained by the best positive-affine transformation of either the linear target $T^{(1)} = \widehat{K}$ or the logarithmic target $T^{(0)} = \log \widehat{K}$, evaluated as follows:

$$A(p) = 1 - \frac{[\mathrm{Cov}_\Omega(C, T^{(p)})]_+^2}{\mathrm{Var}_\Omega(C)\ \mathrm{Var}_\Omega(T^{(p)})}, \qquad D = A(1) - A(0). \tag{8}$$

Here, $A(p)$ is the linear ($p = 1$) or logarithmic ($p = 0$) alignment error and denotes the fraction of variance in C left unexplained by target $T^{(p)}$. A negative alignment contrast $D < 0$ thus indicates alignment with the linear geometry, and $D > 0$ indicates alignment with the logarithmic one.

Because $\widehat{K}$ is estimated from a finite set of anchors, the noise underlying its estimation increases $A(p)$ whenever a model is not fitted to $\widehat{K}$ itself, so that in these cases we use a noise-corrected version of the proposed metric. Details on the rationale, derivation, noise-correction and an empirical evaluation of the proposed metric are available in Appendix B.2.

5.2 Kernel-Informed Geometry Recovery

Figure 2 displays the logarithmic error $A(0)$, the linear error $A(1)$, and the geometric contrast D in separate subplots, with the setup evaluated here denoted as setup **A** and blue and red bars representing spectral and InfoNCE fits, respectively. We find that the spectral fit tracks the linear geometry almost perfectly ($A(1) = 0.022$) while not adhering to the logarithmic one at all ($A(0) = 0.997$). InfoNCE shows the opposite behavior: it does not align with the linear geometry at all ($A(1) = 0.990$) and tracks the logarithmic one more closely ($A(0) = 0.514$), though imperfectly, since $\log K$ is highly indefinite. In additional experiments, we find that variants of both objectives that fit the similarities directly to exact kernel values, instead of inferring them from sampled positives, yield nearly identical results (Appendix B.3). **The stochastic signal of sampled positives thus suffices for both objectives to optimally recover their respective target geometries**.

In the previous experiment, the embeddings were fitted and evaluated against the same matrix $\widehat{K}$. In realistic training, however, models optimize toward the true kernel K, while evaluation is only possible against a noisy estimate of it. To emulate this, we compute two kernel estimates from disjoint anchor sets, fit the objectives to one of them ($\widehat{K}_2$), and evaluate against the other ($\widehat{K}_1$) using the noise-corrected alignment metrics. We find that this estimation noise lowers the alignment with the respective target geometries, most notably for the spectral loss (spectral $A(1)$: $0.022 \to 0.897$; InfoNCE $A(0)$: $0.514 \to 0.519$; setup **B**). However, this is largely caused by the limited agreement between $\widehat{K}_1$ and $\widehat{K}_2$ themselves, which fundamentally limits the alignment attainable by any fit. Nevertheless, as alignment contrasts consistently identify the correct target, **the realized geometries remain clearly identifiable under noisy evaluation targets**.

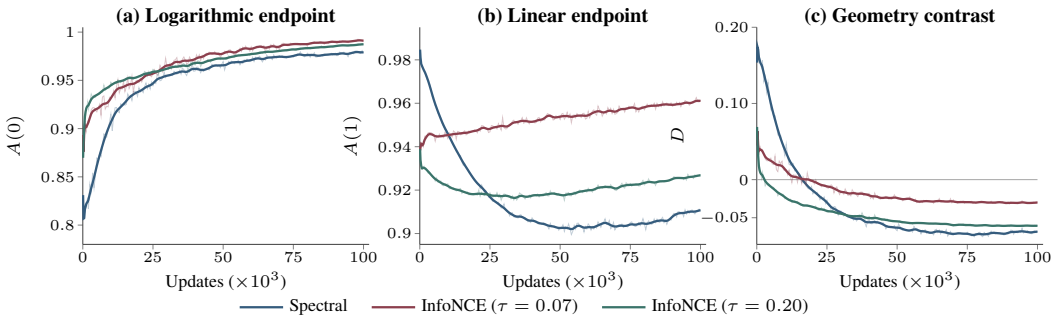

Figure 3: Noise-corrected alignment of predicted embedding similarities vs. $\widehat{K}$ over training.

Finally, we replace the free vectors by a transformer encoder ($d = 768$) that embeds the summary texts themselves, thus introducing the altered optimization dynamics caused by shared encoder parameters. We again train with the sampled objectives, both in the noise-free (setup **C**) and noisy (setup **D**) conditions. Despite the substantially changed optimization dynamics caused by shared encoder parameters, we find that **in both the noise-free and noisy settings, the encoder realizes the geometry predicted for its objective**.

Implementation details, in-depth evaluations and additional results for different embedding dimensionalities are included in Appendix B.3 and B.4.

5.3 Geometry Realization under Standard Contrastive Training

The previous experiments relied on explicit estimates of the CGK to sample positive pairs. We now replace this kernel supervision by sampling positives directly from the joint distribution $p(x, x')$, i.e., we train with the standard spectral and InfoNCE objectives on the original SemCSE training data (Brinner & Zarrieß, 2025), where positives are two summaries generated for the same scientific abstract. Note that the theoretical optima remain exactly K and $\log K$ for spectral and InfoNCE, so that alignment with $\widehat{K}$ remains a meaningful measure of the realized geometry. We track this noise-corrected alignment on the unseen documents of $\widehat{K}$ throughout training, with this progression being visualized in Figure 3.

Under these realistic conditions, only the spectral loss retains a clear alignment with its target geometry. It steadily improves its linear alignment and plateaus at $A(1) \approx 0.90$, close to the attainable optimum imposed by the noise of estimating $\widehat{K}$. In contrast, all models rapidly lose their alignment with the logarithmic target, which explains less than 3% of the variance by the end of training. Both InfoNCE variants therefore end with a negative contrast D, but this seems to rather reflect the loss of logarithmic alignment more than a genuine preference for the linear geometry. We attribute these findings to the following factors:

1. **Only direct neighborhood structure is learned.** All models reliably identify the nearest neighbors under $\widehat{K}$, reaching an NDCG@50 of 0.31–0.34. However, they fail to reproduce the ranking among the many low-similarity pairs, which cannot be reliably estimated from the stochastic minibatch gradients.
2. **Adaptation to the training data.** Documents repeat after fewer than 3K batches, after which all models gradually specialize to the training set, and their alignment with $\widehat{K}$ declines. We find that a stronger repulsion of hard negatives leads to quicker adaptation to the training data.
3. **Temperature constrains the approximation of the logarithmic target.** With unit-normalized embeddings, InfoNCE logits lie in $[-1/\tau, 1/\tau]$. The unconstrained optimum, however, requires logit differences equal to differences in $\log K$, independently of temperature. Increasing τ therefore potentially forces distortions of the realized target's relative spacing; a behavior also observed for higher-temperature free-vector fits (Figure 4).

For the experiment setup and a detailed analysis of these results, see Appendix B.4.3.

5.4 Effects on Downstream Performance

Finally, we also evaluate how the different objectives affect downstream performance. To this end, we train Qwen3.5-9B (Qwen Team, 2026) as an embedding model following the exact setup of the original SemCSE work (Brinner & Zarrieß, 2025), comparing the spectral loss to InfoNCE at $\tau \in \{0.07, 0.2\}$ and to two established InfoNCE variants, SimACE (Jeong et al., 2024) and the filtering of hard in-batch negatives (Solatorio, 2024). We evaluate all models on the SciRepEval benchmark (Singh et al., 2023) and on the semantic benchmark of Brinner & Zarrieß (2025) (full details in Appendix B.5).

First, we find that applying the spectral loss directly to unnormalized encoder outputs yields substantially worse results than all other objectives on both benchmarks (e.g., SciRepEval: 62.51 vs. 65.45 for InfoNCE). We attribute this to the coupling between embedding norms and directions: since both are determined by the same output coordinates, updates that adjust the norm of one embedding can also distort the general directional structure of the embedding space. To reduce this interference, we propose a decoupled architecture that produces a unit-normalized directional embedding and predicts its norm via a separate linear head followed by a softplus. In this way, we recover the dot products as the product of both norms and the cosine similarity of the directions. Norms can thus be adjusted without substantially affecting the directional output, while the encoder remains shared. Only using this decoupling, the spectral loss becomes competitive, reaching a SciRepEval score of 65.68.

Using the decoupled architecture for the spectral loss, we find that all objectives perform similarly in terms of overall benchmark averages, but differ considerably in the task types they favor. The spectral loss and InfoNCE at $\tau = 0.2$ outperform the other objectives on clustering and classification tasks, while the low-temperature variants prefer retrieval settings. We explain this observation by how strongly each objective repels hard negatives: since true labels are unknown during training, hard negatives frequently belong to the same semantic cluster as the anchor, so that strongly repelling them increases the likelihood of breaking up clusters of related documents. Strong repulsion of hard negatives can therefore benefit fine-grained discrimination, while clustering and classification benefit from reduced repulsion. As detailed in Appendix B.5, we find that InfoNCE repels hard negatives exponentially more strongly than easy ones, while this ratio grows only linearly for the spectral loss. Raising the temperature, however, flattens this exponential sufficiently to restore good clustering performance (Wang & Liu, 2021). We therefore attribute the different performance profiles to these specific optimization dynamics instead of a particular property of a realized target geometry.

5.5 Determining the Ranking-Optimal Geometry

Our use of the spectral loss was motivated by the better realizability of its linear target. In practice, however, we found no substantial performance differences between the objectives that could be attributed to their target geometries. This raises the question of whether the linear geometry offers any benefit at all, given that the CGK is of very high rank and can thus only be approximated by low-dimensional embeddings anyways, irrespective of the targeted geometry.

To answer this, we determine which geometry better preserves the ranking structure of $\widehat{K}$ under constrained dimensionality. We optimize free vectors of dimension $d \in \{64, 128, 256, 1946\}$ with a margin loss that solely penalizes violations of the ranking defined by $\widehat{K}$, but only considering for each sample its top-n neighbors according to $\widehat{K}$ (with $n \in \{10, 50, 200\}$). Since $\widehat{K}$ and $\log \widehat{K}$ induce the exact same ranking, this objective favors neither geometry, so that the geometry the resulting embeddings adopt reveals which one is best suited for preserving rankings. We then evaluate these fits with the alignment errors $A(1)$ and $A(0)$ (details in Appendix B.6).

The ranking-optimal embeddings consistently favor the logarithmic geometry over the linear one, across both dot-product and cosine similarities and all evaluated dimensions. For $n \geq 50$, the contrast D ranges from 0.07 to 0.65 in favor of the logarithmic geometry, while the linear geometry never explains more than 7% of the variance. We explain this by the distribution of kernel values. The linear kernel is highly skewed, with a few large similarities dominating its variance, whereas a ranking objective favors evenly spaced similarities, and the logarithm approximates such a spacing far more closely. Under limited embedding dimensionality, a logarithmic geometry is therefore

better suited to preserve the ranking of the most similar pairs, which may offset the realizability advantage of the linear target in practice.

6 Related Work

Optimization Targets and Their Realizability. Our analysis relates closely to the well-known result that InfoNCE drives similarity scores toward the pointwise mutual information of the positive-pair distribution (van den Oord et al., 2019; Poole et al., 2019). More generally, contrastive objectives derived from different f-divergences target different functions of the same density ratio (Nguyen et al., 2010; Lu et al., 2023), e.g., logarithmic for InfoNCE and affine for the Pearson χ^2 divergence, to which the spectral loss is closely related. The spectral loss also appears in the context of least-squares density-ratio estimation (Kanamori et al., 2009), specifically for the ratio between a joint distribution and the product of its marginals (Suzuki et al., 2009). We add a representational perspective, analyzing the inability of dot products to realize the logarithmic target, which differs from the rank-based limitation identified by Uesaka et al. (2025).

Kernels Underlying Contrastive Learning. Several works characterize contrastive learning through kernels induced by the positive-pair distribution. The CGK in particular corresponds to the positive-pair kernel of Johnson et al. (2022), which characterizes the optima of both spectral and InfoNCE-type objectives, and to the generative similarity of Marjieh et al. (2024). It is also closely related to the normalized augmentation graph from which Chen et al. (2021) derive the spectral loss. While these works evaluate their kernels in controlled settings with known generative processes, we compute the kernel directly from LLM likelihoods, which enables an analysis of target realizability and learned geometry on natural text. Related representations evaluate estimated density ratios at landmark documents (Tosh et al., 2021) or compare texts through the likelihoods of their possible continuations (Liu et al., 2024), whereas the CGK compares texts by how likely they are to be generated from the same anchors.

Finally, our work joins a line of related studies showing that theoretical characterizations of contrastive learning do not fully describe trained models (Rusak et al., 2025; Saunshi et al., 2022; HaoChen & Ma, 2023), a gap that the explicit kernel in our setting allows us to quantify directly.

7 Conclusion

In this work, we analyzed generative contrastive learning using SemCSE as an example, showing that the CGK characterizes the optima of both InfoNCE and the spectral loss.

While related analyses exist, the central contribution of our work lies in analyzing a unique contrastive learning paradigm in which a complex optimization target can be estimated empirically, enabled by the explicit parametrization of the positive-pair distribution by an LLM. Computing the CGK directly on natural text then allowed us to establish it as a sensible optimization target, revealing a scaling law by which its quality improves predictably with the size of its feature space. This contrasts with existing kernel-based analyses, in which the quality of the identified kernel as a similarity measure on natural data has remained unassessed.

The empirical CGK estimate further enabled three additional findings: 1) the logarithmic target of InfoNCE was empirically shown to exhibit substantial indefiniteness on natural text, thus creating a fundamental representational bottleneck; 2) both contrastive losses recover their predicted embedding geometries when training provides sufficient information about the target, whereas standard contrastive training primarily recovers local neighborhoods rather than the full similarity structure; and 3) embeddings optimized to preserve neighborhood rankings under limited dimensionality align more closely with the logarithmic target than with the CGK itself.

Our downstream experiments further show that, with a decoupled parameterization of embedding norms and directions, the spectral loss becomes a competitive alternative to InfoNCE, with favorable clustering and classification performance.

Ultimately, this work constitutes an initial exploration of generative contrastive learning, but we believe that the explicitly defined learning setup offers substantial opportunities for future research, as it uniquely allows theoretical analyses to be directly paired with their empirical validation.

AI USE STATEMENT

Generative AI was used as an integrated part of our everyday work, and we did not strictly exclude it from any stage of this work. Beyond obvious uses such as writing code and rephrasing text, we also used AI to double-check our own reasoning and results. As a result, we cannot reliably rule out that AI played some role in any given subtask, even if its influence might have been minor.

In this work, we therefore used generative AI tools for these of the required disclosure tasks: Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results.

Additionally, we used generative AI tools for brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We described experimental details extensively in the appendix. The camera-ready version of this paper will also contain a link to the source code used in the experiments.

REFERENCES

Marc Brinner and Sina Zarrieß. Semcse-multi: Multifaceted and decodable embeddings for aspect-specific and interpretable scientific domain mapping, 2026. URL https://arxiv.org/abs/2510.11599.

