# iBERT: Interpretable Embeddings via Sense Decomposition


**Vishal Anand**

**Milad Alshomary**

**Kathleen McKeown**



EACL 2026


*iBERT.io | github.com/vishalanand/iBERT*

# The Problem with Dense Encoders

Modern sentence encoders **(SBERT, SimCSE)** excel at semantic retrieval but…

- **Dense, opaque embeddings**
  - No visibility into which linguistic attributes are captured
  - Style and semantics are entangled — hard to disentangle by design
- **Style-sensitive NLP is underserved**
  - Authorship attribution, tone control, stylistic retrieval all require disentangled features
  - Post-hoc methods (probing, saliency) yield incomplete or unfaithful explanations

# The Problem with Dense Encoders

Modern sentence encoders **(SBERT, SimCSE)** excel at semantic retrieval but…

- **Dense, opaque embeddings**
  - No visibility into which linguistic attributes are captured
  - Style and semantics are entangled — hard to disentangle by design
- **Style-sensitive NLP is underserved**
  - Authorship attribution, tone control, stylistic retrieval all require disentangled features
  - Post-hoc methods (probing, saliency) yield incomplete or unfaithful explanations

**Key gap:** No encoder where stylistic and semantic representations are ***explicit and controllable by design*** — not post-hoc.

# Research Question

*“Can we design an encoder where semantic and stylistic representations are made explicit and controllable within the embedding space — by design, not post-hoc?”*

# Research Question

> *"Can we design an encoder where semantic and stylistic representations are made explicit and controllable within the embedding space — by design, not post-hoc?"*

*We want embeddings that are…*

**Interpretable**

Each dimension maps to a recognizable stylistic/semantic axis

**Controllable**

Edit specific aspects of a text's representation without affecting others

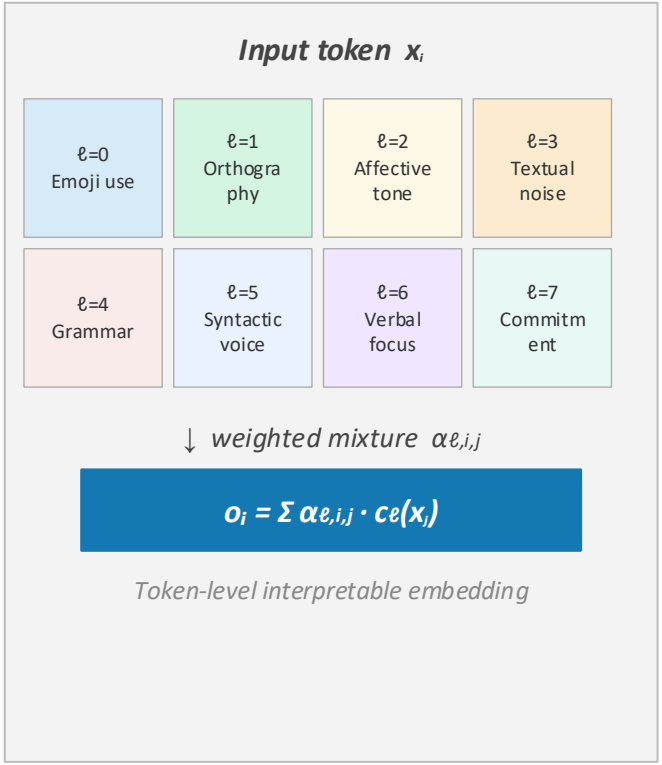
**Performant**

Match or exceed dense baselines on downstream style and semantic tasks

# iBERT: Key Idea

Each input token $x_i$ is represented as a **sparse, non-negative mixture** over $k$ context-independent sense vectors:

- Senses are learned from data
- Mixture weights are context-dependent (via attention)
- Representations are directly inspectable and editable
- Architecture is general — not limited to style

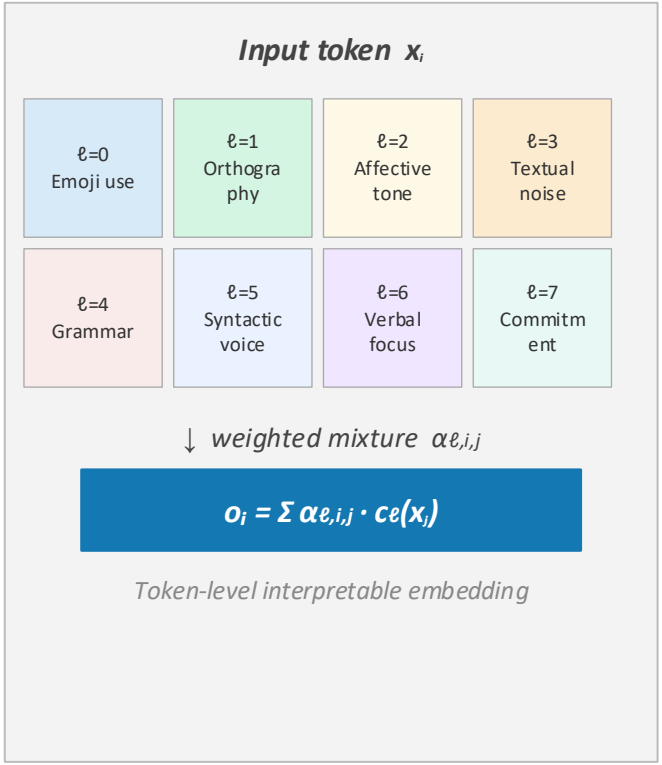


# iBERT: Key Idea

Each input token $x_i$ is represented as a **sparse, non-negative mixture** over $k$ context-independent sense vectors:

- Senses are learned from data
- Mixture weights are context-dependent (via attention)
- Representations are directly inspectable and editable
- Architecture is general — not limited to style

Builds on **Backpack-GPT (Hewitt et al., 2023)**, adapted to a Language Model encoder, alongwith additional global pooling for sentence-encoders



# iBERT Architecture — Language Model

Two parallel streams combine to form interpretable token embeddings:

**Stream 1: Sense Construction**

- Feed-forward layer maps $x_i$ → **k sense vectors**
  $C(x_i) = [c_1(x_i), \ldots, c_k(x_i)] \in \mathbb{R}^{kxd}$
- Produces context-independent senses
- k = 8 in all experiments

⊗

**Stream 2: Contextual Encoder**

- Transformer encoder produces **contextual states**
  $H = [h_1, \ldots, h_n] \in \mathbb{R}^{nxd}$
- Projected into sense-specific Q/K pairs
- Mixture weights $\alpha_{\ell,i,j}$ computed via <K,Q> attention

**Token embedding $o_i = \Sigma_\ell \Sigma_j \alpha_{\ell,i,j} \cdot c_\ell(x_j)$ → mixture of senses**

# iBERT Architecture — Sentence Encoder

**Sentence Embedding:**
Token embeddings $\mathbf{o_i}$ are pooled into a sentence embedding **s** via three strategies:

### v1 — Mean Pooling

$$s = (1/n)\ \Sigma_i\ o_i$$

Uniform average over all token embeddings.

### v2 — Top-Sense Pooling

$$s = (1/n)\ \Sigma_i\ o(\ell^*)_i \quad \text{where } \ell^* = \text{argmax } S_\ell$$

Hard selection of the single dominant sense across all tokens.

### v3 — Softmax Blend

$$s = (1/n)\ \Sigma_i\ \Sigma_\ell\ \pi_\ell(\tau) \cdot o(\ell)_i$$

Dominant senses weighted more than other senses.

τ=1 and τ=10 variants evaluated.
Best empirical performance.

# Two-Stage Training Pipeline

**Stage 0: MLM Pre-training**

- Train iBERT-MLM from scratch on 5% of FineWeb
- 750B tokens — broad linguistic and stylistic diversity
- Standard masked language modeling objective
- Yields 171M-parameter encoder (k = 8 senses)

**Stage 1: Style Representation Learning**

- Fine-tune with contrastive style triplets $\langle s, s^{+}, s^{-} \rangle$
- **StyleSynth (SD):** 50k synthetic triplets, 40 style axes - InfoNCE loss (SD)
- **Wegmann (WG):** 40k author-anchored Reddit triplets - Margin loss (WG)

*Baselines (BERT_SD / BERT_WG / BERT_WG+SD) trained identically — same data, backbone, optimizer — for a controlled comparison.*

# Evaluation Setup

## Benchmarks

| | |
|---|---|
| **STEL** | Multi-class style classification<br>(40 Style axes + 5 Wegmann) |
| **SoC** | Binary polarity separability per style class<br>(formal vs. informal, emoji vs. no emoji, …) |
| **PAN** | Authorship verification (AUC)<br>PAN 2011/13/14/15 — proxy-style task |

## Models Compared

| | |
|---|---|
| BERT_SD | Our BERT baseline · StyleSynth supervision |
| BERT_WG | Our BERT baseline · Wegmann supervision |
| BERT_WG+SD | StyleDistance (Patel et al., 2025) — prior SOTA |
| **iBERT-v1** | iBERT · mean pooling |
| **iBERT-v2** | iBERT · top-sense (hard) |
| **iBERT-v3-1** | iBERT · softmax blend, τ=1 |
| **iBERT-v3-10** | iBERT · softmax blend, τ=10 |

