# Principles of Concept Representation in Sentence Encoders

**Anonymous ACL submission**

## Abstract

What makes a sentence encoder produce good concept representations? We approach this through the lens of representational compositionality: an encoder supports a concept family only when its latent space admits a low-distortion realization of the corresponding semantic operator. This framing predicts both where current encoders succeed and where they are structurally mismatched to their supervision. Through a controlled ablation over encoder conditions trained on 3.3 million synonym and definition pairs from WordNet and Wiktionary, evaluated on three decontaminated splits and a modifier-labeled noun-phrase benchmark, we identify four principles. Fine-tuning recalibrates the latent geometry rather than expanding it (P1). Semantic signal concentrates in the final transformer layer before concept-specific training begins, making cross-layer pooling redundant (P2). Hard negatives improve discrimination and stress-test robustness without improving retrieval ranking, showing that calibration and ranking are independently addressable (P3). Finally, the effectiveness of supervision depends on the composition type of the target concept. Extensional training helps intersective and subsective families while degrading relational and intensional ones, exposing a structural limitation of current training paradigms (P4). We release two new evaluation datasets: a DBpedia semantic-gap benchmark and a modifier-labeled NP paraphrase suite.

## 1 Introduction

Semantic compositionality is the principle in linguistics and philosophy that the meaning of a complex expression arises from the meanings of its individual components together with the way those components are structured and combined(Frege, 1892). In order for encoders to produce faithful concept representations, they must capture conceptual compositionality. Different modifier types such as intersective, subsective, relational, modal, and privative, contribute meaning through fundamentally different semantic operators (Carvalho et al., 2025), yet a sentence encoder must realize all of them inside a single latent geometry scored by one similarity function. Our theoretical lens is that conceptual compositionality in encoders is an approximate homomorphism problem (formalised in Appendix D). An encoder yields good concept representations only when the typed semantic operators required by a concept family admit low-distortion geometric realization. The empirical question this paper asks and answers is which parts of that structure current sentence encoders already support, and which parts remain mismatched to the supervision used to train them.

We study this through concept retrieval, which makes representational quality of concept compositions directly measurable. The encoder must map a query to the same region as any denotationally equivalent target, regardless of surface form. Complex nominals pose a hard challenge. Intersective, relational, and privative modifiers each realize meaning through different operators, yet all must coexist in a single latent geometry scored by one similarity function. This is a fundamental compositional tension that is illustrated geometrically in Figure 1. Evidence that this tension remains an open problem comes from our negation stress test, where the best frozen baseline scores 0.470 ROC-AUC (below chance), assigning higher similarity to negated definitions than to correct ones.

Several design dimensions are natural candidates for closing this gap. Fine-tuning on concept-equivalence pairs may reshape the latent space toward the right geometry. At the readout level, probing studies establish that upper transformer layers encode more semantic information than lower ones (Peters et al., 2018; Jawahar et al., 2019; Tenney et al., 2019), motivating cross-layer pooling as a candidate improvement. Hard negative supervi-

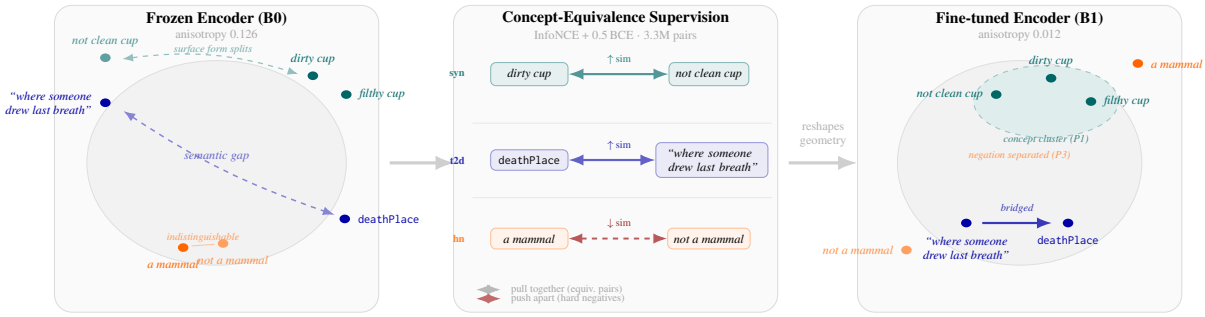


Figure 1: Concept-equivalence fine-tuning, illustrated. **B0:** a frozen encoder splits synonymous paraphrases (teal) by surface form, conflates negations (orange), and leaves NL queries far from structured targets (blue). **Centre:** synonym (*syn*) and term-definition (*t2d*) pairs are pulled together via InfoNCE; hard negatives (*hn*) are pushed apart via BCE. **B1:** the space is recalibrated—paraphrases cluster (P1), the semantic gap bridges, and negation separates (P3).

sion may further sharpen discrimination between near-miss distractors. And theoretically, there is no formal account of when a sentence encoder can support a given modifier composition family, nor of the geometric conditions under which current training objectives succeed or fail.

We structure the empirical investigation around three hypotheses. **H1** *(fine-tuning is necessary):* concept-equivalence fine-tuning substantially improves complex nominal retrieval over frozen baselines. **H2** *(cross-layer pooling):* weighted or input-adaptive mixtures over multiple layers improve over fine-tuned mean pooling. **H3** *(hard negatives):* hard negative supervision improves retrieval ranking in addition to calibration.

**Contributions.** **(T) $\varepsilon_\tau$-compositionality framework:** We introduce a formal characterisation of when a sentence encoder supports a given modifier composition family: $f_\theta$ is $\varepsilon_\tau$-compositional if a low-distortion latent operator $\Phi_\tau$ exists for modifier type $\tau$. This identifies two interacting bottlenecks (representational and objective) and predicts the modifier-family pattern (P4) from first principles. Formal bounds are derived in Appendix D. We additionally identify four empirical principles of compositional concept representation. **(P1) Recalibration, not expansion:** concept-equivalence fine-tuning reshapes the latent geometry, collapsing anisotropy from 0.126 to 0.012 and improving term-to-definition Recall@10 from 0.552 to 0.654, while leaving effective rank unchanged. Fine-tuning recalibrates which regions of the space collapse together, without growing the space. **(P2) Final-layer concentration precedes fine-tuning:** sentence-level pre-training already concentrates semantic signal into the final transformer layer, explaining why cross-layer pooling offers no consistent benefit after concept-equivalence fine-tuning. Confirmed across both NP-paraphrase and term-to-definition tasks. **(P3) Calibration and ranking are dissociable:** hard negatives improve discrimination (ROC-AUC +0.19–+0.46) without improving retrieval ranking, establishing that calibration and ranking are separable training targets. **(P4) Supervision must match composition type:** concept-equivalence training improves intersective and subsective families while degrading relational and intensional types, exposing a fundamental mismatch between equivalence-only supervision and typed semantic operators. We also release two new evaluation datasets: a DBpedia semantic-gap benchmark (3k train / 250 test, zero lexical overlap) and a modifier-labeled NP paraphrase suite (4,000 pairs across five composition families).

## 2 Related Work

**Dense retrieval and contrastive learning.** Sentence-BERT (Reimers and Gurevych, 2019) establishes fine-tuned bi-encoders as practical dense retrievers, and SimCSE (Gao et al., 2021) shows that in-batch contrastive objectives with hard negatives improve representation geometry. We extend this interface to concept-equivalence retrieval and find that semantic signal dominates the final layer in sentence-fine-tuned encoders, limiting the benefit of cross-layer pooling.

**Biomedical concept normalisation.** BioSyn (Sung et al., 2020), SapBERT (Liu et al., 2021), and BioLORD (Remy et al., 2022) demonstrate that synonym-marginalization and definition-aware contrastive training substantially improve biomedical entity retrieval, and Tutubalina et al. (2020) show

that reported accuracy depends heavily on split design. Training on dictionary definitions as concept supervision has prior support (Hill et al., 2016; Carvalho et al., 2023); we use WordNet/Wiktionary synonym and definition pairs in the same vein, but centre our analysis on the mechanism of fine-tuning and introduce controlled modifier-type evaluation absent from prior work.

**Distributional composition and modifier sensitivity.** Compositional distributional semantics (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010) shows that typed composition operators outperform uniform ones and that adjective denotation depends on semantic role—directly motivating the $\varepsilon_\tau$-compositionality framework in §3. The modifier typology underlying our benchmark (intersective, subsective, relational, modal, privative) originates in formal semantics (Partee, 1995), and Ettinger et al. (2018), Shwartz (2019), and Carvalho et al. (2025) show that these distinctions remain problematic for contemporary encoders. Our NP paraphrase benchmark operationalises these concerns as a retrieval task with explicit modifier-family labels, to our knowledge the first to stratify retrieval performance across Montague modifier families.