Marc Felix Brinner and Sina Zarrieß. SemCSE: Semantic contrastive sentence embeddings using LLM-generated summaries for scientific abstracts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32704–32719, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1662. URL https://aclanthology.org/2025.emnlp-main.1662/.

Jeff Z. Hao Chen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.207. URL https://aclanthology.org/2020.acl-main.207/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux,

Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy

Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jeff Z. HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=AuEgNlEAmed.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Tommi Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In M. Kearns, S. Solla, and D. Cohn (eds.), *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998. URL https://proceedings.neurips.cc/paper_files/paper/1998/file/db1915052d15f7815c8b88e879465a1e-Paper.pdf.

Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *J. Mach. Learn. Res.*, 5:819–844, December 2004. ISSN 1532-4435.

Yoo Hyun Jeong, Myeongsoo Han, and Dong-Kyu Chae. A simple angle-based approach for contrastive learning of unsupervised sentence representation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5553–5572, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.318. URL https://aclanthology.org/2024.findings-emnlp.318/.

Daniel D. Johnson, Ayoub El Hanchi, and Chris J. Maddison. Contrastive learning can find an optimal basis for approximately invariant functions. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. URL https://openreview.net/forum?id=ZzngjJb7mLt.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Proceedings of the 34th International Conference on*

Neural Information Processing Systems, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009. URL http://jmlr.org/papers/v10/kanamori09a.html.

Eunhui Kim, Yuna Jeong, and Myung-Seok Choi. Medibiodeberta: Biomedical language model with continuous learning and intermediate fine-tuning. *IEEE Access*, 11:141036–141044, 2023. doi: 10.1109/ACCESS.2023.3341612.

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, 2025a. URL https://arxiv.org/abs/2405.17428.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. Gemini embedding: Generalizable embeddings from gemini, 2025b. URL https://arxiv.org/abs/2503.07891.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026. URL https://arxiv.org/abs/2601.04720.

Tian Yu Liu, Matthew Trager, Alessandro Achille, Pramuditha Perera, Luca Zancato, and Stefano Soatto. Meaning representations from trajectories in autoregressive models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=UyGWafcopT.

Yiwei Lu, Guojun Zhang, Sun Sun, Hongyu Guo, and Yaoliang Yu. f-MICL: Understanding and generalizing infoNCE-based contrastive learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=ZD03VUZmRx.

Raja Marjieh, Sreejan Kumar, Declan Iain Campbell, Liyi Zhang, Gianluca Bencomo, Jake Snell, and Thomas L. Griffiths. Using contrastive learning with generative similarity to learn spaces that capture human inductive biases, 2024. URL https://openreview.net/forum?id=OOt5RMI0JC.

George V. Moustakides and Kalliopi Basioti. Training neural networks for likelihood/density ratio estimation, 2019. URL https://arxiv.org/abs/1911.00405.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theor.*, 56(11):5847–5861, November 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2068870. URL https://doi.org/10.1109/TIT.2010.2068870.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CR1XOQ0UTh-.

Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. InfoNCE: Identifying the gap between theory and practice. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=RNQzrXWhIs.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19250–19286. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/saunshi22a.html.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. SciRepEval: A multi-format benchmark for scientific document representations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5548–5566, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.338. URL https://aclanthology.org/2023.emnlp-main.338/.

Aivin V. Solatorio. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning, 2024. URL https://arxiv.org/abs/2402.16829.

Taiji Suzuki, Masashi Sugiyama, Takafumi Kanamori, and Jun Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10 Suppl 1(Suppl 1):S52, January 2009.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: synthetic images from text-to-image models make strong visual representation learners. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning Vision from Models Rivals Learning Vision from Data . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15887–15898, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.01504. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.01504.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Toshimitsu Uesaka, Taiji Suzuki, Yuhta Takida, Chieh-Hsin Lai, Naoki Murata, and Yuki Mitsufuji. Weighted point set embedding for multimodal contrastive learning toward optimal similarity metric. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.),

International Conference on Learning Representations, volume 2025, pp. 32997–33018, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/514b48017505dc778c6db2d845cbbe24-Paper-Conference.pdf.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2495–2504, 2021. doi: 10.1109/CVPR46437.2021.00252.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024. URL https://arxiv.org/abs/2212.03533.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 641–649, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. Debiased contrastive learning of unsupervised sentence representations. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6120–6130, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.423. URL https://aclanthology.org/2022.acl-long.423/.

A Proofs

A.1 Proof of Theorem 1

This section derives the optimal state of the InfoNCE loss in the limit of infinite negative samples within the SemCSE training framework. Our proof initially follows the proof by Wang & Isola (2020) in analyzing the loss in the limit of infinite negative samples, but then characterizes the optimal state in the context of the specifically defined data-generating distribution, which leads to the CGK arising as central component within the loss.

Deriving Asymptotic Loss. We begin by establishing the form of the log-normalized SemCSE objective: subtracting the constant $\log M$ and separating the positive and negative similarity terms yields:

$$\mathcal{L}_M - \log M \;=\; \mathbb{E}_{\substack{a\sim\mu \\ x,x'\sim p(\cdot|a)}}\left[-\frac{s_{xx'}}{\tau}\right] \;+\; \mathbb{E}\left[\log\left(\frac{1}{M}\,e^{s_{xx'}/\tau} + \frac{1}{M}\sum_{i=1}^{M} e^{s_{xx_i^-}/\tau}\right)\right], \qquad (9)$$

where the outer expectation in the second term is taken over the anchor a, the positive continuations x, x', and the negatives $x_1^-, \ldots, x_M^- \overset{\text{iid}}{\sim} p(\cdot)$.

By the strong law of large numbers, the average over the negative samples converges almost surely to its expectation under the marginal distribution:

$$\frac{1}{M}\sum_{i=1}^{M} e^{s_{xx_i^-}/\tau} \xrightarrow{\text{a.s.}} \mathbb{E}_{x^-\sim p(\cdot)}\left[e^{s_{xx^-}/\tau}\right] \;\coloneqq\; Z(x), \tag{10}$$

while the isolated positive term $\frac{1}{M}e^{s_{xx'}/\tau}$ vanishes as $M \to \infty$.

Therefore we have

$$\begin{aligned}
\mathcal{L}_\infty \;&\coloneqq\; \lim_{M\to\infty}\left(\mathcal{L}_M - \log M\right) \\
&=\; \mathop{\mathbb{E}}_{\substack{a\sim\mu \\ x,x'\sim p(\cdot|a)}}\left[-\frac{s_{xx'}}{\tau}\right] \;+\; \mathbb{E}\left[\lim_{M\to\infty}\log\left(\frac{1}{M}\,e^{s_{xx'}/\tau} + \frac{1}{M}\sum_{i=1}^{M} e^{s_{xx_i^-}/\tau}\right)\right] \\
&=\; \mathbb{E}_{(x,x')\sim p(x,x')}\left[-\frac{s_{xx'}}{\tau}\right] \;+\; \mathbb{E}_{x\sim p(\cdot)}[\log Z(x)]\,.
\end{aligned} \tag{11}$$

Moving the limit inside the expectation is justified by the bounded convergence theorem: since the text space $\mathcal{X}_C$ is finite, the similarities $S = [s_{xx'}]$ form a finite matrix, ensuring the argument of the logarithm is strictly bounded within $[e^{\min s/\tau},\, 2\,e^{\max s/\tau}]$ for all $M \geq 1$. The resulting asymptotic loss therefore has two interpretable components: the first term encourages *alignment* of positive pairs under the joint co-generation distribution $p(x, x') = \mathbb{E}_{a\sim\mu}[p(x \mid a)\,p(x' \mid a)]$, while the second enforces *uniformity* over the marginal distribution $p(x)$.

Deriving the Gradient. By rewriting the expectations in $\mathcal{L}_\infty$ as discrete sums over the finite space $\mathcal{X}_C$, we can compute the partial derivative of the loss with respect to an arbitrary similarity score s_{uv}.

Since the alignment term is linear in s_{uv}, its derivative is simply

$$\frac{\partial}{\partial s_{uv}}\sum_{x,x'} p(x,x')\left(-\frac{s_{xx'}}{\tau}\right) \;=\; -\frac{1}{\tau}\,p(u,v). \tag{12}$$

For the uniformity term, expanding $Z(x) = \sum_{x^-} p(x^-)\,e^{s_{xx^-}/\tau}$ reveals that the variable s_{uv} only appears within the partition function $Z(u)$, specifically when $x^- = v$. Applying the chain rule yields:

$$\frac{\partial}{\partial s_{uv}}\sum_{x} p(x)\log Z(x) \;=\; p(u)\,\frac{1}{Z(u)}\,\frac{\partial Z(u)}{\partial s_{uv}} \;=\; \frac{1}{\tau}\,\frac{p(u)\,p(v)\,e^{s_{uv}/\tau}}{Z(u)}. \tag{13}$$

Summing these two components provides the complete gradient of the asymptotic loss:

$$\frac{\partial \mathcal{L}_\infty}{\partial s_{uv}} \;=\; -\frac{1}{\tau}\,p(u,v) \;+\; \frac{1}{\tau}\,\frac{p(u)\,p(v)\,e^{s_{uv}/\tau}}{Z(u)}. \tag{14}$$

Determining Unconstrained Minimizers. The alignment term is linear in S, while each $\log Z(x)$ is a log-sum-exp of affine functions of S and hence convex. Thus $\mathcal{L}_\infty$ is differentiable and convex over unconstrained scores, so its global minimizers are exactly its stationary points.

Setting Eq. equation 14 to zero and dividing by $p(u)p(v) > 0$ gives

$$e^{s^*_{uv}/\tau} \;=\; Z(u)\,\frac{p(u,v)}{p(u)\,p(v)} \;=\; Z(u)\,k(u,v), \quad \text{i.e.} \quad s^*_{uv} \;=\; \tau\log k(u,v) \;+\; \tau\log Z(u) \tag{15}$$

where $k(u,v) = p(u,v)/(p(u)p(v))$ is the conditional generation kernel of Eq. equation 3. Note that, in this form, $Z(u)$ depends on the row $(s_{uv})_{v\in\mathcal{X}_C}$, so that this equation determines the scores only implicitly.

To make it explicit, we fix arbitrary row offsets b_u, and thus let $s_{uv} = \tau \log k(u,v) + b_u$. Computing $Z(u)$ from this choice of s_{uv} yields:

$$Z(u) = \sum_v p(v) e^{s_{uv}/\tau} = \sum_v p(v) e^{\log k(u,v) + b_u/\tau}$$
$$= e^{b_u/\tau} \sum_v p(v) k(u,v) = e^{b_u/\tau} \sum_v \frac{p(u,v)}{p(u)} = e^{b_u/\tau}. \tag{16}$$

Thus $\tau \log Z(u) = b_u$, confirming that Eq. equation 15 holds for every choice of row offsets b_u.

Characterizing Symmetric Minimizers. Because embedding similarities like the dot product and cosine similarity are inherently symmetric, our optimization is naturally restricted to scores satisfying $s_{uv} = s_{vu}$. We now determine which members of the unconstrained minimizer set $\mathcal{S}^*$ satisfy this symmetry.

First, since $p(u,v) = p(v,u)$, the kernel k is also symmetric. As a result, choosing a constant row offset $b_u = C$ for all u yields a symmetric unconstrained minimizer of $\mathcal{L}_\infty$, confirming that imposing symmetry does not increase the minimum loss, and that the symmetric minimizers are exactly the symmetric members of the unconstrained family $\mathcal{S}^*$.

Applying the symmetry condition to an arbitrary member of $\mathcal{S}^*$ yields:

$$s_{uv} = s_{vu} \quad \Longleftrightarrow \quad \tau \log k(u,v) + b_u = \tau \log k(v,u) + b_v \quad \Longleftrightarrow \quad b_u = b_v. \tag{17}$$

All row offsets must therefore equal a common constant C, so the minimizers over symmetric scores are exactly

$$s^*_{uv} = \tau \log k(u,v) + C, \qquad C \in \mathbb{R}, \tag{18}$$

which proves Theorem 1. □

A.2 Derivation of the Spectral Loss Optimum

We demonstrate that the spectral loss is uniquely minimized by similarity scores equal to the conditional generation kernel.

Rewriting the Loss. Recall the spectral objective from Eq. equation 7:

$$\mathcal{L}_{\text{spectral}} = -\mathbb{E}_{\substack{a\sim\mu \\ x,x'\sim p(\cdot|a)}} [s_{xx'}] + \frac{1}{2}\mathbb{E}_{x,x' \overset{\text{iid}}{\sim} p(\cdot)} [s^2_{xx'}]. \tag{19}$$

Because the continuation space $\mathcal{X}_C$ is finite, these expectations can be rewritten as discrete sums. Positive pairs are distributed according to the joint co-generation distribution $p(x,x')$, whereas independent pairs follow the product of marginals $p(x)p(x')$, yielding:

$$\mathcal{L}_{\text{spectral}} = -\sum_{x,x'} p(x,x') s_{xx'} + \frac{1}{2} \sum_{x,x'} p(x)p(x') s^2_{xx'}. \tag{20}$$

Given the standard full-support softmax generation model, $p(x \mid a) > 0$, which guarantees that $p(x) > 0$ for all continuations x.

Deriving the Gradient. Treating the elements of the matrix $S = [s_{xx'}]$ as independent variables, in each sum, the only summand dependent on a specific entry s_{uv} is the one where $(x,x') = (u,v)$. Computing the partial derivative of both the linear and quadratic components gives:

$$\frac{\partial \mathcal{L}_{\text{spectral}}}{\partial s_{uv}} = -p(u,v) + p(u)p(v) s_{uv}. \tag{21}$$

Setting this gradient to zero and dividing by the strictly positive term $p(u)p(v)$ yields:

$$s^*_{uv} = \frac{p(u,v)}{p(u)p(v)} = k(u,v), \tag{22}$$

where the final step directly applies the definition of the kernel from Eq. equation 3.

Establishing Global Optimality. The objective is a sum of quadratic functions, where each one depends on a single score s_{uv} and has a strictly positive quadratic coefficient $\frac{1}{2}p(u)p(v)$. The loss is therefore strictly convex with respect to S, making the stationary point $S^* = K$ its unique global minimizer.

Furthermore, since K is symmetric, it also satisfies $s_{uv} = s_{vu}$ and remains the unique minimizer when symmetry in the optimal state is required. For this reason, the spectral loss is uniquely minimized by

$$s^*_{uv} = k(u, v). \tag{23}$$

□

A.3 Equivalence of the Spectral Loss and a Weighted Squared Error

We prove that the spectral loss is equal to a weighted squared error between the similarity scores $s_{xx'}$ and the CGK, up to a constant independent of the scores.

Proposition A.2. *For any similarity function $s\colon \mathcal{X}_C \times \mathcal{X}_C \to \mathbb{R}$,*

$$\frac{1}{2}\mathbb{E}_{x,x' \overset{\text{iid}}{\sim} p(\cdot)}\left[(s_{xx'} - k(x, x'))^2\right] = \mathcal{L}_{\text{spectral}} + C, \tag{24}$$
$$C \coloneqq \frac{1}{2}\mathbb{E}_{x,x' \overset{\text{iid}}{\sim} p(\cdot)}\left[k(x, x')^2\right]$$

where the constant C depends only on the data-generating distribution.

Proof. Expanding the square and using $p(x)p(x')k(x, x') = p(x, x')$ gives

$$\begin{aligned}
&\frac{1}{2}\sum_{x,x'} p(x)p(x')\big(s_{xx'} - k(x, x')\big)^2 \\
&\quad = \frac{1}{2}\sum_{x,x'} p(x)p(x')s^2_{xx'} - \sum_{x,x'} p(x, x')s_{xx'} + C \\
&\quad = \mathcal{L}_{\text{spectral}} + C,
\end{aligned} \tag{25}$$

where the last equality follows from Eq. equation 20. □

A.4 Sampling Restricted to Terminated Continuations

The CGK is defined under the length-prior model of Eq. equation 1. In contrast, the SemCSE dataset generation procedure generates continuations until a sentence is complete. We show that this procedure leads to the same kernel on the sampled continuations, assuming that, for every anchor, sentences are completed within L_C tokens almost surely.

Lemma A.1. *Let $\mathcal{S} \subseteq \mathcal{X}_C$ satisfy $\sum_{x\in\mathcal{S}} p(x \mid a) = c > 0$ for every anchor a, with c independent of a. Then replacing $p(\cdot \mid a)$ by its conditional distribution on $\mathcal{S}$ preserves the CGK on $\mathcal{S} \times \mathcal{S}$.*

Proof. For $x, x' \in \mathcal{S}$, conditioning gives $p_{\mathcal{S}}(x \mid a) = p(x \mid a)/c$. With unchanged anchor distribution μ, the resulting marginal and joint probabilities therefore are

$$p_{\mathcal{S}}(x) = \frac{p(x)}{c}, \qquad p_{\mathcal{S}}(x, x') = \frac{p(x, x')}{c^2}. \tag{26}$$

These factors cancel in the kernel ratio:

$$k_{\mathcal{S}}(x, x') = \frac{p_{\mathcal{S}}(x, x')}{p_{\mathcal{S}}(x)p_{\mathcal{S}}(x')} = \frac{p(x, x')}{p(x)p(x')} = k(x, x'). \tag{27}$$

□

Application to Sentence Completion. We define sentence completion as a deterministic stopping rule based on the generated tokens, where generation is terminated at the first sentence boundary reached during the autoregressive sampling procedure. Let $\mathcal{S}$ be the set of sequences at which this rule stops generation. No member of $\mathcal{S}$ is a proper prefix of another, so their generation events are disjoint.

Writing $\tilde{p}(x \mid a) = \prod_{t=1}^{|x|} p(x_t \mid a, x_{<t})$, the assumption of completion within L_C tokens gives

$$\sum_{x \in \mathcal{S}} \tilde{p}(x \mid a) = 1. \tag{28}$$

Since $p(x \mid a) = \tilde{p}(x \mid a)/L_C$, it follows that

$$\sum_{x \in \mathcal{S}} p(x \mid a) = \frac{1}{L_C} \qquad \text{for every } a. \tag{29}$$

Lemma A.1 therefore applies with $c = 1/L_C$, and $p_{\mathcal{S}}(x \mid a) = \tilde{p}(x \mid a)$ is exactly the distribution produced by the stopping procedure. Consequently, the sampled distribution constitutes a restriction of the original CGK to $\mathcal{S} \times \mathcal{S}$, so that Theorem 1 and the spectral optimum apply on this subset with the same kernel k.

A.5 Proof of Proposition 1

We define the repulsion force exerted on a negative within a contrastive loss as the gradient magnitude $g(s_j) = |\partial \mathcal{L}/\partial s_j|$. Since g generally depends on batch-specific normalization terms, we analyze the relative repulsion ratio $R(s_H, s_E)$ of a negative s_H with high similarity to the anchor to an easy negative s_E with low similarity:

$$R(s_H, s_E) = \frac{g(s_H)}{g(s_E)},$$

which therefore quantifies how much more strongly the loss pushes away a hard negative compared to an easy one.

Proposition 1 (Relative Repulsion Ratios)**.** *Let $0 < s_E < s_H$ be the cosine similarities of an easy and a hard negative sample to the same anchor. Under the InfoNCE loss with temperature τ, the relative repulsion ratio is*

$$R_{\textit{InfoNCE}}(s_H, s_E) = e^{(s_H - s_E)/\tau}.$$

Under the spectral loss, writing the dot product as $\sigma_a \sigma_j s_j$ with embedding norms σ, the relative repulsion ratio is

$$R_{\textit{spectral}}(s_H, s_E) = \left(\frac{\sigma_H}{\sigma_E}\right)^2 \frac{s_H}{s_E}.$$

Proof. For an anchor a, let s_j denote the cosine similarity to the j-th negative sample. We define the repulsion force exerted on this negative as $g(s_j) = |\partial \mathcal{L}/\partial s_j|$ and the relative repulsion ratio as $R(s_H, s_E) = g(s_H)/g(s_E)$.

InfoNCE. With positive similarity s_+ and partition function $Z = e^{s_+/\tau} + \sum_k e^{s_k/\tau}$, the loss is $\mathcal{L}_{\text{InfoNCE}} = -s_+/\tau + \log Z$. Differentiating with respect to s_j yields

$$g_{\text{InfoNCE}}(s_j) = \frac{1}{Z}\frac{\partial Z}{\partial s_j} = \frac{1}{\tau Z} e^{s_j/\tau}. \tag{30}$$

Since Z is shared by all negatives of the same anchor, the factor $1/(\tau Z)$ cancels in the ratio, which gives

$$R_{\text{InfoNCE}}(s_H, s_E) = e^{(s_H - s_E)/\tau}. \tag{31}$$

Spectral Loss. The spectral loss operates on dot products, each of which decomposes as $f(a)^\top f(x_j) = \sigma_a \sigma_j s_j$ with embedding norms $\sigma_a = \|f(a)\|$ and $\sigma_j = \|f(x_j)\|$. In the decoupled variant, these norms are predicted by a separate magnitude head. The loss is thus $\mathcal{L}_{\text{spectral}} = -\sigma_a \sigma_+ s_+ + \frac{1}{2}\sum_k (\sigma_a \sigma_k s_k)^2$. Differentiating with respect to s_j at fixed norms yields

$$g_{\text{spectral}}(s_j) = (\sigma_a \sigma_j)^2 \, s_j \tag{32}$$

for $s_j > 0$. The factor σ_a^2 cancels in the ratio, which gives

$$R_{\text{spectral}}(s_H, s_E) = \left(\frac{\sigma_H}{\sigma_E}\right)^2 \frac{s_H}{s_E}. \tag{33}$$

□

B Experimental Details

B.1 CGK Evaluation

We empirically evaluate the suitability of the CGK within a standard retrieval setting. Specifically, we compute the explicit feature representation ϕ underlying the CGK for a set of query-target text pairs, and assess how effectively their dot product (which recovers the kernel k) can retrieve the correct target from a pool of distractors.

B.1.1 Experimental Setup

Data Generation The evaluation is conducted within the scientific domain using sentences that summarize scientific abstracts. This aligns closely with the training setup of the SemCSE model (Brinner & Zarrieß, 2025), which is the main target of our subsequent analyses. We also use the Llama-3 8B model (Grattafiori et al., 2024) to compute the conditional probabilities $p(x \mid a)$, since this model was used for creating the SemCSE training dataset.

The query-target pairs for the retrieval evaluation are constructed as follows: first, we sample $2{,}000$ random scientific abstracts from the corpora underlying the SciRepEval benchmark (Singh et al., 2023). For each abstract, we then generate two summarizing sentences by appending the prompt *"A comprehensive summary for our work would be that"* and making the LLM autoregressively complete the sentence at a temperature of 1.0 (again mirroring the data generation pipeline of SemCSE). Because the two summaries for a given abstract share the same generation context, they are highly semantically related and serve as the query and target for our retrieval evaluation.

Data Filtering To ensure sufficient quality of these query-target pairs, we perform an additional filtering step. This step is needed since continuations can sometimes be highly uninformative (e.g., *"A comprehensive summary for our work would be that our proposed method works very well."*), which does not allow for accurate matching between query and target, irrespective of the quality of the underlying method.

We use the following method for filtering: using the $2{,}000$ source abstracts from which the data was generated, we compute the conditional probability $p(s \mid a)$ for each summary against all $2{,}000$ abstracts. A pair is kept in the dataset only if both of its summaries assign the highest generation probability to their true source anchor. This criterion thus removes generic phrases that match multiple texts equally well. Applying this filter excluded 54 pairs, leaving a final set of $1{,}946$ pairs for evaluation.

Feature Vector Construction We then compute the feature vectors $\phi(s)$ for each summary s in this dataset. As the CGK requires a set of anchor documents $\mathcal{A}$, we sample an additional set of 28K scientific abstracts from the SciRepEval corpora. With these anchors, the feature vector ϕ is constructed from the conditional generation probabilities $p(s \mid a)$, which represent the likelihood of generating summary s given anchor a (and the additional prompt phrase) according to the LLM (Llama-3 8B). Specifically, the feature vectors ϕ are computed as defined in Equation 4, assuming a uniform anchor prior $\mu(a_j) = 1/m$.

Using the full set of anchors $\mathcal{A}$ would yield 28K-dimensional feature vectors ϕ. However, we 1) aim to analyze how the number of anchors impacts retrieval performance, and 2) want to ensure robust estimates independent of any specific anchor selection. To achieve this, we construct lower-dimensional representations by subsampling d anchors from the full candidate pool, allowing us to assemble d-dimensional feature vectors $\phi_d(s)$ for each summary s. The evaluation is then performed across a range of anchor counts ($d \in \{10, 50, 100, 250, 500, 1000, 2000, 5000, 10000, 25000\}$), and we average the results for each d over 50 independent random draws to create more precise estimates.

Retrieval Task Definition Retrieval is performed using the dot product between the resulting vectors, which is mathematically equivalent to computing the CGK over the selected set of anchor documents. We evaluate retrieval in both directions (i.e., each summary in a pair serves once as the query and once as the target), with the task being the retrieval of the correct counterpart from the pool of $1{,}946$ candidate summaries. Finally, we report the mean and median rank at which the correct target is retrieved.

B.1.2 Results and Analysis

Effect of Anchor Count Figure 1 (a) shows the average retrieval rank as a function of the anchor count, falling from 553.60 ± 83.74 at 10 anchors to 18.44 ± 0.37 at $25{,}000$. On a log–log scale (Figure 1 (b)) we see that this curve is well described by a power law rank $\propto A^{-0.53}$, where the linear function is fitted on $A \geq 250$ since below that range the power law cannot hold due to approaching the ceiling of random-guessing (which results in an expected rank of 973). Importantly, the curve has not begun to flatten at the largest anchor count we evaluate, so that additional improvements are expected with increasing anchor count, thus approaching near perfect retrieval performance.

Mean vs. Median Performance Comparison In contrast to the mean retrieval rank of 18.44 at $|\mathcal{A}| = 25{,}000$, the median retrieval rank is just 2.0 ± 0.0. This indicates that for the vast majority of queries, the target is already retrieved at or near the top rank, which therefore demonstrates good retrieval performance by the CGK in these cases. The strong discrepancy between mean and median arises from a small subset of cases that were matched very badly and thus strongly degrade the average rank. This difficulty is likely caused by an absence of closely related anchors for those specific summaries within the sampled set. If a given summary s is an unlikely continuation for all available anchors, its feature vector $\phi(s)$ is less sharply peaked, and its highest-weighted anchors are likely caused not by genuine topical overlap, but by noisy correlations tied to specific phrasing choices. However, these factors naturally vary between different summaries of the same text and are therefore not a consistent signal for the retrieval task.

This hypothesis is supported by an empirical evaluation of the entropy of $\phi(s)$ for pairs with high versus low retrieval performance: the top 10% of samples with the best retrieval ranks exhibit an average entropy of 1.86, whereas the bottom 10% average 3.36. This demonstrates that successful retrieval relies on a small selection of highly informative anchors that provide the necessary discriminative signal in ϕ to correctly match the query to the target. As the total anchor count increases, the likelihood of including such informative anchors in $\mathcal{A}$ naturally rises, thus explaining the continuous improvement in retrieval performance.

Importance of Informative Anchors In the context of the previous paragraph, it becomes clear that selecting the anchor set favorably for a specific task is highly beneficial for retrieval performance. Only if informative anchors have a reasonable probability of being sampled can a limited number of anchors facilitate a sufficient discriminatory signal in the resulting feature vectors ϕ.

Our experiment evaluated the CGK in exactly such a favorable setup: the anchors (i.e., scientific abstracts with a summary-initiating prompt) were specifically matched to the type of target text (i.e., scientific summaries), thus creating a reasonable chance of encountering high conditional probabilities $p(x \mid a)$. If the CGK was instead applied as a general-purpose text embedding approach across diverse domains and text types, the proportion of informative anchors for any given text would drop substantially. However, if the total anchor count is increased, the absolute number of informative anchors still grows, meaning that strong retrieval performance would ultimately still be expected. While relying on such massive anchor sets is computationally infeasible for direct kernel computation in practice, this argument still supports the conclusion that the CGK itself remains highly

semantically informative and therefore constitutes a principled optimization target for more efficient neural embedding models.

B.2 Assessing Linear vs. Logarithmic Kernel Alignment

We introduce a diagnostic metric to assess whether the similarities predicted by an embedding model align more closely with a linear target matrix or its element-wise logarithm. In practice, we use this metric with the empirical Gram matrix $\widehat{K}$ computed in Section 2.2 / Appendix B.1.1 as the reference kernel, and compare trained embeddings to this target.

Requirements To ensure a correct evaluation, the metric must satisfy three main criteria:

1. **Sensitivity to spacing:** Since $\widehat{K}$ and $\log \widehat{K}$ share the exact same similarity ranking, the metric cannot rely solely on comparing the predicted rankings against the targets. Instead, it must accurately evaluate the relative spacing of the similarity values.
2. **Scale and shift invariance:** The metric must be invariant to global positive scaling and additive shifts. This ensures that arbitrary rescaling or constant offsets (such as C in Theorem 1) do not change the assessment of the underlying geometry.
3. **Robustness to reference noise:** Since the empirical kernel $\widehat{K}$ is estimated from a finite set of anchors, it naturally contains estimation noise. The metric must therefore be able to separate the true underlying structure from this random noise.