**Layer distribution and geometry.** Probing studies establish that lower layers of pretrained transformers encode syntax while upper layers encode semantics (Peters et al., 2018; Jawahar et al., 2019; Tenney et al., 2019; Rogers et al., 2020), and Ethayarajh (2019) show this hierarchy is reflected in anisotropy: upper layers are geometrically more uniform and task-ready. We find this distribution collapses in sentence-fine-tuned encoders: prior contrastive training has already concentrated semantic signal into the final layer, and concept-equivalence fine-tuning sharpens it further, leaving nothing for cross-layer readout to exploit. While hyperbolic geometries (Nickel and Kiela, 2017; Valentino et al., 2024) offer theoretically richer hierarchical structure, our results confirm that geometry choice is secondary to training supervision once the space is well-calibrated.

# 3 Representational Compositionality

## 3.1 Problem Setting

Let $\mathcal{X}$ be a space of texts (terms, noun phrases, definitions, ontology labels). We learn an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and a retrieval score $\mathrm{score}_\theta(q, y) \in \mathbb{R}$. Given a query $q$ and a candidate pool $\mathcal{D} = \{y_1, \ldots, y_N\}$, the objective is to rank semantically equivalent candidates first:

$$\mathrm{score}_\theta(q, y^+) > \mathrm{score}_\theta(q, y^-) \quad \text{whenever } [\![q]\!] = [\![y^+]\!] \neq [\![y^-]\!],$$

where $[\![\cdot]\!]$ denotes the concept denotation. In practice, positives are synonym pairs, term-definition pairs, and cross-source definition pairs from the same concept. The hard case is a zero-overlap triplet where the correct answer $y^+$ shares no surface tokens with $q$.

## 3.2 From Semantic Composition to Representational Composition

Semantic compositionality alone does not guarantee compositional representations. Let $f_\theta : \mathcal{X} \to \mathbb{R}^d$ be a sentence encoder. We say that $f_\theta$ is $\varepsilon_\tau$-compositional for modifier family $\tau$ if there exists a latent operator

$$\Phi_\tau : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$$

such that for all valid modifier–head pairs $(m, h)$ of type $\tau$,

$$\| f_\theta(m \circ_\tau h) - \Phi_\tau(f_\theta(m), f_\theta(h)) \| \leq \varepsilon_\tau.$$

This definition makes explicit the bridge between conceptual and geometric compositionality. Semantic composition is typed, but the encoder must realize all such operators in a shared latent space. Concept retrieval then depends on low distortion of the relevant $\Phi_\tau$, and a scoring function that ranks denotationally equivalent expressions above non-equivalent distractors.

Appendix D formalizes these definitions and derives theoretical bounds on retrieval distortion for each composition family.

## 3.3 Why Pooled Embeddings Struggle

Following Montague semantics (Carvalho et al., 2025), the denotation of a complex nominal is:

$$[\![m \circ_\tau h]\!] = C_\tau([\![m]\!], [\![h]\!]),$$

where $\tau$ is the modifier composition type and $C_\tau$ is a type-specific operator. Different modifiers types instantiate materially different $C_\tau$. These five types constitute the standard complete typology of modifier composition under Montague generative semantics (Carvalho et al., 2025), spanning the full range from extensional set-intersection (intersective) to intensional non-instantiation (privative). A

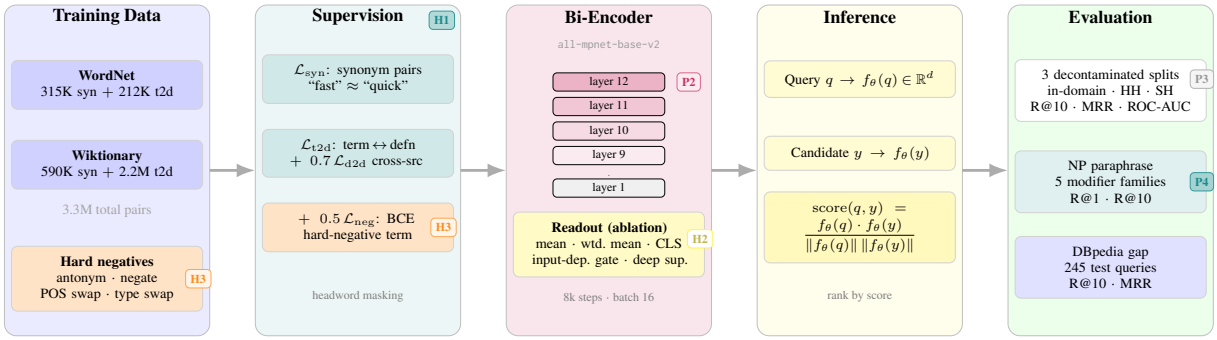


Figure 2: End-to-end pipeline. WordNet/Wiktionary synonym and definition pairs (3.3M) are combined with five hard-negative rule types in a joint InfoNCE + BCE objective, training a bi-encoder (`all-mpnet-base-v2`) under eight ablation conditions varying readout strategy and hard-negative supervision. Evaluation covers three decontaminated term-definition splits (in-domain, HH, SH), modifier-sensitive NP paraphrase retrieval, and zero-lexical-overlap DBpedia property retrieval. Coloured badges indicate which component each hypothesis and principle targets.

pooled vector $z(x) = g(H^{(1)}, \ldots, H^{(L)})$ cannot expose the type $\tau$ explicitly, but encodes it implicitly in the geometry.

This creates two interacting bottlenecks. The **representation bottleneck**: structured token-role information is collapsed into one vector, and multiple semantic relations must coexist in one shared neighborhood structure. The **objective bottleneck**: standard sentence embedding training rewards broad distributional semantic similarity, not operator-sensitive conceptual precision. Our study isolates and investigates these two aspects. Through contrastive fine-tuning on synonym and definition pairs directly, we target the objective bottleneck, and a cross-layer readout ablation tests whether the representation bottleneck matters after fine-tuning.

## 4 Proposed Method

### 4.1 Model and Conditions

The models used in our ablations are based on the bi-encoder sentence-transformer architecture (Reimers and Gurevych, 2019) where both query and candidate texts are encoded independently, and similarity is computed via cosine distance. We define readout (how a single vector is extracted from the transformer's hidden states) as the primary design variable. Table 1 defines the eight conditions evaluated.

The full eight-condition ablation is conducted on a single backbone, `all-mpnet-base-v2`, to isolate the contribution of each readout and supervision design choice without conflating backbone effects. To test whether the layerwise structure and the B0→B1 gain generalise beyond this choice, we additionally train B0 and B1 on three further backbones: `paraphrase-mpnet-base-v2` (same MPNet architecture, paraphrase-oriented pre-training objective), `e5-base-v2` (a different model family with strong general-purpose text embeddings), and `jina-embeddings-v5-text-nano` (PEFT-based, trained with a distinct contrastive objective outside the standard sentence-transformer paradigm). Together, these four backbones provide same-architecture/different-objective variation (`all-mpnet` vs. `para-mpnet`), cross-family generalisation (`e5-base`), and a deliberate architectural contrast case (`jina-nano`).

| ID | Pooling | HN | Note |
|---|---|---|---|
| B0 | Mean (frozen) | | Frozen baseline |
| B0-WM | Wtd. mean (frozen) | | Frozen wtd. mean |
| B1 | Mean | ✓ | Fine-tuned ref. |
| B2 | CLS token | ✓ | |
| B3 | Weighted mean | | Pos. only |
| M1 | Weighted mean | ✓ | |
| M2 | Input-dep. gate | ✓ | MLP-gated |
| M3 | Deep supervision | ✓ | Per-layer heads |

Table 1: Encoder conditions (shared backbone: `all-mpnet-base-v2`, 8k steps, batch 16, single seed). B0/B0-WM are frozen and differ only in pooling, isolating cross-layer averaging without fine-tuning. Weighted mean pools the last $K$=4 layers; M2 uses a per-input MLP gate; M3 attaches per-layer projection heads during training.

### 4.2 Cross-Layer Readout

Let $H^{(l)}(x) \in \mathbb{R}^{T \times d}$ be the token hidden states at layer $l$ for an input of length $T$. The B1 baseline uses standard mean pooling over the final layer only.

The cross-layer variants instead form a weighted mixture over the last $K{=}4$ layers before pooling:

$$z(x) = P\left(\sum_{l=L-K+1}^{L} \alpha_l \, H^{(l)}(x)\right),$$
$$\alpha = \mathrm{softmax}(w),$$

where $P(\cdot)$ denotes mean pooling, $w \in \mathbb{R}^K$ are learned scalar weights shared across all inputs (B3, M1), and $z(x)$ is L2-normalised before scoring.

**M2 (input-dependent gating).** The global weights $w$ are replaced by a per-input MLP:

$$\alpha(x) = \mathrm{softmax}\Big(\mathrm{MLP}\Big(P(H^{(L)}(x))\Big)\Big),$$

so different inputs can weight layers differently.