Target Matrices Let C_{ij} denote the similarity score between documents i and j predicted by an embedding model. Depending on the evaluation, this can be the raw dot product ($C_{ij} = v_i^\top v_j$) or the cosine similarity. We aim at evaluating whether the similarity structure of this predicted matrix C aligns more with a reference matrix $\widehat{K}$ or its element-wise logarithm $\log \widehat{K}$. For this reason, we use these exact matrices as the two references to compare against:

$$T_{ij}^{(1)} = \widehat{K}_{ij} \qquad \text{and} \qquad T_{ij}^{(0)} = \log \widehat{K}_{ij}. \tag{34}$$

Metric Definition We aim to measure the alignment of our predicted similarities C with both potential targets, $T^{(0)}$ and $T^{(1)}$, by quantifying how much of the variance in C can be explained by each target. To ensure the evaluation is invariant to scale and shift, we evaluate this alignment after applying the optimal positive-affine transformation of the target, $aT^{(p)} + b$ (with $a \geq 0$), meaning that we choose a and b to minimize the difference between C and $aT^{(p)} + b$.

Statistically, this fraction of variance explained by an unconstrained affine fit (which would allow for $a < 0$) is exactly equivalent to the squared Pearson correlation coefficient, computed as:

$$R^2 = \frac{\mathrm{Cov}_\Omega(C, T^{(p)})^2}{\mathrm{Var}_\Omega(C)\,\mathrm{Var}_\Omega(T^{(p)})}, \tag{35}$$

where Cov_Ω and Var_Ω denote the usual covariance and variance computed over all off-diagonal matrix entries $\Omega = \{(i, j) : i \neq j\}$.

This means that we can use the Pearson correlation, as long as we can enforce $a \geq 0$, and thus guarantee that a negatively correlated score matrix is not incorrectly classified as aligned. We do this by simply allowing solely for positive covariance values, resulting in the positive-affine alignment score $S(p)$ defined as:

$$S(p) = \frac{[\mathrm{Cov}_\Omega(C, T^{(p)})]_+^2}{\mathrm{Var}_\Omega(C)\,\mathrm{Var}_\Omega(T^{(p)})}. \tag{36}$$

This score thus estimates the structural alignment between the model and the target, independent of arbitrary scoring scales. Assuming non-zero variances, $S(p)$ is bounded within $[0, 1]$, where a value of 1 indicates a perfect positive-affine fit. We convert this alignment metric into an error for a given

target $T^{(p)}, p \in \{0, 1\}$ by simply defining $A(p) = 1 - S(p)$, which is exactly the percentage of unexplained variance in C by the corresponding target $T^{(p)}$.

Noise-Corrected Estimation The standard score in Equation 36 assumes the target matrix $T^{(p)}$ is exact. However, empirical estimates of $\widehat{K}$ (as used in our later experiments) are noisy due to limited anchor count, which artificially increases the error $A(p)$. For this reason, we correct for this by computing two independent estimates of the target, $T^{(p,1)}$ and $T^{(p,2)}$, using disjoint subsets of anchors, and extract the consistent signal of alignment between our predicted similarities and either noisy target.

In detail this is done by modeling each noisy estimate as $T^{(p,r)} = \mu^{(p)} + \epsilon^{(p,r)}$, where $\mu^{(p)}$ represents the true underlying target structure and $\epsilon^{(p,r)}$ is independent zero-mean estimation noise. Under this assumption computing the standard variance equates to $\mathrm{Var}_\Omega(T^{(p,r)}) = \mathrm{Var}_\Omega(\mu^{(p)}) + \mathrm{Var}_\Omega(\epsilon^{(p,r)})$, which (due to the additional term $\mathrm{Var}_\Omega(\epsilon^{(p,r)})$) incorrectly penalizes the model for failing to predict the random reference noise.

However, because the estimation noise is independent across replicas, we have in expectation $\mathbb{E}[\mathrm{Cov}_\Omega(T^{(p,1)}, T^{(p,2)})] = \mathrm{Var}_\Omega(\mu^{(p)})$. Further, since C is uncorrelated with $\epsilon^{(p,r)}$, we similarly have in expectation $\mathbb{E}[\mathrm{Cov}_\Omega(C, T^{(p,r)})] = \mathrm{Cov}_\Omega(C, \mu^{(p)})$. Substituting these estimates into our original formulation then yields the corrected alignment score:

$$S_{\mathrm{corr}}(p) = \frac{[\mathrm{Cov}_\Omega(C, T^{(p,1)})]_+ [\mathrm{Cov}_\Omega(C, T^{(p,2)})]_+}{\mathrm{Var}_\Omega(C)\, \mathrm{Cov}_\Omega(T^{(p,1)}, T^{(p,2)})}. \tag{37}$$

The noise-corrected error is then defined analogously as $A_{\mathrm{corr}}(p) = 1 - S_{\mathrm{corr}}(p)$.

Final Contrast To summarize the relative preference between the two candidate target geometries, we compute the difference in their errors (substituting A_{corr} when noise correction is applied):

$$D = A(1) - A(0). \tag{38}$$

A negative contrast ($D < 0$) indicates that the model's similarity structure aligns more closely with the linear kernel, whereas a positive contrast ($D > 0$) indicates stronger alignment with the logarithmic target. A value near zero implies no clear preference. Ultimately, this metric directly answers which underlying geometry best describes the learned representation, explicitly disregarding absolute, unnormalized reconstruction errors.

Empirical Validation To validate that the proposed metric correctly distinguishes linear from logarithmic kernel geometries, we use the data from Section 2.2 / Appendix B.1.1, which constructed CGK feature vectors ϕ for $1{,}946$ pairs of summaries with respect to an anchor set consisting of 28K anchors. For this evaluation, we use only the first summary of each pair and randomly partition the anchor pool into three disjoint groups of $9{,}333$ anchors, computing an independent empirical kernel $\widehat{K}_i \in \mathbb{R}^{1946 \times 1946}$ from each. One of these kernel estimates is then treated as prediction C, while the other two are used as two independent estimates $\widehat{K}_1$ and $\widehat{K}_2$ of the true underlying kernel K. Note that we use two estimates since we apply noise correction.

We then create two versions of the estimate C: the linear estimate $C_{\mathrm{lin}} = C$ and the logarithmic estimate $C_{\mathrm{log}} = \log C$. Our evaluation now tests whether the metric correctly identifies the linear geometry for C_{lin} and the logarithmic geometry for C_{log}.

Evaluating the noise-corrected contrast $D = A_{\mathrm{corr}}(1) - A_{\mathrm{corr}}(0)$ across multiple random partitions, we observe a mean value of $D = -0.1054$ for C_{lin} and $D = +0.6665$ for C_{log}, meaning that the target geometry is, on average, estimated correctly. More importantly, all evaluations return the expected sign for both the linear and the logarithmic test matrices, thus showing that the metric reliably determines the underlying target structure despite the estimation noise inherent in the independently sampled kernels.

Robustness to Noise As a final check, we assess the metric's robustness to increasing noise in the evaluated matrix C. We artificially degrade the test matrices C_{lin} and $C_{\log}$ by progressively reducing their underlying anchor counts, while keeping the two reference estimates fixed at $9{,}333$ anchors each. At $2{,}000$ test anchors, the mean contrasts remain correctly separated at -0.0701 and $+0.5992$, and even when reduced to just 500 anchors, the metric successfully identifies the correct geometry in all evaluations, yielding correctly signed mean scores of -0.0164 and $+0.4982$. The metric therefore maintains strong discriminative power under substantial estimation noise, which confirms its suitability for evaluating noisy model predictions.

B.3 Free-Vector Kernel Recovery

To empirically validate the theoretically derived differences between the InfoNCE and spectral objectives, we evaluate the geometries of the embedding spaces they produce in an idealized setup designed to eliminate potential generalization gaps. We achieve this by evaluating free-vector fits, thereby removing any shared parameters (e.g., in the form of a shared encoder) that could influence the optimization dynamics. Furthermore, we both fit and evaluate the free vectors against the exact same empirical CGK Gram matrix $\widehat{K}$, which eliminates the need for generalization to unseen data. In combination, this setup therefore allows both objectives to fully realize their intended target geometries, which enables a precise and unconfounded comparison.

General Setup We again utilize the data from Section 2.2 / Appendix B.1.1 to construct the empirical Gram matrix $\widehat{K} \in \mathbb{R}^{1946\times 1946}$ by computing the feature vectors ϕ across the full set of 28K anchors as described in Appendix B.2. This matrix $\widehat{K}$ serves both as the optimization target and as reference for the evaluation. For a chosen embedding dimension d, we then assign a randomly initialized free vector $v_i \in \mathbb{R}^d$ to each of the $1{,}946$ documents, which are then optimized by one of the training objectives listed below. Once training is done, we evaluate the resulting embedding geometry by evaluating the error metrics $A(0)$ and $A(1)$.

B.3.1 Training Objectives

We evaluate the target geometries of both spectral and InfoNCE by optimizing the embeddings v_i with respect to $\widehat{K}$ using different spectral- and InfoNCE-derived objectives. Specifically, for both spectral and InfoNCE, we compare two methods of utilizing the empirical kernel $\widehat{K}$ during optimization.

1. The *explicit-target* formulations use exact kernel values within the objectives to specify the desired similarities or conditional probabilities. In this way, we facilitate a highly informative gradient signal towards the optimum, so that this approach allows us to test whether the InfoNCE and spectral objectives are fundamentally capable of recovering the target geometry at all in this idealized setting.
2. In contrast, the *sampled-positive* formulations use the kernel solely to inform the drawing of positive pairs, thus providing a stochastic estimate of the general objective. Evaluating this approach therefore verifies whether the differences in the target geometries are still realized when the target distribution is conveyed through stochastic sampling rather than explicit guidance.

Spectral with Explicit Kernel Targets The most direct way to recover the kernel geometry is to minimize the squared difference between the inner products and the trainable embeddings and the corresponding entries of $\widehat{K}$. Let $s_{ij} = v_i^\top v_j$ denote the dot product between the embeddings, where the vector norms are freely optimized. Given a uniformly sampled batch of documents I of size b, we define the explicit-target spectral loss as:

$$\mathcal{L}^{\text{explicit}}_{\text{spectral}}(I) = \frac{1}{2b(b-1)} \sum_{\substack{i,j\in I \\ i\neq j}} \left(s_{ij} - \widehat{K}_{ij}\right)^2 . \tag{39}$$

As shown in Appendix A.3, this formulation is equivalent to the standard spectral objective up to an additive constant. However, providing the exact kernel values allows for direct supervision of the inner products, which therefore eliminates the variance inherent in recovering these target similarities solely from the stochastic co-occurrence of sampled positive pairs.

Spectral with Sampled Positives We also evaluate the standard spectral objective specified in Equation 7, utilizing $\widehat{K}$ solely to determine the sampling probabilities for positive pairs. Under the spectral paradigm, positive pairs are drawn from the joint co-generation distribution $p(x, y)$. To estimate this distribution from the empirical target $\widehat{K}$, we use the defining relationship:

$$k(x, y) = \frac{p(x, y)}{p(x)p(y)}. \tag{40}$$

Assuming a uniform distribution over the documents for negative sampling, the denominator is constant across all document pairs, so that $p(i, j) \propto \widehat{K}_{ij}$. We can therefore estimate the joint positive-pair probabilities over our finite document pool by normalizing the off-diagonal kernel entries:

$$\widehat{p}(i, j) = \frac{\widehat{K}_{ij}}{\sum_{u \neq v} \widehat{K}_{uv}}, \qquad i \neq j. \tag{41}$$

To sample a positive pair from this estimated joint distribution, we first draw a query document $I_\ell = i$ according to its marginal probability:

$$\Pr(I_\ell = i) = \sum_{j \neq i} \widehat{p}(i, j) = \frac{\sum_{j \neq i} \widehat{K}_{ij}}{\sum_{u \neq v} \widehat{K}_{uv}}. \tag{42}$$

Given this query, we then draw its positive counterpart $J_\ell = j$ from the corresponding conditional distribution:

$$\Pr(J_\ell = j \mid I_\ell = i) = \frac{\widehat{p}(i, j)}{\sum_{k \neq i} \widehat{p}(i, k)} = \frac{\widehat{K}_{ij}}{\sum_{k \neq i} \widehat{K}_{ik}}. \tag{43}$$

This ancestral sampling process is repeated independently to generate a batch of b positive pairs $\{(I_\ell, J_\ell)\}_{\ell=1}^{b}$.

For the uniformity term that requires negative samples, we draw an independent batch U of b distinct documents uniformly from the finite pool. The resulting stochastic loss is:

$$\mathcal{L}_{\text{spectral}}^{\text{sampled}} = -\frac{1}{b} \sum_{\ell=1}^{b} s_{I_\ell J_\ell} + \frac{1}{2b(b-1)} \sum_{\substack{i,k \in U \\ i \neq k}} s_{ik}^2. \tag{44}$$

Thus, $\widehat{K}$ determines only which positive pairs are sampled, while no kernel-dependent weights enter the loss, exactly mirroring the standard spectral setting.

InfoNCE with Explicit Kernel Targets Analogous to the explicit spectral formulation, we evaluate a version of InfoNCE that utilizes the exact kernel values, thus avoiding the need to recover the target distribution from sampled positives. Unlike spectral, which directly fits inner products to the kernel entries, InfoNCE applies a softmax over similarity scores to classify positive candidates. This means, that we provide explicit supervision by minimizing the divergence between the predicted similarities and the kernel-derived conditional probabilities.

Specifically, for a uniformly sampled batch of documents I of size b, let $u_i = v_i / \|v_i\|$ denote the unit-normalized embeddings. The target probabilities Q_{ij}^I and the predicted probabilities P_{ij}^I over the available candidates are defined as:

$$Q_{ij}^I = \frac{\widehat{K}_{ij}}{\sum_{k\in I\setminus\{i\}} \widehat{K}_{ik}}, \qquad P_{ij}^I = \frac{\exp(u_i^\top u_j/\tau)}{\sum_{k\in I\setminus\{i\}} \exp(u_i^\top u_k/\tau)}, \qquad j \in I \setminus \{i\}. \tag{45}$$

We then minimize the average Kullback-Leibler (KL) divergence between these distributions across the batch:

$$\mathcal{L}_{\mathrm{NCE}}^{\mathrm{explicit}}(I) = \frac{1}{b}\sum_{i\in I}\sum_{j\in I\setminus\{i\}} Q_{ij}^I \log \frac{Q_{ij}^I}{P_{ij}^I}. \tag{46}$$

Note that this form is equivalent to minimizing the standard InfoNCE loss for every candidate positive in the batch, weighted by a kernel-derived probability:

$$\mathcal{L}_{\mathrm{NCE}}^{\mathrm{explicit}}(I) \;\equiv\; \frac{1}{b}\sum_{i\in I}\sum_{j\in I\setminus\{i\}} Q_{ij}^I \left[-\log \frac{\exp(u_i^\top u_j/\tau)}{\sum_{k\in I\setminus\{i\}} \exp(u_i^\top u_k/\tau)}\right]. \tag{47}$$

The optimal unconstrained scores for this objective are $s_{ij}^* = \tau \log \widehat{K}_{ij} + c_i$, so that this kernel-informed formulation exactly preserves the logarithmic target of InfoNCE established by Theorem 1. We evaluate this objective using temperatures $\tau \in \{0.07, 0.20\}$.

InfoNCE with Sampled Positives We also evaluate the standard unweighted InfoNCE loss, utilizing $\widehat{K}$ solely to calculate the distribution of positive pairs. This is done by first drawing a uniform document batch I of size b. For each query $i \in I$, we then sample one positive J_i from the other documents in the batch according to the kernel-derived conditional distribution:

$$\Pr(J_i = j \mid I) = Q_{ij}^I = \frac{\widehat{K}_{ij}}{\sum_{k\in I\setminus\{i\}} \widehat{K}_{ik}}, \qquad j \in I \setminus \{i\}. \tag{48}$$

All other in-batch samples that were not selected as positive pair are then used as negatives in the standard InfoNCE formulation:

$$\mathcal{L}_{\mathrm{NCE}}^{\mathrm{sampled}}(I) = -\frac{1}{b}\sum_{i\in I} \log \frac{\exp(u_i^\top u_{J_i}/\tau)}{\sum_{k\in I\setminus\{i\}} \exp(u_i^\top u_k/\tau)}. \tag{49}$$

This formulation ultimately replaces the explicit target supervision of the explicit version with a stochastic estimate derived from sampled positives, which allows us to test whether the InfoNCE target geometry can still be recovered from a noisy, implicit training signal.

Full-Batch and Minibatch Comparisons For both spectral and InfoNCE, we evaluate three optimization protocols to systematically isolate the sources of training noise:

1. **Explicit full-batch:** We evaluate the explicit spectral and InfoNCE objectives using all $N = 1946$ documents in every update, which eliminates batch sampling noise and thus provides a purely deterministic gradient towards the optimal state.
2. **Explicit minibatch:** We evaluate the explicit objectives using random minibatches of size $b = 64$, which enables assessing the effect of gradient variance introduced by batch sampling.
3. **Sampled minibatch:** We evaluate the sampled objectives using minibatches of size $b = 64$, which closely mirrors the standard setups used in normal model training.

B.3.2 Experimental Setup

Data and Kernel Construction We perform the free-vector experiments on the same 1,946 documents used to construct the empirical CGK Gram matrix $\widehat{K}$ for evaluating the metric in Appendix

Objective	Protocol	τ	$A(0)$	$A(1)$	D
Spectral	Explicit full-batch	–	0.9973	0.0276	-0.9697
Spectral	Explicit minibatch	–	0.9972	0.0352	-0.9619
Spectral	Sampled-positive minibatch	–	0.9971	0.0475	-0.9496
InfoNCE	Explicit full-batch	0.07	0.4983	0.9876	$+0.4893$
InfoNCE	Explicit minibatch	0.07	0.4549	0.9913	$+0.5364$
InfoNCE	Sampled-positive minibatch	0.07	0.5150	0.9908	$+0.4758$
InfoNCE	Explicit full-batch	0.20	0.8145	0.9859	$+0.1714$
InfoNCE	Explicit minibatch	0.20	0.6754	0.9938	$+0.3183$
InfoNCE	Sampled-positive minibatch	0.20	0.7007	0.9937	$+0.2930$

Table 1: Alignment metrics of the free-vector fits for the three evaluated objectives and all the training protocols. Each entry averages results over eight embedding dimensions and five independent fits.

B.2, meaning that we use only the first summary from each pair, but this time constructing ϕ using all $28{,}000$ anchors.

We then fit vectors of dimensions $d \in \{64, 128, 256, 512, 768, 1024, 2048, 4096\}$ to $\widehat{K}$ using the six setups described in the previous subsection. To ensure robust results, we perform five independent runs from random initializations for every combination of objective, training protocol, and dimension, and report the average. Naturally, spectral embeddings are optimized with free vector norms, whereas InfoNCE embeddings are constrained to unit norm.

Optimization All embeddings are optimized using Adam, with learning rates tuned separately for each objective and sampling strategy. Minibatch experiments use a batch size of $b = 64$ and run for a maximum of $200{,}000$ updates, while the full-batch setup is allowed a reduced number of $10{,}000$ optimization steps due to quicker convergence. To select the best checkpoint, we evaluate the training objective over all available documents every 100 steps and save the fit with the lowest loss. Thus, we evaluate the point at which each objective best matched the target $\widehat{K}$.

Evaluation The goal of the experiments is to evaluate whether the InfoNCE and spectral objective realize their theoretically derived target geometries. Thus, we construct the predicted similarity matrix C, which consists of raw dot products for spectral and cosine similarities for InfoNCE, and evaluate it against the target $\widehat{K}$ using the metrics introduced in Appendix B.2. Because these embeddings are fitted and evaluated against the exact same empirical kernel $\widehat{K}$, we do not require noise correction.

Results are always reported as the mean over the five random initializations, and whenever results across dimensions are aggregated, each evaluated dimension receives equal weight.

B.3.3 Results

Recovery of the Predicted Geometries We first analyze the overall alignment with the theoretically derived target geometries, averaged across all embedding dimensions, which is displayed in Table 1.

Starting with the **explicit full-batch** setup, a deterministic training paradigm that optimizes the exact kernel values or probabilities over the entire document pool, we see that spectral nearly perfectly recovers its intended linear target with an error of just $A(1) = 0.0276$, while the non-matching logarithmic target remains entirely uninformative, yielding $A(0) = 0.9973$. Similarly, InfoNCE successfully aligns with its logarithmic target, achieving $A(0) = 0.4983$ at $\tau = 0.07$ (and 0.8145 at $\tau = 0.20$), while showing essentially zero alignment with the linear target ($A(1) \approx 0.99$).

The higher absolute error for InfoNCE relative to spectral is a direct consequence of the representational bottleneck identified in Section 3: because $\log K$ is heavily indefinite, it cannot be perfectly realized as a Gram matrix in any Euclidean space, thus leaving a strict lower bound on the achievable error even under optimal deterministic training.

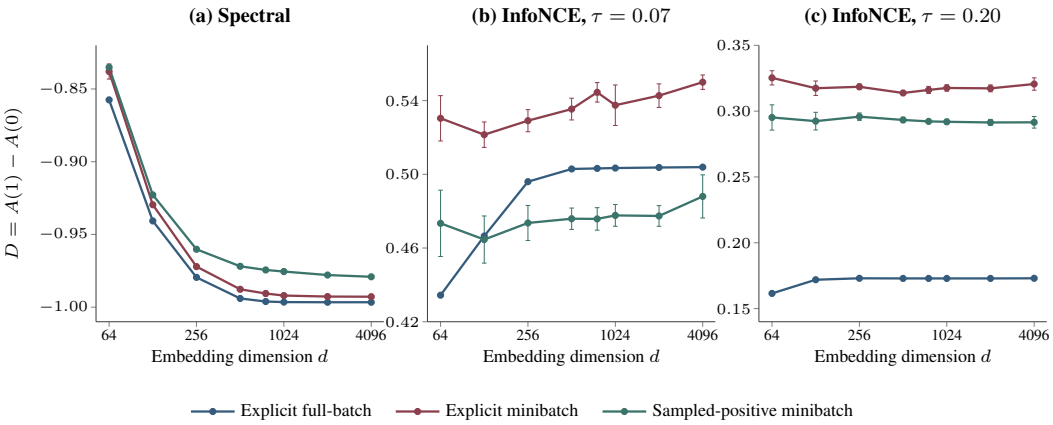

Figure 4: Alignment difference $D = A(1) - A(0)$ as a function of embedding dimension for spectral and InfoNCE under three training protocols.

Next, we evaluate the **explicit minibatch** setup, which introduces noise in the gradient estimation through the sampling of minibatches, while retaining the more informative kernel-informed gradient signal. For spectral, this stochasticity slightly reduces the precision of the fit, increasing the linear error to $A(1) = 0.0352$, but the logarithmic error remains near 1.0. For InfoNCE, explicit minibatch optimization yields a logarithmic error of $A(0) = 0.4549$ at $\tau = 0.07$ (and 0.6754 at $\tau = 0.20$), with the linear target again remaining uninformative ($A(1) \approx 0.99$). Both objectives thus strongly preserve their respective geometric preferences despite the additional batch sampling variance.

The **sampled-positive minibatch** setup then further increases the noise in the gradient estimation by using the kernel solely for selecting positive samples stochastically. This setup therefore mirrors standard contrastive training, where the similarity structure needs to be inferred solely from co-occurrence statistics. Interestingly, relying on this weaker signal still preserves the predicted linear geometry for spectral, achieving a linear fit of $A(1) = 0.0475$ alongside a completely uninformative logarithmic fit ($A(0) = 0.9971$). As expected, this alignment is slightly less precise than the explicit minibatch variant due to the increased sampling noise, but it still demonstrates a highly accurate recovery of the linear target.

For InfoNCE, using the sampled-positive objective slightly degrades the logarithmic fit relative to the explicit minibatch setup, raising $A(0)$ to 0.5150 at $\tau = 0.07$ and 0.7007 at $\tau = 0.20$. Nevertheless, the linear error remains near 1.0, which confirms that pure stochastic sampling preserves the underlying logarithmic geometry.

Effect of Embedding Dimension We also analyze the effect of embedding dimensionality on the alignment with the respective target geometries. Figure 4 displays the target error difference D as a function of the embedding dimension d across all training paradigms.

For the spectral objective, the contrast D initially drops sharply toward the linear endpoint of -1 as d increases, but plateaus around $d = 512$ once the embeddings achieve near-perfect alignment with the linear target. This general trajectory is highly consistent across all three training setups, but the sampled-positive variant consistently shows a slightly weaker (but still very strong) alignment, likely caused by the increased stochasticity in the gradient estimation.

For the InfoNCE setups, on the other hand, the scaling behavior is very different. The improvement in alignment with the logarithmic target with increasing dimensionality d is minimal, with D growing by less than 0.04 for five of the six evaluated paradigms. This early saturation is caused by the nature of the logarithmic target. The saturation of spectral already indicates that roughly 512 dimensions are sufficient to capture almost all of the structure of $\widehat{K}$, which is far fewer than its full rank of $1{,}946$. The logarithmic target, however, is additionally highly indefinite, meaning that no reconstruction arising from embeddings via a dot product can reproduce it exactly. And since only the part of its structure that a positive semidefinite Gram matrix can represent is recoverable, this

Objective	Protocol	τ	$A(0)$	$A(1)$	D
$\widehat{K}_2$ control	Cross-replica	–	0.9961	0.8932	-0.1028
$\log \widehat{K}_2$ control	Cross-replica	–	0.3005	0.9491	$+0.6485$
Spectral	Explicit minibatch	–	0.9965	0.9101	-0.0864
Spectral	Sampled-positive minibatch	–	0.9966	0.9093	-0.0874
InfoNCE	Explicit minibatch	0.07	0.4890	0.8747	$+0.3857$
InfoNCE	Sampled-positive minibatch	0.07	0.5224	0.8675	$+0.3451$
InfoNCE	Explicit minibatch	0.20	0.6020	0.8571	$+0.2551$
InfoNCE	Sampled-positive minibatch	0.20	0.6215	0.8560	$+0.2345$

Table 2: Noise-corrected alignment after fitting free vectors on $\widehat{K}_2$ and evaluating against independent estimates constructed from the held-out anchor group $\widehat{K}_1$. The first two rows additionally show the control of evaluating the alignment of $\widehat{K}_2$ and $\log \widehat{K}_2$ with the independent $\widehat{K}_1$.

part is likely captured with even fewer dimensions, thus explaining stagnation starting already in the range of $d = 64$-256.

B.3.4 Evaluation under Anchor Noise

General Setup The preceding experiments optimize and evaluate the embeddings on the exact same empirical kernel matrix $\widehat{K}$. However, if subsequent experiments train models in realistic contrastive learning settings, they implicitly optimize towards the true CGK defined by the underlying data-generating distribution, while the alignment evaluation can only use the empirical matrix $\widehat{K}$, which is just a noisy estimate. To verify that our metric can reliably identify the correct geometric preferences when evaluating against a noisy approximation of the true target, we now evaluate a setup where models are optimized against one kernel estimate ($\widehat{K}_2$) but are evaluated against an entirely independent estimate ($\widehat{K}_1$) of the same kernel. The remaining experimental setup remains identical to the previous section, except for restricting the evaluation to the more realistic minibatch protocols.

Creating Target Matrices In practice, the two estimates $\widehat{K}_1$ and $\widehat{K}_2$ are constructed by partitioning the $28{,}000$ available anchors into two equally sized disjoint sets. We assign the first $14{,}000$ anchors to construct the training target matrix $\widehat{K}_2$, while the remaining $14{,}000$ anchors are further divided into two disjoint subsets to form two independent reference replicas, $\widehat{K}_1^{(1)}$ and $\widehat{K}_1^{(2)}$. These two independent estimates are required to evaluate the predicted embeddings using the noise-corrected metric defined in Equation equation 37.

Cross-Replica Controls To establish an upper bound on the alignment that any trained embedding can reasonably achieve, we first evaluate the direct alignment of the raw matrices $\widehat{K}_2$ and $\log \widehat{K}_2$ against the $\widehat{K}_1$ estimates, with results being included in Table 2. The raw linear alignment of $\widehat{K}_2$ with $\widehat{K}_1$ is present but weak ($A(1) = 0.8932$), but due to having essentially zero alignment with the logarithmic target ($A(0) = 0.9961$), it still produces a clearly negative contrast of $D = -0.1028$. For the logarithmic estimate $\log \widehat{K}_2$, alignment with the intended target is much stronger ($A(0) = 0.3005$), resulting in a highly positive contrast of $D = +0.6485$. In summary, this confirms that our noise-corrected metric can successfully identify both the linear and logarithmic geometries despite severe anchor noise.

Results Alignment results for the free-vector fits under anchor noise are presented in Table 2, averaged across all evaluated dimensions. Despite the noisy training target, both spectral setups display alignment that almost matches the linear control baseline, yielding contrasts of $D = -0.0864$ (explicit) and $D = -0.0874$ (sampled-positive). Similarly, all InfoNCE protocols remain robustly on the predicted (logarithmic) side, with contrasts ranging between $D = +0.2345$ and $D = +0.3857$.

B.3.5 ALIGNMENT CONSISTENCY

We find that across both the noise-free and noisy experiments, the alignment contrast D has the theoretically expected sign in every single run: all 264 spectral fits (120 noise-free, 144 under anchor noise) yield $D < 0$, and all 528 InfoNCE fits (240 noise-free, 288 under anchor noise) yield $D > 0$. This perfect consistency holds regardless of the training protocol (full-batch, explicit minibatch, or sampled-positive), temperature parameter, and embedding dimension. These results therefore robustly support the following conclusions:

1. Both spectral and InfoNCE fundamentally drive embeddings toward their theoretically derived target geometries. Since these targets are completely independent from any hyperparameter or sampling protocol, it is an intrinsic property of the objective itself rather than an artifact of any specific experimental setup.
2. The structural distinction between linear and logarithmic fits remains fully measurable even when only a stochastic estimate of the kernel is available, thus supporting the use of the estimated CGK matrix $\widehat{K}$ in subsequent experiments that analyze standard contrastive embedding model training.