**M3 (deep supervision).** During training, a separate linear projection head $\phi_l$ is attached to each of the last $K$ layers, and an InfoNCE loss is computed at every layer independently:

$$\mathcal{L}_{\text{deep}} = \sum_{l=L-K+1}^{L} \mathcal{L}_{\text{InfoNCE}}\Big(\phi_l\Big(P(H^{(l)})\Big)\Big).$$

At inference the projection heads are discarded and the standard weighted-mean readout is used, forcing all layers to develop useful representations during training rather than letting the gradient concentrate at the final layer.

### 4.3 Training Objective

The positive loss is a weighted InfoNCE sum over three supervision views:

$$\mathcal{L}_{\text{pos}} = \mathcal{L}_{\text{syn}} + \mathcal{L}_{\text{t2d}} + 0.7\,\mathcal{L}_{\text{d2d}},$$

where $\mathcal{L}_{\text{syn}}$, $\mathcal{L}_{\text{t2d}}$, and $\mathcal{L}_{\text{d2d}}$ are InfoNCE losses over synonym pairs, term-definition pairs, and cross-source definition-definition pairs respectively. The 0.7 weight on d2d reflects its noisier cross-source alignment signal.

When hard negatives are used (all conditions except B0 and B3), a binary cross-entropy term targets five rule-based negative types per definition (see Appendix G for full rule definitions and examples). The complete objective is:

$$\mathcal{L} = \mathcal{L}_{\text{pos}} + 0.5\,\mathcal{L}_{\text{neg}},$$

where the 0.5 coefficient moderates the contribution of the hard-negative term relative to the positive InfoNCE objective. Unlike the batch-level InfoNCE loss, the BCE term applies a gradient of fixed magnitude to each negative pair independently of batch size, without down-weighting, its aggregate contribution would scale with the number of negative pairs and risk overshadowing the ranking objective. The coefficient was not ablated, however Appendix F provides evidence of its importance by evaluating a variant (B1-unified) in which hard negatives are folded into the InfoNCE denominator, eliminating the need for an explicit weight at the cost of a substantial reduction in calibration.

### 4.4 Training Details

Full hyperparameters are listed in Appendix E. All experiments use AdamW with linear warmup (100 steps) and linear decay to zero. A layerwise LR decay of 0.90 per transformer layer mitigates catastrophic forgetting (Howard and Ruder, 2018); the pooling head and cross-layer weights receive $5\times$ the base rate.

### 4.5 Data

**Training.** 3.3 million synonym and term-to-definition pairs from **WordNet** (527K: 315K syn + 212K t2d) and **Wiktionary** (2.8M: 590K syn + 2.2M t2d). Cross-source definition–definition pairs (d2d) are generated on-the-fly by grouping definitions of the same term across both sources (up to 6 per term; not counted in the 3.3M total). A headword-masking transform removes the defined term from each definition before training to prevent lexical shortcutting. Representative examples of all three pair types are given in Appendix B.

**Evaluation splits.** Three decontaminated splits, each with test concepts strictly absent from training. **In-domain** (t2d / d2t): held-out WordNet/Wiktionary pairs partitioned at the synset level. **Head-holdout** (HH): the most frequent synsets withheld entirely from training, testing generalisation to high-frequency unseen concepts. **Source-holdout** (SH): trained on Wiktionary, evaluated on WordNet, testing cross-source generalisation. Each split includes hard-negative stress tests (Appendix G).

**DBpedia semantic-gap benchmark.** A new evaluation dataset of (query, property label) pairs over the DBpedia ontology, consisting of 3,063 training queries generated by GPT-4o-mini with zero lexical overlap with property labels, and 245 manually reviewed test pairs with strictly disjoint train/test properties.

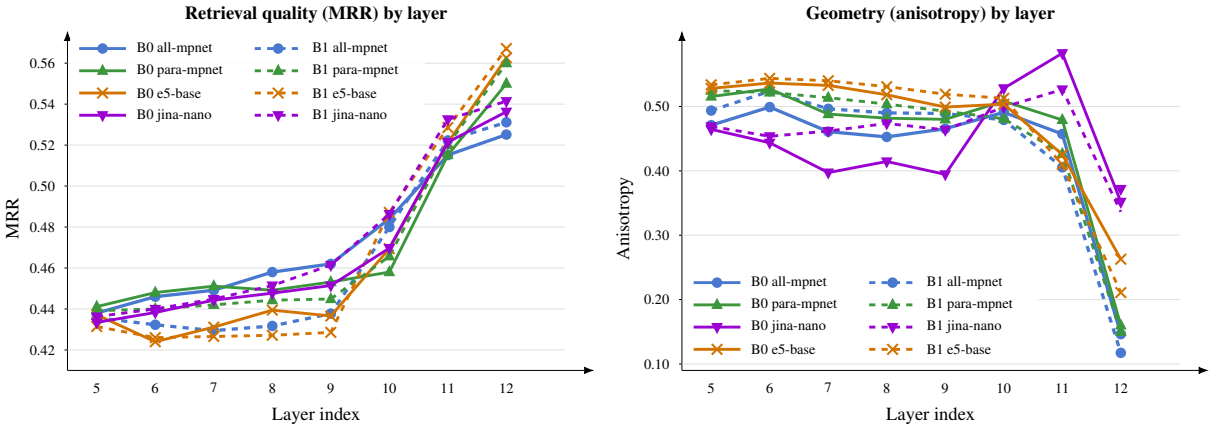


Figure 3: Per-layer MRR (left) and anisotropy (right) for transformer layers 5–12 on NP paraphrase pairs (solid = B0 frozen, dashed = B1 fine-tuned). Final-layer MRR dominance pre-exists fine-tuning across all tested backbones: layer-12 MRR far exceeds earlier layers, with anisotropy dropping sharply at the final layer. Fine-tuning sharpens both effects without redistributing signal to earlier layers, leaving nothing for cross-layer readout to exploit.

**NP paraphrase benchmark.** A new modifier-sensitive evaluation suite of 4,000 noun-phrase paraphrase pairs balanced across five composition families, namely intersective (900), subsective (800), modal (800), privative (800), relational (700). These cover the complete Montague modifier typology (Section 3). Pairs are drawn from two equivalence regimes of 2,000 each, to create **strict** (GPT-4o-mini-generated paraphrases with zero lexical overlap enforced after head masking) and **near** (rule-based head-broadening or relation-abstraction transforms applied to the strict pairs).

## 5 Empirical Analysis

### 5.1 Layer-12 Dominance: The Mechanism Behind H2

We present the layerwise analysis before the H1 ablation to establish the mechanistic explanation that directly predicts the H2 null result and motivates the $K{=}4$ window choice in the cross-layer conditions. B1 results appear here alongside B0 for contrast; that B1 represents a substantial retrieval improvement over B0 is confirmed in §5.2.

Figure 3 shows that the final layer dominates in both B0 and B1. MRR at layer 12 far exceeds layers 5–7, while anisotropy drops sharply at the last two layers. Fine-tuning sharpens this concentration, in that earlier layers see lower MRR and higher anisotropy after training, but training does not redistribute information across layers. The pattern holds across all tested sentence-transformer backbones.

Probing studies (Jawahar et al., 2019; Tenney et al., 2019) established the layer hierarchy on raw pretrained models, where signal *is* distributed and cross-layer readout could plausibly help. In an already-sentence-fine-tuned encoder, that distribution collapses before our fine-tuning even begins, and concept-equivalence training sharpens it further. There is no distributed multi-layer signal for a cross-layer readout to exploit. The H2 ablation (Table 3, §5.3) confirms this directly.

### 5.2 H1: Fine-Tuning Substantially Improves Concept Retrieval

Table 2 shows that B0 (frozen `all-mpnet-base-v2` with mean pooling) achieves t2d R@10 = 0.552 and negate-stress ROC-AUC = 0.470 (below chance). We highlight this failure mode in the introduction, where the frozen encoder assigns *higher* similarity to negated definitions than to correct ones. Concept-equivalence fine-tuning resolves both, as B1 reaches t2d R@10 = 0.654 and negate ROC-AUC = 0.980, confirming that supervision is necessary and that it specifically addresses the negation failure. However, fine-tuning is not uniformly beneficial, given that B0 already achieves NP paraphrase R@10 = 0.704, exceeding all fine-tuned conditions (B1: 0.656, B3: 0.655). The backbone's prior sentence-level training is well-suited to paraphrase retrieval. Concept-equivalence fine-tuning improves cross-modal concept matching precisely where the backbone is weak, at the cost of paraphrase-level similarity where it was already well-calibrated.