B.4 TRANSFORMER EMBEDDING KERNEL RECOVERY

The free-vector results demonstrated that the spectral and InfoNCE objectives reliably push toward their theoretically derived target geometries. We now empirically evaluate whether this property holds when the objectives are used to train transformer encoders, where shared parameters across samples substantially alter the optimization dynamics.

General Setup To systematically isolate the effects of 1) the shared encoder and 2) generalization of the target geometries to unseen data, we evaluate two different setups:

1. The first experiment evaluates an idealized setup where a transformer is fitted to, and evaluated against, the exact same empirical kernel $\widehat{K}$. This setup therefore differs from the free-vector experiments mainly through the introduction of a shared neural encoder operating on the actual textual inputs, which therefore still removes the need to generalize the learned geometry to unseen data.
2. The second experiment then trains embedding models using the standard InfoNCE and spectral objectives on the much larger SemCSE dataset, thus evaluating both target geometries under the realistic settings of normal contrastive learning. Since no information from $\widehat{K}$ is used during training, a subsequent evaluation on the documents underlying $\widehat{K}$ allows us to assess whether the models learn an embedding geometry that successfully generalizes to unseen data.

B.4.1 MODEL DETAILS

Across both experiments, we use a SciDeBERTa model (Kim et al., 2023) as base encoder, which is a transformer encoder model with 183 million total parameters and embedding dimension $d = 768$ pre-trained on large-scale scientific text. This model is chosen to align with the original SemCSE research, as well as due to providing solid performance with low computational cost. Note that we perform additional training runs with a larger Qwen3.5-9B (Qwen Team, 2026) to evaluate the actual downstream performance in Appendix B.5.

Embeddings are created by first processing the `[CLS]` token embedding of the base encoder via another linear projection that preserves the output dimension, and the purely directional embedding is then obtained via subsequent L_2 normalization. For methods that only require normalized embeddings (e.g., InfoNCE), this constitutes the final output of the model.

For the spectral objective, which requires unnormalized embeddings, the network includes an additional scalar magnitude head. This head applies a linear transformation to the normalized output embedding, followed by a softplus activation to predict the strictly positive magnitude. This design explicitly decouples the directional component from the embedding magnitude, an architectural choice we empirically found to substantially improve training stability and downstream performance.

Objective	Fit → evaluation	$A(0)$	$A(1)$	D
Spectral	$\widehat{K} \to \widehat{K}$	0.9900	0.6310	-0.3591
	$\widehat{K}_2 \to \widehat{K}_2$	0.9839	0.5263	-0.4576
	$\widehat{K}_2 \to \widehat{K}_1$	0.9900	0.8563	-0.1337
InfoNCE$_{\tau=0.07}$	$\widehat{K} \to \widehat{K}$	0.4214	0.9926	$+0.5712$
	$\widehat{K}_2 \to \widehat{K}_2$	0.4082	0.9842	$+0.5760$
	$\widehat{K}_2 \to \widehat{K}_1$	0.4697	0.8869	$+0.4171$
InfoNCE$_{\tau=0.2}$	$\widehat{K} \to \widehat{K}$	0.6720	0.9939	$+0.3219$
	$\widehat{K}_2 \to \widehat{K}_2$	0.6556	0.9872	$+0.3316$
	$\widehat{K}_2 \to \widehat{K}_1$	0.6029	0.8601	$+0.2573$

Table 3: Alignment of an encoder trained directly on the kernel targets using the sampled minibatch protocol for 100k updates. Models are either trained and evaluated on $\widehat{K}$ ($\widehat{K} \to \widehat{K}$), or trained on an estimate $\widehat{K}_2$ and evaluated both on the same kernel ($\widehat{K}_2 \to \widehat{K}_2$) and on an independent estimate $\widehat{K}_1$ ($\widehat{K}_2 \to \widehat{K}_1$).

B.4.2 Evaluating Supervised Kernel Recovery

This section trains an encoder using explicit supervision from the empirical kernel estimates, thus mirroring the setup from the free-vector fits.

Experimental Setup We fit the encoder directly to a target kernel using the sampled loss function specified in Appendix B.3.1. To establish results both in the known-target and noisy-target setups discussed before, we perform two different experiments:

1. In one training run, we compute the kernel $\widehat{K}$ using all $28{,}000$ anchors, using this kernel both as optimization target and as evaluation reference (mirroring Appendix B.3.2 / B.3.3)
2. A second run creates two independent estimates $\widehat{K}_1$ and $\widehat{K}_2$ of the underlying K and optimizes towards $\widehat{K}_2$ while evaluating against $\widehat{K}_1$ (mirroring Appendix B.3.4)

The kernel matrix creation is identical to the free-vector references. In contrast to the earlier experiments, each of the $1{,}946$ documents is now represented by the first of its generated summaries, truncated to 128 tokens.

All models are optimized with Adam (learning rate $3 \cdot 10^{-5}$) for $100{,}000$ updates, tracking the alignment metrics throughout training to get an accurate picture of the change in the realized target geometry over time.

Results Table 3 displays the alignment evaluations of the trained encoders at the end of training. When trained on and evaluated against the full kernel $\widehat{K}$, both spectral and InfoNCE are identified as targeting the same geometry observed for the free vectors: spectral yields $D = -0.3591$ while InfoNCE with $\tau = 0.07$ yields $D = 0.5712$.

These initial results suggest that the shared encoder affects the two objectives differently, since InfoNCE recovers the logarithmic geometry better than the free vectors ($A(0)$ of 0.4214, compared to 0.45-0.52 for free-vector minibatch fits), while the linear alignment of spectral is considerably weaker ($A(1)$ of 0.6310, compared to 0.04-0.05 for free vectors).

This difference can, however, be explained by inspecting the progression of the alignment metrics over the course of training, displayed in Figure 5. We see that the InfoNCE variants reach their final geometric contrast already quickly after the start of training, while the spectral loss displays substantially slower convergence, with the optimum not being reached at the end of the training budget. We therefore conclude that the reduction in linear alignment is mostly caused by the limited number of training steps, and not by a fundamental limitation imposed by the shared encoder.

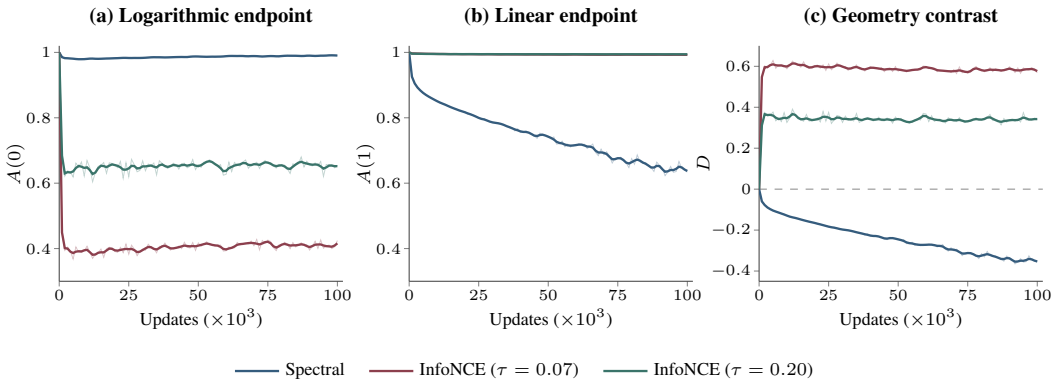

Figure 5: Alignment of the encoder's predicted similarities with $\widehat{K}$ across training, evaluated on the full kernel.

The general conclusion that the intended target geometries are reliably realized by both objectives is further strengthened by noting that the alignment metric D consistently has the expected sign for the matching target geometry even under the noisy conditions of training on one kernel estimate $\widehat{K}_2$ and evaluating on an independent kernel estimate $\widehat{K}_1$.

B.4.3 Geometry Recovery under Standard Contrastive Training

The previous experiments demonstrated that both the spectral and InfoNCE objectives reliably drive representations toward their theoretically derived target geometries, even in the presence of sampling noise or shared parameters in the form of a text encoder. However, all of these setups used explicit information from the empirical kernel $\widehat{K}$ (or an independent estimate of it) during training, either by directly fitting its exact values or by using it to define the positive sampling distribution.

This section now instead evaluates a fully realistic training paradigm by training embedding models using the standard spectral and InfoNCE objectives on an unrelated, large-scale dataset, thus providing no access to $\widehat{K}$ during optimization. We then evaluate whether these models still learn their distinct embedding geometries in a way that successfully generalizes to the unseen documents underlying the evaluation kernel $\widehat{K}$.

Experimental Setup We use the exact same model architecture as in the previous kernel-informed transformer-fit experiments. The network is optimized using either the standard InfoNCE objective (Equation equation 5) evaluated at temperatures $\tau \in \{0.07, 0.20\}$, or the spectral objective (Equation equation 7).

As training dataset, we use the dataset from the original SemCSE experiments (Brinner & Zarrieß, 2025), which consists of approximately 350,000 scientific abstracts with three summarizing sentences generated for five different summary-initiating prompts. However, to ensure that the objectives optimize toward the exact same theoretical target kernel that is approximated by $\widehat{K}$, we only use the three summaries corresponding to the same prompt that was used to generate the data underlying $\widehat{K}$.

All models are trained for 100K steps using a batch size of 128. Optimization is performed using Adam with a learning rate of 10^{-5}. Rather than evaluating a single final checkpoint, we track multiple metrics (including the linear and logarithmic alignment metrics on the unseen evaluation documents of $\widehat{K}$) throughout the entire training process. In this way, we can gain a deeper insight into the underlying optimization dynamics and their influence on the resulting geometry.

Alignment with the Target Geometries. Figure 7 displays the progression of the individual alignment errors $A(0)$ and $A(1)$ and of the alignment contrast D over the course of training. While the objectives generally show very different geometric trajectories, they share one interesting similarity:

all models rapidly lose alignment with the logarithmic target, as seen by the increasing logarithmic error $A(0)$. As a result, by the end of training, this logarithmic geometry explains less than 3% of the overall variance. While this behavior is naturally expected for the spectral loss that targets the linear structure, it is highly surprising for the InfoNCE variants, which completely fail to track their theoretically derived target geometry.

For the linear geometry, however, all objectives display a different behavior from each other. As expected by the previous analyses, the spectral loss rapidly increases its alignment with the linear target, reaching a plateau after approximately $50{,}000$ updates at a residual error of $A(1) \approx 0.90$ (with a slight decrease afterwards) and thus indicating a successful partial recovery of the intended linear structure. While an absolute error of 0.90 may appear high, we need to contextualize it by remembering that the evaluation matrix $\widehat{K}$ is itself just a noisy approximation of the true underlying CGK. As previously established in Table 2, evaluating two completely independent empirical estimates $\widehat{K}_1$ and $\widehat{K}_2$ against each other yields a baseline linear error of $A(1) = 0.8932$, despite both matrices provably approximating the exact same target. This cross-replica value can therefore be seen as a reasonable estimate for the maximum achievable linear alignment. We note, however, that the current experiment uses slightly larger reference sets ($14{,}000$ anchors per replica compared to $7{,}000$ previously, caused by now using a two-way rather than a three-way anchor split). This should improve the estimate's precision, but does not remove the fundamental estimation limit. In this context, the spectral loss therefore achieves a reasonably high alignment with the linear target geometry, at least as far as can be analyzed with the given precision available for this estimation.

Interestingly, the InfoNCE variant at $\tau = 0.20$ also initially increases its alignment with the linear endpoint. But it does so to a lesser degree than the spectral loss, and shifts from its maximum alignment to a steady decrease earlier during training. In contrast, the InfoNCE objective at $\tau = 0.07$ exhibits a steady decrease in linear alignment already right from the start. Combining both $A(0)$ and $A(1)$, both InfoNCE variants yield a negative overall alignment contrast ($D < 0$). However, rather than interpreting it as a genuine preference for the linear geometry, we recognize that this negative contrast is primarily driven by the fast and almost full loss of their logarithmic alignment.

In this context, the next paragraphs focus on explaining this unexpected lack of logarithmic alignment for the InfoNCE models. Furthermore, we analyze potential causes of why the spectral objective recovers its target geometry more accurately, and try to identify factors that drive the distinct differences between the InfoNCE variants at different temperatures.

Memorization of Training Data. A natural explanation for both the failure to recover the logarithmic target geometry and for the reduction of the spectral loss's linear alignment later in training is an overly strong adaptation to the finite training dataset. With a batch size of 128 and approximately 350K generating anchors, the first repeated document is encountered after just $2{,}714$ updates. Also, because only three summaries are available per anchor, the subsequent epochs not only repeatedly train the model on the exact same general topics, but even on the same exact texts. Figure 6 plots the training loss as an exponential moving average alongside the validation loss. We see that training loss falls below the validation loss for all methods, with this point being reached at around 14K updates for InfoNCE at $\tau = 0.07$, compared to 19.5K for $\tau = 0.20$ and 32K for the spectral objective. Additionally, the difference between the training and validation loss grows much more slowly for the spectral objective.

We attribute this distinct ordering in part to how the respective objectives treat hard negative samples. In Appendix B.5, we analyze how strongly hard negatives (i.e., negatives with a high similarity to the anchor) are pushed away from the anchor relative to easy negatives, quantified via the relative repulsion ratio. We find that low-temperature InfoNCE concentrates most of its gradient signal on repelling hard negatives compared to higher-temperature InfoNCE or the spectral loss. Because existing studies identify hard negatives as a primary factor for contrastive learning (Wang & Liu, 2021; Kalantidis et al., 2020; Robinson et al., 2021), adapting strongly to these pairs within the first encounters can lead to a quick adaptation to the specifics of the training data (thus quickly reducing the training loss) while not necessarily benefiting generalization. The fact that for InfoNCE at $\tau = 0.20$ the training loss falls below the validation loss later than for $\tau = 0.07$ perfectly aligns with this explanation, since the only difference between these settings is the temperature, which solely affects the relative repulsion ratio. The even lower relative repulsion of the spectral loss then

leads to an even slower training-specific adaptation, which could explain why its linear alignment continues to improve for substantially longer before eventually deteriorating.

This interpretation is further supported by the neighborhood-ranking evaluation in Figure 8(a), which measures the NDCG@50 agreement between the top 50 nearest neighbors predicted by the models and those defined by the empirical kernel $\widehat{K}$. Initially, all three objectives successfully increase the alignment of their neighborhood structures with $\widehat{K}$, leading to NDCG@50 values in the range of 0.33-0.38. However, this alignment eventually peaks and then trends downward continuously. The important observation is here, that the order of these peaks again aligns with the relative repulsion ratios, and thus also with how quickly the train loss becomes substantially lower than the validation loss. Specifically, InfoNCE at $\tau = 0.07$ peaks very quickly at around 3K steps, followed by InfoNCE at $\tau = 0.20$ (peaking at about 21K steps), while the spectral loss continues to improve considerably longer (peaking at about 50K steps). This pattern therefore again indicates that a strong adaptation to the hard negatives leads to an overly strong adaptation to training data, which speeds up the transition from learning generalizable structure towards memorizing training data.

Ultimately, while the hypothesized difference in transition from generalization to memorization can explain the timing of specific peaks and declines, it does not explain why higher-temperature InfoNCE develops a preference for the linear endpoint, or why both InfoNCE variants begin losing their logarithmic alignment immediately at the start of training, long before any duplicate data has been observed.

Neighborhood Learning and Broad Kernel Ordering. Figure 8 gives more insight into which structural components of the kernel are successfully learned and how they progress during training. Specifically, Figure 8 (b) displays the Spearman correlation between the similarity values of $\widehat{K}$ and the model's predictions, computed for each row independently and averaged afterwards. We see that the correlation is high initially but then begins to decline, which is consistent across all objectives. Final values at the end of training are below 0.06 for all methods, and even near 0 for InfoNCE with $\tau = 0.20$, thus suggesting almost no correlation between the trained embeddings and $\widehat{K}$.

In contrast, NDCG@50 improves initially and then decreases only moderately, stabilizing between approximately 0.31 and 0.34 across objectives. These scores thus remain far above the random ranking baseline of 0.021, and are close to the value of 0.38 measured if evaluating two estimates $\widehat{K}_1$ and $\widehat{K}_2$ against each other (which serves as a realistic upper bound for the maximum achievable score under the available precision).

The fact that overall correlation is low and NDCG@50 is reasonably high therefore suggests that all models learn to reliably identify the closest neighbors according to $\widehat{K}$, but are incapable of learning the more fine-grained ranking of most of the low-similarity pairs at the lower end of the spectrum. To fully confirm this hypothesis, we evaluate the Spearman correlation just on the top 50 neighbors according to $\widehat{K}$, resulting in values between 0.17 and 0.24 that remain stable throughout training. In contrast, evaluating Spearman solely outside this top 50, the correlation falls to between -0.03 and 0.03, indicating no alignment at all.

The likely explanation for this inability to learn the more fine-grained ranking structure is the highly stochastic and thus noisy training signal caused by the limited batch size. For example, each training batch only has one positive per anchor, thus moving anchor and positive closer together. If each batch contained many positives for each anchor, then the direction in which this anchor is moved is instead determined by the centroid of these positive embeddings, meaning that the resulting update could even increase the distance of that anchor to certain "outlier" positives. Also, the positives themselves could increase their distance to the anchor if they appear as positives to other anchors within the same update.

In the noisy minibatch setup, these dynamics have to instead be estimated from occurrence statistics across batches, which is impossible for the most fine-grained relationships. In combination with the fact that the limited training data does not provide full coverage of the target distribution, the result is a learning process biased toward capturing strong, highly frequent associations rather than learning the precise rankings of the exact kernel.

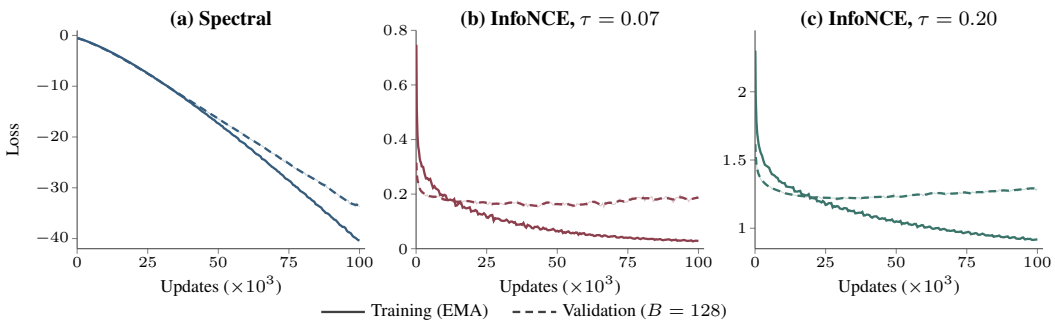

Figure 6: Training loss EMA and validation loss over the course of training for the Spectral loss and both InfoNCE variants.

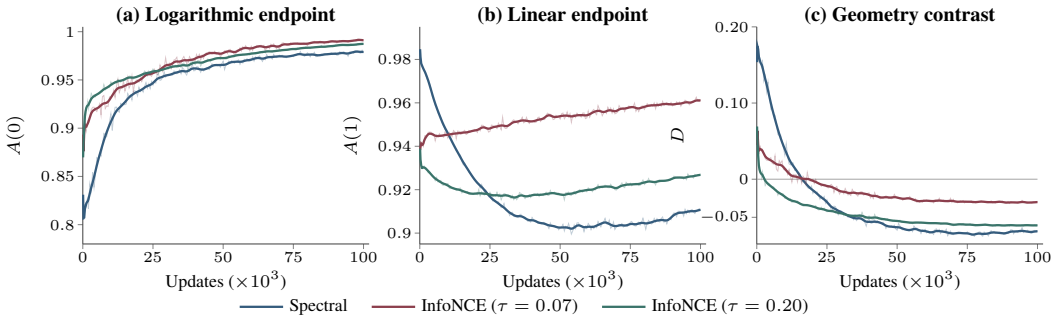

Figure 7: Noise-corrected alignment metrics of predicted embedding similarities vs. $\widehat{K}$ over training.

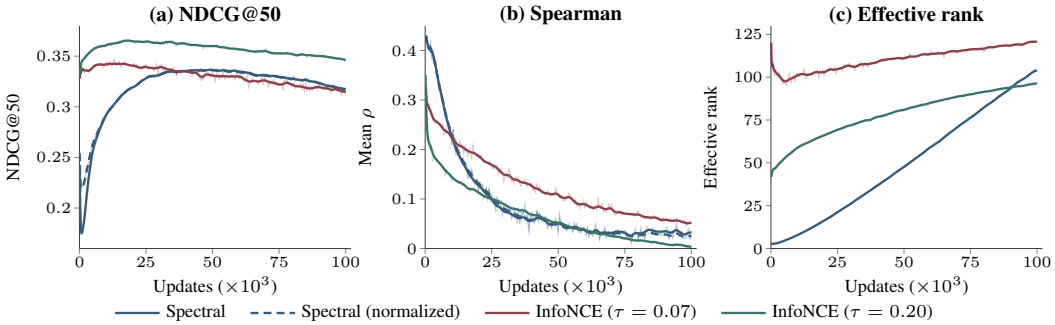

Figure 8: Alignment of the trained embedding similarities with the empirical kernel $\widehat{K}$, evaluated via (a) NDCG@50 and (b) Spearman correlation, alongside (c) the effective rank of the embeddings. All metrics are tracked over the course of training for the spectral objective and both InfoNCE variants.

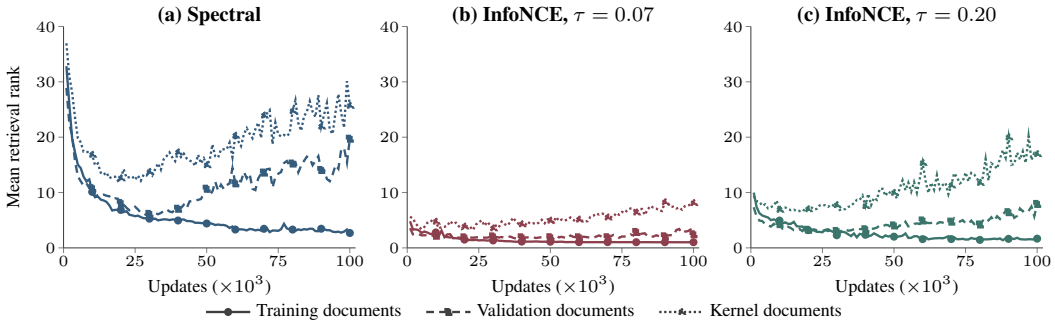

Figure 9: Mean retrieval rank on summary-pair retrieval from a pool of 600 distractors.

Effect of the InfoNCE Temperature. In several metrics we observe a clear difference between the two InfoNCE temperatures $\tau = 0.07$ and $\tau = 0.20$, even though the theoretical optimum according to Theorem 1 is equivalent up to a simple rescale. The important difference is however introduced by restricting the optimization to unit-normalized embeddings, in which case the temperature parameter fundamentally alters the value ranges that the model is able to predict. Specifically, a higher temperature compresses the range of the possible logits and thus limits how strongly the model can distinguish candidate probabilities. Together with the weaker relative repulsion of hard negatives, this bottleneck can favor preserving general neighborhood structures over optimizing for finer distinctions between samples.

The stronger apparent preference for the linear endpoint at $\tau = 0.20$ may thus reflect this structural shift, even though the theoretical optimum of the unrestricted objective remains strictly logarithmic. Supporting this analysis is the fact that this temperature-dependent degradation of the logarithmic target is already observed in the earlier free-vector experiments, where sampled-positive fits yielded a logarithmic error of $A(0) = 0.5150$ at $\tau = 0.07$, compared to a substantially worse 0.7007 at $\tau = 0.20$ (Table 1). As a result, we conclude that the shift towards a more linear behavior for higher InfoNCE temperatures is likely an intrinsic property of the optimization dynamics rather than just being caused by the generalization to unseen documents.

Overall Findings. Standard contrastive training transfers useful neighborhood information, but does not reproduce the kernel's broad similarity structure. Spectral partially recovers its linear target, approaching the earlier independent-kernel benchmark before declining, whereas both InfoNCE variants progressively lose their logarithmic alignment. Limited positive coverage and increasing training-data specialization provide a consistent explanation for this gap between the theoretical optima and the geometries learned on unseen data.

B.5 Analyzing Downstream Performance

The previous sections focused solely on evaluating whether spectral and InfoNCE realize their theoretically derived target geometries under increasingly realistic training conditions. While these analyses suggest that neither objective realizes its target geometry particularly well in the most realistic contrastive learning setup using the SemCSE dataset, the two objectives still have very different optimization dynamics, which might result in substantial differences in downstream performance. This section therefore centers around evaluating the downstream performance of embedding models trained with different contrastive objectives, thus yielding insight into their specific strengths and weaknesses.

B.5.1 Contrastive Objectives

We evaluate both the spectral and InfoNCE objectives, as they are the central subjects of analysis in this study. For InfoNCE, we evaluate a temperature of $\tau = 0.07$, as this is the choice made by the original SemCSE paper and is also a commonly used default value. We additionally evaluate $\tau = 0.2$, since Wang & Liu (2021) demonstrated that this value can lead to improved clustering performance. Because spectral performs especially well on clustering tasks, evaluating InfoNCE in this favorable setting for this task type is an important comparison, and thus also motivated the use of this temperature throughout the previous experiments.

In addition to these main objectives, we additionally evaluate two modifications of the standard InfoNCE objective that were shown to have favorable properties in previous studies:

1. The SimACE objective (Jeong et al., 2024), which replaces the cosine similarity in the InfoNCE logits with an angular similarity, $\pi/2 - \arccos(\cos(a, b))$, and subtracts an additive angular margin from the positive pair's logit. Following the original work, we use $\tau = 0.05$ and a margin of $10°$.
2. Multiple previous works have shown that omitting in-batch negatives with high similarity to the anchor can improve performance (Solatorio, 2024; Zhou et al., 2022), so we adopt this setup as a further baseline. Specifically, we use the existing SemCSE model, as trained by the original work, as a judge model on each batch and filter out any in-batch negative whose cosine similarity to the anchor exceeds 0.9. We use $\tau = 0.07$ for this variant.

B.5.2 MODEL

We choose largely the same setup as in the previous experiments to predict embeddings. Specifically, we use a pretrained model as the encoder, whose output is mapped through an additional linear projection that we initialize as a whitening matrix computed on the encoder's outputs for a subset of the training texts. The projected embedding is then again normalized to unit length, constituting the final output for most methods. For spectral in particular, a separate norm-prediction head (a linear layer followed by a softplus) additionally predicts the embedding's magnitude from the normalized direction, and pairs are scored by the product of the two magnitudes and the cosine similarity (thus mirroring a dot product).

In contrast to the previous experiments, we use a Qwen3.5-9B model (Qwen Team, 2026) because of the superior knowledge it acquired during pretraining compared to the smaller encoder model used before. We embed a text by appending an `EOS` token to it and using that token's final hidden state as input to the projection. The encoder is fully fine-tuned together with the projection and norm head.

The hidden dimension of the base model is 4,096, but since certain task evaluations of the SciRepEval benchmark (Singh et al., 2023) increase substantially in their computational cost with increased embedding dimension, we chose to set the final linear projection to have a final output dimension of 2,048.

B.5.3 TRAINING SETUP

We largely adopt the training setup of the original SemCSE paper, including the following design choices:

1. **Data.** Approximately 350K scientific abstracts from six source corpora from the SciRepEval benchmark (Singh et al., 2023), each paired with 15 LLM-generated summarizing sentences.
2. **Positive pairs.** A positive pair consists either of two summaries of the same abstract (50% probability), a summary and the paper's title (15%), or a summary and a randomly chosen sentence of the abstract (35%).
3. **Weighted sampling.** Source datasets are sampled with probability proportional to the square root of their size, which evens out the topic distribution.
4. **Model selection.** The model is validated every 1,000 updates on a retrieval task using a held-out validation set of abstracts and summaries. Training stops once validation performance has not improved for five consecutive evaluations.
5. **Input truncation.** All inputs are truncated to at most 128 input tokens.

We further choose to use a batch size of 32, AdamW with learning rate 10^{-5} and weight decay 10^{-4}, as well as gradient clipping at norm 1.0.

B.5.4 BENCHMARKS

We evaluate the trained models on two benchmarks, with the first being the SciRepEval benchmark (Singh et al., 2023). Covering diverse scientific domains, this benchmark comprises 23 different tasks for evaluating scientific document representations, split into four different main categories: *classification* of papers into topics or categories, *regression* (i.e., the prediction of continuous properties such as citation counts), *proximity*-based ranking of related papers, and ad-hoc *search* (i.e., matching search queries to papers). We report the average score per task category as well as the overall average across all tasks.

Since many SciRepEval tasks rely on auxiliary signals of similarity like citations, authorship or click-through data, which do not necessarily align perfectly with actual semantic relatedness, we additionally evaluate the models on the semantic benchmark introduced by Brinner & Zarrieß (2025) alongside the development of the SemCSE method. This benchmark aims at evaluating more accurately if embedding models actually encode the semantic content of the underlying scientific text. Three of its tasks are formulated as retrieval: matching a paper's title to its abstract, matching the two halves of a split abstract to each other, and matching a search query to its abstract. For each, we

Objective	Classification	Regression	Proximity	Ad-hoc Search	Overall
InfoNCE ($\tau = 0.07$)	65.19	32.75	79.47	78.41	65.45
InfoNCE ($\tau = 0.2$)	66.01	32.07	**80.04**	**78.87**	**65.80**
InfoNCE (filtered)	65.16	32.42	79.39	78.43	65.34
SimACE	65.07	**32.80**	79.38	78.51	65.41
Spectral	**66.41**	31.69	79.70	78.82	65.68
Spectral w/o norm head	61.59	26.01	78.35	77.67	62.51

Table 4: SciRepEval results of Qwen3.5-9B embedding models ($d = 2048$) trained with different contrastive objectives. Columns show the average score per task category and the overall average across all 23 tasks. The best value per column is shown in bold.