The B3 vs. M1 comparison reveals that B3

| ID | Condition | t2d R@10 | t2d MRR | Pair ROC-AUC | HH d2t R@10 | Negate | NP R@10 |
|---|---|---|---|---|---|---|---|
| B0 | Frozen baseline (mean) | 0.552 | 0.432 | 0.718 | 0.546 | 0.470 | 0.704 |
| B0-WM | Frozen baseline (wtd. mean) | 0.392 | 0.303 | 0.690 | 0.438 | 0.461 | 0.675 |
| B1 | Mean pool + hard neg. | 0.654 | 0.537 | **0.963** | **0.686** | 0.980 | **0.656** |
| B3 | Weighted mean, no hard neg. | **0.657** | **0.544** | 0.769 | 0.677 | 0.528 | 0.655 |
| M1 | Weighted mean + hard neg. | 0.651 | 0.532 | 0.961 | 0.673 | **0.984** | 0.623 |

Table 2: H1 ablation results. t2d = term-to-definition (in-domain); HH = head-holdout; Negate = negation stress-test ROC-AUC; NP R@10 = noun-phrase paraphrase ($n$=4,000). Bold: best fine-tuned condition per column.

matches M1 on retrieval ranking (0.657 vs. 0.651 t2d R@10) but is far weaker on calibration (pair ROC-AUC 0.769 vs. 0.961) and semantic stress tests (negate ROC-AUC 0.528 vs. 0.984). Hard negatives train the model to *discriminate*, not to *rank*. Whether hard negatives are necessary depends entirely on the downstream use case.

Fine-tuning collapses anisotropy from 0.126 to 0.012 while leaving effective rank essentially unchanged (247.0 $\rightarrow$ 247.2), confirming that concept-equivalence training reorganises the representation space rather than expanding it. Hard negatives are the dominant factor, given that B3 without hard negatives reaches only 0.166 anisotropy under an otherwise identical regime. Full geometry statistics are in Appendix H.

### 5.3 H2: Cross-Layer Pooling Offers No Consistent Benefit

Table 3 reveals cross-layer pooling to be redundant. No pooling variant consistently outperforms the fine-tuned mean-pool baseline B1. CLS pooling (B2) and the wider K=6 window underperform B1 on all splits. Input-dependent gating (M2) is the only variant that matches or marginally exceeds B1 (+0.005 t2d R@10), and the margin is $<1\%$ absolute. Deep supervision (M3) improves over the static weighted mixture M1 but does not reach B1. This is the expected consequence of the layer-concentration principle (Section 5.1).

### 5.4 H3: Hard Negatives Improve Calibration, Not Ranking

Table 4 shows that hard negatives improve discrimination, not ranking. B3 and M1 provide the cleanest isolation, differing only in whether the BCE hard-negative term is active. B3 achieves t2d R@10 = 0.657 vs. M1's 0.651 ($-0.006$, effectively unchanged), while negate ROC-AUC jumps from 0.528 to 0.984 (+0.456) and type-swap from 0.472 to 0.976 (+0.504). Ranking and calibration are separable properties of the concept representation space, governed by different components of the training objective.

| | In-domain | | HH | |
|---|---|---|---|---|
| | t2d R@10 | d2t R@10 | t2d R@10 | d2t R@10 |
| B1 | 0.654 | 0.661 | 0.673 | **0.686** |
| B2 | 0.650 | 0.657 | 0.666 | 0.679 |
| M1 | 0.651 | 0.655 | 0.665 | 0.673 |
| M2 | **0.659** | **0.664** | **0.676** | **0.688** |
| M3 | 0.649 | 0.654 | 0.667 | 0.674 |

Table 3: H2 retrieval results (R@10). HH = head-holdout; t2d = term-to-definition; d2t = definition-to-term. No cross-layer variant consistently beats B1; M2 matches or marginally exceeds it by $<1\%$ absolute. Bold: best per column.

| | No hard negatives | | Hard negatives | |
|---|---|---|---|---|
| Metric | B0 | B3 | B1 | M1 |
| t2d R@10 | 0.552 | **0.657** | 0.654 | 0.651 |
| Pair ROC-AUC | 0.718 | 0.769 | **0.963** | 0.961 |
| Antonym flip | 0.894 | 0.946 | **0.965** | 0.964 |
| Negate | 0.470 | 0.528 | **0.986** | 0.984 |
| Same-POS (rnd.) | 0.903 | **0.962** | **0.962** | **0.962** |
| Same-POS (pfx.) | 0.856 | 0.919 | 0.926 | **0.928** |
| Type swap | 0.453 | 0.472 | **0.981** | 0.976 |

Table 4: H3 stress-test results (ROC-AUC, in-domain). Upper: retrieval ranking and calibration; lower: per-type stress tests. B3 vs. M1 isolates hard-negative supervision (same pooling and budget). Bold: best per row.

We conclude that hard negatives are warranted when the downstream task requires semantic discrimination. These tasks include pair scoring, similarity thresholding, and robustness to adversarial paraphrase. For ranking-only pipelines, where the goal is Recall@K rather than a calibrated similarity score, B3 matches or exceeds M1, and the additional complexity of rule-based hard-negative generation and BCE weighting is unnecessary.

### 5.5 Modifier-Family Analysis: Extensional Supervision Helps Extensional Families

Table 5 breaks down NP Recall@1 by modifier type, revealing a pattern that mirrors the semantic

| Modifier | B0 | B1 | B3 | M2 |
|---|---|---|---|---|
| Intersective | 0.332 | **0.403** | 0.377 | 0.364 |
| Subsective | 0.229 | 0.279 | 0.275 | **0.284** |
| Relational | **0.356** | 0.194 | 0.271 | 0.181 |
| Modal | 0.194 | 0.196 | 0.188 | **0.201** |
| Privative | 0.065 | 0.070 | **0.077** | **0.077** |
| Overall | 0.235 | 0.234 | **0.240** | 0.226 |

Table 5: NP paraphrase R@1 by modifier family ($n$=4,000 pool). Bold: best per row. ROC-AUC is uniformly high (0.945–0.999) and omitted.

| Model | R@1 | R@10 | MRR |
|---|---|---|---|
| B0 (frozen) | 0.220 | 0.608 | 0.349 |
| B1 | **0.257** | 0.649 | 0.393 |
| M2 | 0.253 | **0.665** | **0.397** |
| *DBpedia 300-step* | *0.376* | *0.845* | *0.550* |

Table 6: Cross-domain transfer to DBpedia (245 queries, 2,849 candidates). Non-italic = WordNet/Wiktionary-only training; italic = 300 in-domain DBpedia steps (upper bound). Bold: best cross-domain.

type hierarchy in Section 3 (see Appendix C for a plain-language description of each family). Extensional families (intersective, subsective) allow the correct paraphrase to be recovered from proximity to the head-noun class, while relational, modal, and privative families introduce implicit arguments, possible-world reference, or head-extension negation that make head-noun neighbours actively misleading. Privative and modal accordingly remain the hardest families (below 0.08 and 0.21 respectively) across all conditions.

B1 shows the largest intersective gain over B0 (+0.071 R@1) but substantially hurts relational recall (−0.162); no fine-tuned condition recovers relational performance. This follows directly from the supervision signal, as synonym and definition pairs encode extensional equivalence, improving families whose $C_\tau$ is extensional (intersective, subsective) while degrading or leaving unchanged those with relational or intensional $C_\tau$. We call this the **supervision–composition matching principle**: a concept embedding benefits from fine-tuning only when the supervision encodes the same semantic composition structure as the target concept class. Closing the gap for relational and intensional families requires type-matched supervision, which remains an open problem.

The aggregate results in Table 5 (mechanistic breakdown in the Appendix I) confirm that extensional supervision improves intersective and subsective families (+0.064 and +0.041 R@1) while degrading relational (−0.162, 311 regressions) and leaving modal and privative largely unchanged, consistent with P4 and the geometric account in Section 3.2.

### 5.6 Cross-Domain Transfer: P1 and P4 Generalise Beyond the Training Domain

Table 6 confirms both principles in a new domain. Generic fine-tuning (B1, trained on WordNet/Wiktionary) transfers partially. DBpedia R@10 improves from 0.608 to 0.649, consistent with P1—the recalibrated geometry clusters concept-equivalent expressions more tightly regardless of domain. But 300 steps of DBpedia-specific supervision reaches R@10 = 0.845, a further +0.196 gain that generic fine-tuning cannot close. This gap is the cross-domain expression of P4, showing that WordNet/Wiktionary encodes lexicographic equivalence between synonyms and dictionary definitions, while DBpedia requires mapping natural-language queries to terse property labels (`deathPlace`, `populationTotal`) whose structure never appears in lexicographic training. We validate that P1 generalises across encoder architectures in Appendix J. Contrastive fine-tuning improves every tested backbone, with larger gains where the frozen baseline is weaker.

## 6 Conclusion

In this paper we investigate and identify four principles that govern concept-equivalent retrieval using sentence encoders, namely the training signal, the readout strategy, the use of hard negatives, and the match between supervision and semantic composition type. These principles offer concrete guidance to practitioners: use a sentence-fine-tuned encoder with mean pooling, add hard negatives only when calibrated scoring matters more than ranking, and match supervision to the semantic composition structure of the target concept class.

## Limitations

**Sentence-encoder backbones.** All H1/H2 ablations use `all-mpnet-base-v2`, which is already sentence-level contrastively trained; the H2 null result may therefore be specific to already-fine-tuned backbones. On a raw pretrained transformer, cross-layer readout may still recover meaningful signal. **Single seed.** All conditions use one training seed. H1 margins are large enough to be ro-

bust; H2 margins (<1% absolute) may not be stable across seeds. **Type-matched supervision.** The supervision–composition matching principle (Section 5) indicates the need for relation-typed and negation-aware supervision pairs, which do not currently exist at scale, so relational, modal, and privative concept types remain systematically underserved. Constructing such resources may be a valuable next step to extend this work.

## References


Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are functions: Experiments with compositional models of meaning. In *Proceedings of EMNLP 2010*, pages 1183–1193.

Danilo S. Carvalho, Edoardo Manino, Julia Rozanova, Lucas Cordeiro, and André Freitas. 2025. Montague semantics and modifier consistency measurement in neural language models. In *Proceedings of COLING 2025*.

Danilo S. Carvalho, Giangiacomo Mercatali, Yingji Zhang, and André Freitas. 2023. Learning disentangled representations for natural language definitions. In *Findings of EACL 2023*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? In *Proceedings of EMNLP 2019*.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of COLING 2018*.

Gottlob Frege. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP 2021*.

Felix Hill, KyungHyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of ACL 2018*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of ACL 2019*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of NAACL 2021*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems 30*.

Barbara H. Partee. 1995. Lexical semantics and compositionality. In Lila Gleitman and Mark Liberman, editors, *An Invitation to Cognitive Science: Language*, volume 1, pages 311–360. MIT Press, Cambridge, MA.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL 2018*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019*.

Franck Remy, Kris Demuynck, and Thomas Demeester. 2022. BioLORD: Learning ontological representations from definitions for biomedical concepts. In *Findings of EMNLP 2022*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Vered Shwartz. 2019. A systematic comparison of English noun compound representations. In *Proceedings of the MWE-WN Workshop, ACL 2019*.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of ACL 2020*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL 2019*.

Elena Tutubalina, Artur Kadurin, and Zulfat Miftahutdinov. 2020. Fair evaluation in concept normalization: a large-scale comparative analysis for BERT-based models. In *Proceedings of COLING 2020*.

Marco Valentino, Danilo Carvalho, and André Freitas. 2024. Multi-relational hyperbolic word embeddings from natural language definitions. In *Proceedings of EACL 2024*.

## A Reproducibility

All backbone models used in this work (`all-mpnet-base-v2`, `paraphrase-mpnet-base-v2`, `e5-base-v2`, `jina-embeddings-v5-text-nano`, and `paraphrase-MiniLM-L6`) are publicly available from the Sentence-Transformers library (Reimers and Gurevych, 2019). Training data are drawn exclusively from WordNet and Wiktionary, both publicly available; the exact pair-generation procedure is described in the Method section and Appendix B. Full hyperparameters (learning rate, batch size, warmup, layerwise decay schedule, pooling window $K$) are listed in Appendix E. The two evaluation benchmarks introduced in this work—the DBpedia semantic-gap benchmark and the modifier-labeled NP paraphrase suite—are released alongside the paper. Training code, configuration files, and model checkpoints for all eight conditions are released publicly.[1]

## B Training Data Examples

Table 7 illustrates the three pair types used during training. Synonym pairs align near-synonymous terms from the same source. Term-to-definition (t2d) pairs align a term with its dictionary definition; the headword is masked from the definition before training to prevent lexical shortcutting. Cross-source definition–definition (d2d) pairs are generated on-the-fly by pairing definitions of the same concept across WordNet and Wiktionary.

## C Modifier Composition Types

The five modifier families used throughout this paper originate in formal semantics (Partee, 1995; Carvalho et al., 2025). Table 8 provides an accessible reference with concrete adjective–noun examples and plain-language descriptions of the underlying meaning mechanism, for readers less familiar with this typology.

## D Theoretical Foundations of Conceptual Compositionality

This appendix formalizes the theoretical perspective used in the main text. It states the structural assumptions under which the observed modifier-family differences, pooling null result, and fine-tuning effects follow naturally.

[1] Code release is withheld during double-blind review to preserve author anonymity.

### D.1 Setup and Notation

Let $(\mathcal{X}, \circ_\tau)$ denote a typed expression algebra, where $\tau$ indexes composition types such as intersective, subsective, relational, modal, and privative modification. Let $[\![\cdot]\!] : \mathcal{X} \to \mathcal{C}$ be a semantic interpretation function into a concept domain $\mathcal{C}$. Let $f : \mathcal{X} \to \mathbb{R}^d$ be a sentence encoder, and let $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the similarity score used for retrieval. We write $x \sim y$ when $[\![x]\!] = [\![y]\!]$.

We use the $\varepsilon_\tau$-compositionality definition from §3: encoder $f$ is $\varepsilon_\tau$-compositional for type $\tau$ if there exists $\Phi_\tau : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ such that $\|f(m \circ_\tau h) - \Phi_\tau(f(m), f(h))\| \leq \varepsilon_\tau$ for all valid modifier–head pairs $(m, h)$ of type $\tau$. The quantity $\varepsilon_\tau$ measures the distortion incurred when a typed semantic operator is realized in a single geometric representation space. Small $\varepsilon_\tau$ indicates that the encoder supports the corresponding composition family; large $\varepsilon_\tau$ indicates an operator–geometry mismatch.

### D.2 Low-Distortion Homomorphism Requirement

**Theorem D.1** (Approximate Homomorphism Requirement)**.** *Suppose retrieval over a composition family $\tau$ is stable under denotational equivalence: for any $x = m \circ_\tau h$ and any $y$ such that $x \sim y$, the representations $f(x)$ and $f(y)$ lie within radius $\delta$ of one another. Then there exists a latent operator $\Phi_\tau$ whose empirical distortion over observed compositions is at most the optimal prediction error from constituent embeddings:*

$$\sup_{(m,h)} \|f(m \circ_\tau h) - \Phi_\tau(f(m), f(h))\| \leq \varepsilon_\tau^*,$$

*where $\varepsilon_\tau^*$ is the minimum achievable reconstruction error over operators on $\mathbb{R}^d \times \mathbb{R}^d$.*

*Proof.* Let $Z_\tau = \{(f(m), f(h), f(m \circ_\tau h))\}$ be the set of observed constituent–composition triples. Define $\Phi_\tau$ as any minimizer of worst-case reconstruction error over $Z_\tau$; existence on finite support is immediate by table-lookup construction, assigning each observed input pair to a centroid of the corresponding target set. It remains to bound $\varepsilon_\tau^*$ using the stability assumption. For each pair $(f(m), f(h))$, the target $f(m \circ_\tau h)$ may vary across surface realizations of the same composition. Stability under equivalence ensures all such realizations lie within a $\delta$-ball of one another: $\|f(x) - f(y)\| \leq \delta$ whenever $[\![x]\!] = [\![y]\!]$. Any

| Type | Source | Left | Right |
|---|---|---|---|
| Synonym | WN | fast | quick |
| | WN | joyful | elated |
| | Wkt | physician | doctor |
| t2d | WN | democracy | a political system in which the supreme power lies in a body of citizens who can elect people to represent them |
| | WN | photosynthesis | the process in green plants by which carbohydrates are synthesised from carbon dioxide and water using light as an energy source |
| | Wkt | pandemic | a disease that affects a very large number of people throughout the world |
| d2d | WN↔Wkt | a hot drink made by infusing the dried crushed leaves of the tea plant in boiling water | an aromatic beverage prepared by steeping cured leaves of the tea plant |
| | WN↔Wkt | a state of complete mental and physical well-being | the state of being comfortable, healthy, or happy |

Table 7: Representative training pair examples for each supervision type. WN = WordNet; Wkt = Wiktionary. Headwords are masked from definitions during training.

| **Type** | **Example** | **Paraphrase** | **Meaning mechanism** |
|---|---|---|---|
| Intersective | *red car* | *a car that is red* | Intersection of sets: the referent is in both the modifier and head extension. |
| Subsective | *large ant* | *an ant that is large for an ant* | Restriction within the head extension; the modifier is relative to the noun class, not absolute. |
| Relational | *electric car* | *a car powered by electricity* | Introduces an implicit relation argument absent from the surface string; meaning depends on context. |
| Modal | *alleged thief* | *someone alleged to be a thief* | References possible worlds; the head denotation is not entailed by the compound. |
| Privative | *toy gun* | *something that looks like a gun but is not a gun* | Negates the head extension: the referent is explicitly excluded from the head class. |

Table 8: The five modifier composition families with adjective–noun examples and plain-language meaning descriptions. Each family requires a structurally different semantic operator $C_\tau$ (Section 3), which is why a single latent geometry cannot support all of them equally well under uniform supervision.

operator that maps each input pair to a representative point of the corresponding $\delta$-ball therefore achieves worst-case reconstruction error at most $\delta$ over the observed support. Hence $\varepsilon_\tau^* \leq \delta$, establishing the bound. Conversely, if no low-distortion $\Phi_\tau$ exists ($\varepsilon_\tau^* \gg \delta$), then the composed representations are not determined by their constituents, and expressions with identical semantic structure but different surface forms scatter to unrelated regions—contradicting retrieval stability. □

**Interpretation.** The theorem makes precise why compositional retrieval is stronger than ordinary sentence similarity: success requires not only clustering equivalent expressions, but also realizing a predictable latent operator for the relevant composition family.