Objective	Title–Abstract ↓	Abstract Segments ↓	Query ↓	Clustering ↑
InfoNCE ($\tau = 0.07$)	9.86	7.80	8.04	0.720
InfoNCE ($\tau = 0.2$)	10.56	**5.32**	**6.94**	0.735
InfoNCE (filtered)	**9.69**	7.68	8.03	0.721
SimACE	10.20	8.90	8.38	0.721
Spectral	21.09	8.48	7.11	**0.739**
Spectral w/o norm head	37.65	11.69	14.63	0.714

Table 5: Results of the same models on the semantic benchmark from the SemCSE paper. The three matching tasks report the mean rank of the correct match (lower is better); clustering reports the mean k-NN accuracy ($k = 5$, higher is better). The best value per column is shown in bold.

report the average rank of the correct match within the pool of distractors. The fourth task evaluates semantic clustering on the SciDocs MAG dataset (Cohan et al., 2020) and reports the percentage of each paper's five nearest neighbors that share its thematic category.

B.5.5 Results and Analysis

Results for both the SciRepEval and the semantic benchmark are displayed in Tables 4 and 5, respectively.

Overall, we see that the differences between the different objectives are small: all five models lie within 0.5 points of each other on the SciRepEval average, with InfoNCE at $\tau = 0.2$ (65.80) and spectral (65.68) performing best. The two modifications of InfoNCE (i.e., filtering and SimACE) perform equally to standard InfoNCE at $\tau = 0.07$ across nearly all task types and do not yield a consistent improvement. The main difference between objectives therefore is the type of task they perform particularly well in.

The clearest difference can be seen in tasks that depend on the cluster structure of the embedding space, with these being the *clustering* task of the semantic benchmark and the *classification* tasks in SciRepEval. The latter benefit from a good clustering because they are evaluated by training a linear classifier on top of the predicted embeddings, so that good linear separability is the key requirement for good classification performance. The best scores for both of these task types are achieved by spectral, with a classification average on SciRepEval of 66.41 and a score of 0.739 for semantic clustering. Interestingly, InfoNCE with the standard temperature of 0.07 falls substantially short of spectral for these tasks, while using a temperature of 0.2 closes the gap substantially.

We explain these results by analyzing how strongly each objective pushes certain negative samples away from the anchor. For a given anchor a and a negative sample x_j with cosine similarity s_j, we define the *repulsion force* as the gradient magnitude $g(s_j) = |\partial \mathcal{L} / \partial s_j|$, meaning that a larger gradient corresponds to a stronger push. Since g generally depends on batch-specific normalization terms, we analyze the *relative* repulsion inflicted on a negative s_H with high similarity to the anchor compared to an easy negative s_E with low similarity by defining the *relative repulsion ratio*:

$$R(s_H, s_E) = \frac{g(s_H)}{g(s_E)},$$

This ratio therefore quantifies how much more strongly the loss pushes away a hard negative compared to an easy one. Analyzing this ratio for both InfoNCE and spectral results in the following proposition:

Proposition 1 (Relative Repulsion Ratios)**.** *Let $0 < s_E < s_H$ be the cosine similarities of an easy and a hard negative sample to the same anchor. Under the InfoNCE loss with temperature τ, the relative repulsion ratio is*

$$R_{\textit{InfoNCE}}(s_H, s_E) = e^{(s_H - s_E)/\tau}.$$

Under the spectral loss, writing the dot product as $\sigma_a \sigma_j s_j$ with embedding norms σ, the relative repulsion ratio is

$$R_{\textit{spectral}}(s_H, s_E) = \left(\frac{\sigma_H}{\sigma_E}\right)^2 \frac{s_H}{s_E}.$$

The proof of this proposition is given in Appendix A.5. We note that Wang & Liu (2021) already performed this analysis for InfoNCE, terming its stronger repulsion force on hard negatives as the *hardness-aware property*.

This difference in relative repulsion now likely has an effect on clustering since true labels are unknown during unsupervised training, so that hard negatives frequently belong to the same semantic cluster as the anchor. The more emphasis a loss then puts on pushing away these hard negatives, the more likely it is that clusters of semantically related samples are shattered, thus reducing clustering and classification performance.

To see the strength of this effect, consider a hard negative with $s_H = 0.6$ and an easy negative with $s_E = 0.1$. InfoNCE at $\tau = 0.07$ repels the hard negative approximately $1{,}265$ times more strongly than the easy one, whereas this factor drops to approximately 12 at $\tau = 0.2$ and to 6 for spectral if both embeddings have comparable magnitudes. The specific repulsion properties of the spectral loss therefore enable good clustering and classification performance, but adapting the temperature of InfoNCE can lead to comparable behavior and thus similar results.

The few-shot settings of SciRepEval provide further support for this explanation. Three of the classification tasks additionally evaluate the embeddings with only a handful of labeled training examples per class, and averaged over just these few-shot settings, spectral reaches 30.46 while InfoNCE at $\tau = 0.2$ reaches 29.75 and InfoNCE at $\tau = 0.07$ reaches only 24.42. In contrast, when the full training data is available, all objectives perform much more similarly (all in the range of 74.91 to 75.74). This exact behavior is expected if the objectives differ mainly in how well they preserve clusters: with sufficient labeled data, a linear classifier can still separate classes whose members are somewhat more distributed across the embedding space, whereas with only a few (randomly selected) labeled examples, it relies on members of the same class forming compact groups.

The individual task results also show a similar story. While InfoNCE at $\tau = 0.2$ and spectral perform best on the SciDocs proximity tasks, where relevant papers have to be ranked above random candidates, the low-temperature variants perform better when only subtle differences between related documents determine the correct result. On the *Highly Influential Citations* task, which ranks the citations of a paper by their influence, they reach 43.11 to 43.34 compared to 40.8 for spectral and InfoNCE at $\tau = 0.2$. They also achieve the highest precision at five under the strict criterion of the *Paper-Reviewer Matching* task (34.02 to 34.39 vs. 32.71 for spectral) and best match titles to their exact abstracts in the semantic benchmark. This suggests that the strong repulsion of hard negatives described by Proposition 1 is beneficial whenever the single best match has to be identified within a set of similar candidates.

B.5.6 Ablation

Throughout this work, we use a specialized architecture for the spectral loss that first predicts a unit-normalized directional embedding and then applies an additional linear projection followed by

a softplus to independently predict the vector norm. Dot products are then recovered by multiplying the cosine similarity of two directional embeddings with the product of their predicted norms.

To assess the importance of this "decoupled" architecture for the spectral loss, we train an additional model identical to the spectral model described above, except that the norm head is removed, and the raw output of the linear projection is used without unit normalization. Thus, the resulting architecture is the naive implementation of the spectral loss that simply predicts an unnormalized embedding and uses the dot product to compute similarities.

The results in Tables 4 and 5 show that removing the norm head substantially degrades performance across all task types, which we attribute to the way the spectral loss interacts with embedding norms. Since the loss operates on unnormalized dot products, both the uniformity and the alignment terms can be reduced in two different ways: by arranging the embedding directions so that related texts point in similar directions, or simply by changing the embedding norms. When the raw output of the encoder is used as the embedding, both mechanisms act on the same vector, so that every gradient update mixes changes to the semantic structure with changes to the scale. Since all embeddings are produced by shared weights, even a change that is only meant to adjust the norm of a single embedding is implemented by modifying these weights, for example by amplifying the output dimensions along which this embedding points. This amplification, however, also applies to all other embeddings, changing the relative weighting of its components, and thus its direction rather than just its norm.

We empirically found this entanglement to make training considerably less stable. The norm head resolves this by allowing changes in embedding norm to be realized by adapting the weights of the norm prediction head. At the same time, the directional component can independently leverage the same gradients to freely adapt the directional structure of the embedding space.

B.6 Determining the Ranking-Optimal Geometry

In practice, the true target kernel K, which constitutes the theoretical optimum of the spectral loss, is of very high rank. As a result, while an embedding model with an equally high dimensionality could exactly realize this target, realistic models constrained to dimensions $d \leq 4096$ can just approximate it. Because such an approximation is therefore inevitable irrespective of whether K or $\log K$ is targeted, this raises the question of whether the linear geometry actually constitutes a superior target under these capacity constraints. To answer this question empirically, we train low-rank fits to the empirical kernel matrix $\widehat{K}$ that explicitly optimize for the preservation of its ranking structure, and then analyze which geometric target these optimal embeddings naturally align with.

B.6.1 Experimental Setup

General Setup. We fit free vectors to the same empirical kernel matrix $\widehat{K}$ that was used in the previous free-vector experiments. We evaluate embedding dimensionalities $d \in \{64, 128, 256, 1946\}$, where $d = 1946$ is chosen as it matches the rank of $\widehat{K}$. To evaluate both the setups underlying InfoNCE and the spectral loss, we do separate runs where similarities are computed either as dot products or as cosine similarities.

Each configuration is optimized from two random initializations using Adam with an initial learning rate of 10^{-2} for the dot-product setup and 10^{-3} for the cosine setup. The learning rate is halved whenever the training loss fails to improve by a relative threshold of 10^{-4} over $1{,}000$ consecutive updates. The loss is always computed over all embeddings (i.e., no mini-batches), thus providing the most informative gradient signal. Training terminates if no progress is made after halving the learning rate four times, or if a maximum budget of $10{,}000$ updates is reached. For each setup, we evaluate the best-performing state achieved during optimization.

Objective Functions We design a loss that optimizes strictly for ranking preservation compared to the target matrix $\widehat{K}$. To this end, we use a triplet margin loss that, for a given query document i, orders all candidate documents according to their reference similarities $\widehat{K}_{ij}$ and strictly penalizes ranking violations in the current prediction. To maintain tractability, we restrict the higher-ranked document of each pair to the top-k neighbors of the query.

Similarity	Loss	k	$d=64$ $A(1)$	$d=64$ $A(0)$	$d=64$ $A(\mathrm{rank})$	$d=128$ $A(1)$	$d=128$ $A(0)$	$d=128$ $A(\mathrm{rank})$	$d=256$ $A(1)$	$d=256$ $A(0)$	$d=256$ $A(\mathrm{rank})$	$d=1946$ $A(1)$	$d=1946$ $A(0)$	$d=1946$ $A(\mathrm{rank})$
Dot product	Absolute	10	0.989	0.960	**0.953**	0.983	0.949	**0.941**	0.974	0.924	**0.914**	0.938	0.897	**0.890**
		50	0.994	0.893	**0.867**	0.990	0.787	**0.737**	0.987	0.694	**0.629**	0.983	0.572	**0.488**
		200	0.991	0.441	**0.352**	0.991	0.555	**0.482**	0.991	0.602	**0.521**	0.988	0.394	**0.297**
	Squared	10	0.990	0.965	**0.958**	0.984	0.959	**0.953**	0.977	0.942	**0.935**	0.939	0.930	**0.929**
		50	0.994	0.898	**0.872**	0.990	0.813	**0.767**	0.988	0.728	**0.668**	0.983	0.575	**0.497**
		200	0.994	0.413	**0.322**	0.991	0.550	**0.479**	0.991	0.603	**0.522**	0.987	0.413	**0.313**
Cosine	Absolute	10	0.995	0.981	**0.978**	0.994	0.988	**0.987**	0.991	0.988	**0.987**	**0.961**	0.977	0.979
		50	0.995	0.814	**0.772**	0.992	0.753	**0.700**	0.989	0.692	**0.627**	0.983	0.594	**0.513**
		200	0.996	0.351	**0.275**	0.994	0.419	**0.344**	0.991	0.594	**0.513**	0.987	0.446	**0.348**
	Squared	10	0.997	0.990	**0.989**	0.997	**0.995**	**0.995**	**0.997**	**0.997**	**0.997**	**0.980**	0.991	0.992
		50	0.995	0.820	**0.779**	0.995	0.908	**0.885**	0.995	0.921	**0.903**	0.984	0.895	**0.876**
		200	0.996	0.349	**0.272**	0.994	0.395	**0.322**	0.992	0.676	**0.604**	0.991	0.591	**0.493**

Table 6: Alignment of ranking-optimal free-vector fits with the linear ($A(1)$), logarithmic ($A(0)$), and rank ($A(\mathrm{rank})$) geometry of $\widehat{K}$. The best-aligned geometry per setting is shown in bold.

Specifically, for every document a among the top-k neighbors of query i, and every document b ranked strictly below a according to the target $\widehat{K}$, we define the pairwise violation as:

$$\ell_{iab} = \max\big(0,\ m - (s_{ia} - s_{ib})\big). \tag{50}$$

Here, s denotes the predicted similarity (either dot product or cosine) and m is a scalar margin. The total loss is computed as the mean over all such triplets using either the absolute violation ℓ_{iab} or the squared violation ℓ_{iab}^2, thereby penalizing misranked pairs with either constant or increasing pressure, respectively. We fit embeddings for $k \in \{10, 50, 200\}$ to evaluate the effect of varying degrees of ranking that have to be preserved.

Importantly, 1) $\widehat{K}$ and $\log \widehat{K}$ have the identical similarity ranking, and 2) the loss function optimizes exclusively with regard to this exact ordering. In combination, these factors make the optimization objective entirely independent of any specific target geometry. As a result, the geometric structure ultimately assumed by the trained embeddings reveals the preferred geometry for maintaining optimal rankings, without being biased towards any specific target.

Evaluation. For the fitted vectors, we compute $A(1)$, $A(0)$, and D against $\widehat{K}$ as in the previous experiments to identify the targeted embedding geometry. In addition, we introduce the *rank geometry* $A(\mathrm{rank})$, which is computed like $A(1)$ but with the entries of $\widehat{K}$ replaced by their ranks among the evaluated pairs. We evaluate this geometry since it is the one that a margin-based objective favors when all constraints can be satisfied.

B.6.2 Results

Results are displayed in Table 6. Considering first the case where the optimization is restricted to just the top $k = 10$ neighbors of each query, we see that the objective imposes only minimal constraints on the embeddings, so that none of the evaluated geometries show a particularly low error. However, for more extensive ranking constraints of $k \in \{50, 200\}$, the embeddings clearly prefer the logarithmic geometry over the linear one, producing positive geometric contrasts D ranging from 0.07 to 0.65. Additionally, we also see that in absolute terms the fits are never well described by the linear geometry, indicated by the consistently high errors of $A(1) \geq 0.93$ across all settings.

As expected, the pure rank geometry describes these fits most accurately across almost all settings. But crucially, the logarithmic geometry approximates the optimal, evenly spaced structure of the ranking geometry much more closely than the highly skewed linear geometry does. Ultimately, because this strong preference for the logarithmic structure compared to the linear target holds consistently across both similarity functions, both loss formulations, and all evaluated dimensionalities, it points to a fundamental geometric principle. When an embedding model of limited capacity must approximate a kernel of substantially higher rank, a logarithmic fit is inherently preferable to a linear one for successfully preserving the similarity ranking of the most similar pairs.