### D.3 Identifiability from Equivalence Supervision

**Theorem D.2** (Supervision Identifiability Boundary)**.** *If training supervision consists only of equivalence constraints of the form $x \sim y \Rightarrow f(x) \approx f(y)$, then latent operators $\Phi_\tau$ are identifiable only up to equivalence-preserving transformations. In particular, relation-typed and intensional parameters not expressed in the equivalence labels are underdetermined.*

*Proof.* Let $[x]_\sim = \{y \in \mathcal{X} : [\![y]\!] = [\![x]\!]\}$ denote the denotational equivalence class of $x$. Equivalence supervision constrains $f$ only through labeled pairs: it enforces $f(x) \approx f(y)$ when $x \sim y$, and $f(x) \not\approx f(z)$ when $x \not\sim z$. This determines the quotient geometry of $f$ over $\sim$—the relative placement of equivalence class centroids—but leaves the internal compositional structure of $f$ underdeter-

mined.

Formally, let $T : \mathbb{R}^d \to \mathbb{R}^d$ be any invertible map satisfying: $T(v)$ lies within the $\delta$-ball of $[x]_\sim$ whenever $v = f(x)$. Such transformations form a non-trivial family; for example, any orthogonal rotation within each class neighborhood qualifies. The transformed encoder $\tilde{f} = T \circ f$ satisfies all equivalence constraints, since $\|\tilde{f}(x) - \tilde{f}(y)\| = \|Tf(x) - Tf(y)\| \leq \delta$ whenever $f(x)$ and $f(y)$ already lie within $\delta$ of one another. For any operator $\Phi_\tau$ compatible with $f$, the conjugated operator

$$\tilde{\Phi}_\tau = T \circ \Phi_\tau \circ (T^{-1} \times T^{-1})$$

satisfies $\tilde{\Phi}_\tau(\tilde{f}(m), \tilde{f}(h)) = T(\Phi_\tau(f(m), f(h)))$, matching $\tilde{f}(m \circ_\tau h)$ whenever $\Phi_\tau$ matched $f(m \circ_\tau h)$. Since $T$ can be chosen to permute the internal geometry arbitrarily within equivalence neighborhoods, the composition operator is identifiable only up to the family of such $T$. In particular, whether a match arises from intersection, relation binding, modal evaluation, or privative exclusion cannot be inferred from equivalence labels alone, unless those distinctions are explicitly represented in the training signal. □

**Interpretation.** This result formalizes the supervision–composition matching principle: synonym and definition pairs identify quotient geometry over denotational equivalence, but they do not identify all typed semantic operators.

### D.4 Head-Preservation Bias

**Theorem D.3** (Geometric Head Bias)**.** *Consider a retrieval geometry in which composed noun phrases are encouraged to remain close to their heads: $s(f(m \circ_\tau h), f(h))$ is high relative to unrelated heads. Then composition families whose denotations preserve head membership admit lower distortion under this geometry than composition families requiring head exclusion or non-local relation binding.*

*Proof sketch.* The argument proceeds by examining the semantic relationship between composition and head extension for each family type. For intersective modification, $[\![m \circ h]\!] = [\![m]\!] \cap [\![h]\!] \subseteq [\![h]\!]$: every instance of the composition is also an instance of the head. Geometric proximity $s(f(m \circ h), f(h)) \geq \alpha$ is therefore semantically coherent—the representation should lie in the same region as the head. For subsective modification, the composition is head-preserving in extension ($[\![m \circ h]\!] \subseteq [\![h]\!]$), so the same argument applies. For relational modification, $[\![m \circ h]\!]$ depends on an implicit relation argument $R$ not encoded in either $f(m)$ or $f(h)$ alone; head proximity provides no information about $R$, so the prediction from constituent embeddings incurs additional residual. For modal modification, the composition introduces an evidential variable indicating uncertainty about head membership; proximity to $f(h)$ conflates semantically certain and uncertain instances. For privative modification, $[\![m \circ_{\text{priv}} h]\!] \cap [\![h]\!] = \emptyset$: proximity to $f(h)$ is maximally misleading. The distortion $\varepsilon_\tau$ therefore increases monotonically across the hierarchy intersective, subsective, relational, modal, privative, following the operator–geometry alignment. □

**Interpretation.** This theorem explains why the modifier hierarchy in the experiments is not merely empirical: the ranking follows from how well each operator family aligns with head-centered similarity.

### D.5 Privative Incompatibility under a Single Global Metric

**Theorem D.4** (Privative Metric Conflict)**.** *Assume the encoder is consistent with respect to head instances: for all genuine instances $y \in [\![h]\!]$, $s(f(y), f(h)) \geq \alpha - \eta$ for small $\eta > 0$ (instances cluster near their category head). Then no single global metric can simultaneously satisfy, for all privative constructions $m \circ_{\text{priv}} h$, the following three constraints: (i) $s(f(m \circ_{\text{priv}} h),\ f(h)) \geq \alpha$ (high proximity to lexical head); (ii) $[\![m \circ_{\text{priv}} h]\!] \cap [\![h]\!] = \emptyset$ (semantic exclusion from head extension); (iii) $s(f(m \circ_{\text{priv}} h),\ f(y)) < \beta$ for all $y \in [\![h]\!]$, with $\beta < \alpha - \eta$ (retrieval isolation from true head instances).*

*Proof.* Suppose (i) and the consistency assumption both hold. Let $y \in [\![h]\!]$ be any genuine head instance. By the consistency assumption, $f(y)$ lies within distance $\eta$ (in the metric induced by $s$) of $f(h)$. By the triangle inequality, proximity of $f(m \circ_{\text{priv}} h)$ to $f(h)$ implies proximity to $f(y)$:

$$s(f(m \circ_{\text{priv}} h),\ f(y)) \geq \alpha - \eta - \epsilon,$$

where $\epsilon > 0$ is determined by the metric geometry. For $\eta$ and $\epsilon$ small relative to the gap $\alpha - \beta > 0$, this contradicts constraint (iii), which requires $s(f(m \circ_{\text{priv}} h), f(y)) < \beta < \alpha - \eta$. Hence (i) and (iii) cannot hold simultaneously. Relaxing

(i) to allow $s(f(m \circ_{\text{priv}} h), f(h)) < \alpha$ sacrifices lexical anchoring. Since the same global score $s$ governs both, no uniform choice satisfies all three constraints without additional type-discriminative structure. □

**Interpretation.** Privative failures are therefore not just cases of insufficient training. They expose a representational conflict in ordinary embedding spaces: the model must be close to the head in one sense and far from it in another.

### D.6 Equivalence Training Reorganizes Geometry

**Theorem D.5** (Capacity-Reorganization Principle)**.** *Let contrastive equivalence training update an encoder primarily through pairwise attraction of positives and repulsion of negatives in a fixed $d$-dimensional embedding space. Then, absent architectural expansion or new representational channels, improvements in retrieval arise from reorganization of the existing geometry rather than from increased representational capacity.*

*Proof.* The encoder output dimension remains $d$ throughout fine-tuning, so the ambient embedding space $\mathbb{R}^d$ and its capacity are fixed. Contrastive updates are gradient steps on pairwise cosine similarities: positive pairs receive attraction gradients and negative pairs receive repulsion gradients, modifying the weight matrices of the transformer but not adding independent output coordinates. The effect on the embedding geometry is a change in the covariance structure of $\{f(x)\}_{x \in \mathcal{X}}$: specifically, the singular value distribution of the embedding matrix shifts. Anisotropy (the fraction of variance captured by the leading singular direction) decreases as attraction and repulsion spread the embedding cloud more uniformly across directions. Crucially, the effective rank (number of dimensions carrying non-negligible variance, measured for example by the participation ratio $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ over singular values $\lambda_i$) may remain approximately stable even as the distribution across directions changes. Contrastive training therefore reorganizes the geometry (changing which directions carry semantic signal) without expanding it (the number of independent representational axes is bounded by $d$ before and after). □

**Interpretation.** The theorem supports the claim that concept-equivalence fine-tuning recalibrates a pre-existing semantic space rather than expanding it.

### D.7 Distortion and Operator Complexity

**Theorem D.6** (Operator Distortion Lower Bound)**.** *Let $\kappa(\tau)$ denote the number of independent semantic degrees of freedom required to realize composition type $\tau$, including latent relation variables, modal parameters, and exclusion constraints. If these degrees of freedom are not encoded in the supervision or representation, then any realization of $\tau$ in a $d$-dimensional embedding space incurs nonzero distortion. Moreover, distortion must increase as the unsupervised degrees of freedom of $\tau$ increase relative to the information preserved by $f$.*

*Proof sketch.* An encoder can faithfully realize $C_\tau$ only to the extent that the $\kappa(\tau)$ semantic variables determining that operator are preserved in the representation. If two expressions differ in a latent variable required by $C_\tau$ (a relation argument, an intensional parameter, or a privative exclusion flag) but the encoder maps them to near-identical vectors, then no downstream $\Phi_\tau$ can recover the distinction. Each such collapsed variable represents at least one dimension of the target space that is rendered inaccessible. By a counting argument over the required variable assignments: if $k$ independent binary distinctions collapse (each mapped to indistinguishable vectors), then at least $k$ semantic contrasts cannot be expressed at the output, and the resulting retrieval errors are bounded away from zero for each. Since the number of collapsed variables is bounded below by $\kappa(\tau)$ minus the number of variables identified by the training signal, distortion increases monotonically as $\kappa(\tau)$ grows relative to supervised information. Making this argument tight requires an information-theoretic lower bound on the minimum encoding error for a $d$-dimensional representation with partially-observed supervision, which we defer to future work. □

**Interpretation.** This result gives a formal reason why relational, modal, and privative families remain difficult even when extensional concept equivalence improves.

### D.8 Readout Invariance under Final-Layer Concentration

**Theorem D.7** (Readout Invariance)**.** *Let $h_l(x) \in \mathbb{R}^d$ denote the output of transformer layer $l$ for expression $x$, and let $f_r(x) = \sum_{l=1}^{L} w_l \, h_l(x)$ be the*

*pooled representation under readout $r = \{w_l\}_{l=1}^{L}$ with $\sum_l w_l = 1$. Suppose semantic content concentrates in the final layer: $h_l(x) = h_L(x) + \xi_l(x)$ where $\|\xi_l(x)\| \le \gamma$ for all $l < L$ and all $x$. Then for any two readout strategies $r$ and $r'$,*

$$\|f_r(x) - f_{r'}(x)\| \le 2\gamma.$$

*In particular, when $\gamma$ is small, the choice of readout has negligible effect on retrieval.*

*Proof.*

$$\begin{aligned}
\|f_r(x) - f_{r'}(x)\| &= \left\| \sum_l (w_l - w_l')\, h_l(x) \right\| \\
&= \left\| \sum_l (w_l - w_l') \big( h_L(x) + \xi_l(x) \big) \right\| \\
&= \Bigg\| h_L(x) \underbrace{\sum_l (w_l - w_l')}_{=0} \\
&\qquad + \sum_l (w_l - w_l')\, \xi_l(x) \Bigg\| \\
&\le \sum_l |w_l - w_l'| \cdot \|\xi_l(x)\| \;\le\; 2\gamma,
\end{aligned}$$

where $\sum_l (w_l - w_l') = 0$ because both weight vectors sum to 1, and $\sum_l |w_l - w_l'| \le 2$ because the $L_1$ distance between two probability vectors is at most 2. □

**Interpretation.** This theorem formalizes the P2 finding that pooling strategy is a null factor in fine-tuned encoders. Equivalence training concentrates semantic content in the final transformer layer, driving $\gamma$ toward zero: intermediate layers carry only low-variance residuals, so the choice among mean pooling, CLS, weighted mean, or gated readout produces embeddings within $2\gamma$ of one another. Pre-fine-tuning, layer representations are more heterogeneous ($\gamma$ larger) and readout strategies diverge correspondingly. The theorem also predicts that readout sensitivity should correlate with anisotropy: both reflect the distribution of semantic content across encoder layers and directions.

### D.9 Summary

The theoretical picture is that sentence encoders approximate a homomorphic mapping from a typed semantic algebra into a fixed geometric space. The seven theorems above provide formal grounding for the four empirical principles of the main paper. **P1** (recalibration, not expansion) is supported by the Capacity-Reorganization Principle (Theorem 5). Contrastive fine-tuning changes the covariance structure of the space but not its dimension, so improvements arise from redistribution rather than augmentation. **P2** (final-layer concentration) is supported by the Readout Invariance theorem. Once semantic content concentrates in the final layer, all linear pooling strategies converge to within $2\gamma$ of one another. **P3** (calibration and ranking separable via hard negatives) is supported by the Approximate Homomorphism and Head-Preservation theorems (Theorems 1, 3). Stable retrieval requires a low-distortion composition operator, and head-preserving families achieve this under the existing geometry while operator-complex families require explicit negative supervision. **P4** (supervision must match composition type) is supported by the Supervision Identifiability theorem (Theorem 2). Equivalence labels identify geometry only up to equivalence-preserving transformations; composition-type-specific operators are underdetermined without targeted supervision. The Privative Metric Conflict (Theorem 4) and Operator Distortion Lower Bound (Theorem 6) additionally explain why privative, modal, and relational families remain hard even after equivalence fine-tuning. They require representational structure that a single global metric over synonym-definition pairs cannot provide.

## E Training Hyperparameters

Table 9 lists the full set of hyperparameters used across all conditions.

## F Loss Variant: Unified InfoNCE

A simplified variant (B1-unified) folds hard negatives directly into the InfoNCE denominator, eliminating the BCE term. Results are shown in Table 10. B1-unified improves retrieval ranking (t2d R@10 +0.012, t2d MRR +0.014) but substantially reduces calibration (pair ROC-AUC −0.058) and hard-negative stress-test robustness (negate ROC-AUC −0.182). The BCE term provides a fixed-magnitude gradient per hard negative regardless of batch size; the unified formulation dilutes the hard-negative signal in proportion to batch size, reducing its effectiveness for calibration. The primary BCE variant is preferred when downstream use requires

| Hyperparameter | Value |
|---|---|
| Optimizer | AdamW |
| Base learning rate | $2\times10^{-5}$ |
| LR schedule | Linear warmup + decay |
| Warmup steps | 100 |
| Layerwise LR decay | 0.90 per layer |
| Pooling LR multiplier | $5\times$ base |
| Weight decay | 0.01 |
| Gradient clip norm | 1.0 |
| Batch size | 16 |
| Training steps | 8,000 |
| Max sequence length | 128 tokens |
| InfoNCE temperature $\tau$ | 0.05 |
| Hard-neg margin | 0.20 |
| Layer window $K$ | 4 |

Table 9: Hyperparameters shared across all fine-tuned conditions.

calibrated pair scoring or robustness to semantic perturbations.

## G Hard-Negative Rules

Five rule-based transforms are applied to positive definitions to generate hard negatives. The same five rules are used during both training (BCE hard-negative term) and evaluation (stress-test ROC-AUC), so the stress tests directly probe the discriminations the model was trained on. Each rule is applied once per positive definition; all generated negatives are used.

**Antonym flip.** A key content word in the definition is identified and replaced by its antonym using a WordNet antonymy lookup. This tests whether the model can distinguish conceptually opposite definitions that share the same syntactic frame (*"a sustained increase in prices"* → *"a sustained decrease in prices"*).

**Lexical negation.** The first finite verb or copula in the definition is negated by inserting *not* (*"is a warm-blooded vertebrate"* → *"is not a warm-blooded vertebrate"*). This is the most challenging type: the negative differs from the positive by a single token, and the models must assign lower similarity to what is linguistically a denial of the positive concept.

**Same-POS random swap.** The positive definition is replaced with a randomly sampled definition from a different word sharing the same coarse POS tag (noun, verb, adjective, adverb). This tests broad categorical discrimination: the model must not rely solely on POS-level similarity cues to rank candidates.

**Same-POS prefix swap.** As above, but the replacement definition is drawn from a word sharing the same morphological prefix (*un-*, *re-*, *over-*, etc.). The prefix overlap makes this harder than random same-POS swaps, as it introduces surface cues that can mislead lexical-similarity-based models.

**Ontological type swap.** The definition is replaced with one from a concept of a different ontological type, where type is drawn from a small closed set (person, organization, location, event, artifact, biological entity, abstract concept). This tests whether the model has encoded the coarse ontological category of a concept rather than merely its surface distribution (*"the capital city of France"* → *"a multinational technology corporation"*).

**Evaluation procedure.** For each test pair (query $q$, positive definition $d^+$), a negative $d^-$ is generated by applying one rule to $d^+$. The model scores $\cos(f(q), f(d^+))$ and $\cos(f(q), f(d^-))$; ROC-AUC is computed over the resulting binary discrimination across all test pairs for that rule. A score of 1.0 indicates perfect discrimination; 0.5 is chance.

## H Geometry Diagnostics

We measure two complementary geometry properties on the in-domain split. **Anisotropy** is the mean cosine similarity over 10,000 random pairs of encoded texts: lower values indicate a more isotropic, uniformly spread space. **Effective rank** is $\exp(H(\sigma))$, the exponential Shannon entropy of the normalised singular-value spectrum: higher values indicate that more dimensions carry meaningful variance. Together these characterise whether the model is using the available dimensions efficiently and without degenerate clustering.

**Why B0 already has good geometry.** The frozen mean-pool baseline (B0) achieves surprisingly well-calibrated geometry (anisotropy 0.126, effective rank 247.0) without any task-specific training. This is a direct consequence of the backbone's history: `all-mpnet-base-v2` was itself contrastively trained on 1 billion sentence pairs for sentence-level similarity (Reimers and Gurevych, 2019), which spreads representations isotropically and concentrates signal in the final layer. The per-layer anisotropy plot (Figure 3) makes this concrete: layer-12 anisotropy in the frozen encoder (0.146) is far below layers 5–11 (0.45–0.50). We inherit this structure without modification; the B0 baseline is

| Condition | t2d R@10 | t2d MRR | Pair AUC | HH d2t | Negate | NP R@10 |
|---|---|---|---|---|---|---|
| B1 | 0.654 | 0.537 | **0.963** | 0.686 | **0.980** | 0.656 |
| B1-unified | **0.666** | **0.551** | 0.905 | **0.692** | 0.798 | 0.656 |

Table 10: B1 vs. B1-unified across the same evaluation columns as Table 2. Unified InfoNCE improves retrieval ranking but substantially reduces calibration and stress-test robustness. Bold: best per column.

| ID | Condition | Aniso. | Eff. rank |
|---|---|---|---|
| B0 | Frozen mean | 0.126 | 247.0 |
| B0-WM | Frozen wtd. mean | 0.303 | 177.0 |
| B1 | Mean + HN | **0.012** | **247.2** |
| B3 | Weighted mean | 0.166 | 193.9 |
| M1 | Weighted mean + HN | 0.015 | 239.4 |

Table 11: Geometry diagnostics (in-domain split). Anisotropy: mean cosine over random pairs (lower = better). Effective rank: $\exp(H(\sigma))$, entropy of the normalised eigenspectrum (higher = better). Hard negatives are the dominant geometric factor (B1 vs. B3).

strong not because of anything we did, but because the backbone was already well-calibrated.

**Why B0-WM is substantially worse.** B0-WM (frozen uniform weighted mean over the last four layers) has anisotropy 0.303 and effective rank only 177.0—substantially worse than B0 despite using the same frozen backbone. The reason is straightforward: averaging layers 9–11 into the layer-12 output introduces high-anisotropy (0.45–0.50) intermediate representations into the mixture, diluting the well-calibrated final layer. This is the same mechanism as layer-12 dominance, operating at the pooling level: even without gradient flow, choosing the wrong readout is geometrically costly. The retrieval penalty mirrors the geometry penalty—B0-WM achieves only 0.392 t2d R@10 vs. B0's 0.552 (Table 2).

**Fine-tuning reorganises, not expands.** The most striking geometry finding is that fine-tuning on 3.3 million concept pairs leaves effective rank essentially unchanged (B0: 247.0 → B1: 247.2) while collapsing anisotropy (0.126 → 0.012). Contrastive training does not recruit new dimensions; it redistributes variance across existing ones, rotating and rescaling directions so that equivalent concepts land in the same neighbourhood and hard negatives are pushed apart. The representational capacity (effective rank ≈ 247) was already present in the frozen encoder, contributed by the prior sentence-transformer training. Concept-equivalence fine-tuning directs it, but cannot create it.

**Hard negatives are the dominant geometric factor.** The B3 vs. B1 comparison isolates the contribution of hard-negative supervision (same backbone, same pooling, same budget; only the BCE hard-negative term differs): B3 anisotropy is 0.166, effective rank 193.9; B1 achieves 0.012 and 247.2. Hard negatives require the model to push semantically distinct but geometrically close representations apart, which directly reduces anisotropy and expands the effective rank. By contrast, varying the pooling architecture while holding supervision fixed has a much smaller effect: M1 (weighted mean + HN) achieves anisotropy 0.015 and rank 239.4, only marginally below B1 (mean pool + HN) despite a different readout. This confirms that the geometry improvement is driven by the training signal, not the readout architecture—consistent with the H2 retrieval null result.

## I NP Paraphrase Error Examples

Table 12 lists the three hardest B1 failures per modifier family, sorted by rank of the correct candidate (descending). Rank is computed over the full 4,000-item pool. "Correct sim" and "Retrieved sim" are cosine similarities to the correct paraphrase and to the rank-1 retrieved candidate respectively. The final column classifies the failure mode as discussed in Section I.

## J Backbone Grid on DBpedia

Table 13 reports R@10 and MRR for all five backbones on DBpedia (245 test queries), both frozen and after 300 in-domain fine-tuning steps.

## K Use of AI Writing Assistance

During the preparation of this work the author(s) used Claude (Anthropic) in order to assist with experiment implementation (coding), writing, editing, and LaTeX preparation of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

| Family | Query | Correct paraphrase | Cor. sim | Rank | Retrieved rank-1 | Ret. sim | Failure mode |
|---|---|---|---|---|---|---|---|
| INTERSECTIVE | clean plate | free of dirt plate | 0.487 | 83 | clean dish | 0.832 | *adj. synonym* |
| | dirty temple | not clean temple | 0.577 | 59 | filthy temple | 0.937 | *adj. synonym* |
| | dirty cup | not clean cup | 0.583 | 58 | filthy cup | 0.838 | *adj. synonym* |
| SUBSECTIVE | effective student | high impact student | 0.689 | 48 | effective teacher | 0.892 | *head-noun confusion* |
| | effective trainer | high impact trainer | 0.726 | 42 | highly effective trainer | 0.933 | *degree-modifier variant* |
| | effective manager | high impact manager | 0.739 | 30 | highly effective manager | 0.918 | *degree-modifier variant* |
| RELATIONAL | community cup | container owned by local communities | 0.253 | 864 | city cup | 0.899 | *compound cross-retrieval* |
| | city cup | container associated with municipal service | 0.264 | 637 | community cup | 0.899 | *compound cross-retrieval* |
| | summer account | profile related to the warm season | 0.335 | 580 | student account | 0.771 | *compound cross-retrieval* |
| MODAL | alleged culprit | not proven culprit | 0.633 | 94 | purported culprit† | 0.929 | *annotation limit* |
| PRIVATIVE | toy ticket | pass lookalike | 0.347 | 497 | play ticket‡ | 0.807 | *annotation limit* |
| | fraudulent account | profile lookalike | 0.349 | 470 | forged account | 0.912 | *adj. synonym* |
| | spurious ticket | pass lookalike | 0.372 | 437 | bogus ticket | 0.927 | *adj. synonym* |

Table 12: Hardest B1 failures per modifier family (up to 3; 4,000-item pool). Rank = position of correct paraphrase; lower = harder failure. Failure modes: *adj. synonym* = same-adjective surface variant ranked above definitional paraphrase; *head-noun confusion* = correct modifier, wrong noun; *degree-modifier variant* = degree form ranked above definition; *compound cross-retrieval* = adjacent compound preferred over relational-clause paraphrase; *annotation limit* = retrieved item is a valid paraphrase under multi-reference annotation. † *purported culprit* is a valid paraphrase of *alleged culprit*; counted wrong due to single-reference annotation. ‡ *play ticket* is a valid paraphrase of *toy ticket*; counted wrong because the annotated reference is *pass lookalike*.

| | Frozen | | Fine-tuned | |
|---|---|---|---|---|
| Backbone | R@10 | MRR | R@10 | MRR |
| `all-mpnet-base-v2` | 0.608 | 0.349 | **0.845** | **0.550** |
| `jina-v5-nano` | **0.784** | **0.431** | 0.829 | 0.468 |
| `e5-base-v2` | 0.551 | 0.299 | 0.816 | 0.519 |
| `para-mpnet-base-v2` | 0.624 | 0.307 | 0.816 | 0.508 |
| `para-MiniLM-L6` | 0.539 | 0.248 | 0.678 | 0.360 |

Table 13: Backbone grid on DBpedia (245 test queries, 2,849 candidates). All fine-tuned models use 300 DBpedia-specific steps. Bold: best per column. Contrastive fine-tuning improves every backbone tested.