# Results

| Data | Models | 128 tokens | | |
|---|---|---|---|---|
| | | STEL↑ | SoC↑ | PAN↑ |
| SD-ONLY | BERT | 31.6 ± 0.78 | 91.5 ± 0.56 | *60.57* |
| | iBERT-v1 | *37.5* ± 0.41 | *92.3* ± 0.33 | **60.88** |
| | iBERT-v2 | 28.5 ± 1.13 | 88.7 ± 0.68 | 58.95 |
| | iBERT-v3-1 | **38.3** ± 0.62 | **92.8** ± 0.48 | 58.15 |
| | iBERT-v3-10 | *38.1* ± 0.44 | *92.5* ± 0.30 | 58.29 |
| WG+SD | BERT | 30.6 ± 0.93 | 88.9 ± 0.53 | 58.96 |
| | iBERT-v1 | **38.6** ± 1.57 | *90.0* ± 0.22 | 58.96 |
| | iBERT-v2 | 33.4 ± 0.54 | 89.1 ± 0.50 | 58.47 |
| | iBERT-v3-1 | *38.0* ± 0.85 | **90.2** ± 0.30 | 58.56 |
| | iBERT-v3-10 | *38.0* ± 0.88 | *89.6* ± 0.65 | **60.96** |
| WG-ONLY | BERT | 27.1 ± 0.70 | **27.5** ± 1.18 | 58.20 |
| | iBERT-v1 | *28.5* ± 0.86 | 8.0 ± 0.62 | *59.29* |
| | iBERT-v2 | 25.9 ± 0.76 | 6.7 ± 0.65 | 57.77 |
| | iBERT-v3-1 | *28.1* ± 0.59 | 7.4 ± 0.42 | *59.47* |
| | iBERT-v3-10 | **28.5** ± 0.58 | 7.3 ± 0.41 | **59.87** |

**STEL:** In all cases iBERT outperforms corresponding BERT baselines

→ iBERT-v3-1 achieves **38.3%** STEL vs. **31.6%** for BERT

— a gain of **+6.7 points** (SD-ONLY, 128 tokens)

# Results

| Data | Models | 128 tokens | | |
|---|---|---|---|---|
| | | STEL↑ | SoC↑ | PAN↑ |
| SD-ONLY | BERT | 31.6 ± 0.78 | 91.5 ± 0.56 | *60.57* |
| | iBERT-v1 | *37.5* ± 0.41 | *92.3* ± 0.33 | **60.88** |
| | iBERT-v2 | 28.5 ± 1.13 | 88.7 ± 0.68 | 58.95 |
| | iBERT-v3-1 | **38.3** ± 0.62 | **92.8** ± 0.48 | 58.15 |
| | iBERT-v3-10 | *38.1* ± 0.44 | *92.5* ± 0.30 | 58.29 |
| WG+SD | BERT | 30.6 ± 0.93 | 88.9 ± 0.53 | 58.96 |
| | iBERT-v1 | **38.6** ± 1.57 | *90.0* ± 0.22 | 58.96 |
| | iBERT-v2 | 33.4 ± 0.54 | 89.1 ± 0.50 | 58.47 |
| | iBERT-v3-1 | *38.0* ± 0.85 | **90.2** ± 0.30 | 58.56 |
| | iBERT-v3-10 | *38.0* ± 0.88 | *89.6* ± 0.65 | **60.96** |
| WG-ONLY | BERT | 27.1 ± 0.70 | **27.5** ± 1.18 | 58.20 |
| | iBERT-v1 | *28.5* ± 0.86 | 8.0 ± 0.62 | *59.29* |
| | iBERT-v2 | 25.9 ± 0.76 | 6.7 ± 0.65 | 57.77 |
| | iBERT-v3-1 | *28.1* ± 0.59 | 7.4 ± 0.42 | *59.47* |
| | iBERT-v3-10 | **28.5** ± 0.58 | 7.3 ± 0.41 | **59.87** |

**STEL:** In all cases iBERT outperforms corresponding BERT baselines

→ iBERT-v3-1 achieves **38.3%** STEL vs. **31.6%** for BERT

— a gain of **+6.7 points** (SD-ONLY, 128 tokens)

**SoC:** Results are mixed

→ iBERT-v3-1: **92.8%** vs BERT 91.5% (SD-ONLY, 128 tokens)

# Results

| Data | Models | 128 tokens | | |
|---|---|---|---|---|
| | | STEL↑ | SoC↑ | PAN↑ |
| SD-ONLY | BERT | 31.6 ± 0.78 | 91.5 ± 0.56 | *60.57* |
| | iBERT-v1 | *37.5* ± 0.41 | *92.3* ± 0.33 | **60.88** |
| | iBERT-v2 | 28.5 ± 1.13 | 88.7 ± 0.68 | 58.95 |
| | iBERT-v3-1 | **38.3** ± 0.62 | **92.8** ± 0.48 | 58.15 |
| | iBERT-v3-10 | *38.1* ± 0.44 | *92.5* ± 0.30 | 58.29 |
| WG+SD | BERT | 30.6 ± 0.93 | 88.9 ± 0.53 | 58.96 |
| | iBERT-v1 | **38.6** ± 1.57 | *90.0* ± 0.22 | 58.96 |
| | iBERT-v2 | 33.4 ± 0.54 | 89.1 ± 0.50 | 58.47 |
| | iBERT-v3-1 | *38.0* ± 0.85 | **90.2** ± 0.30 | 58.56 |
| | iBERT-v3-10 | *38.0* ± 0.88 | *89.6* ± 0.65 | **60.96** |
| WG-ONLY | BERT | 27.1 ± 0.70 | **27.5** ± 1.18 | 58.20 |
| | iBERT-v1 | *28.5* ± 0.86 | 8.0 ± 0.62 | *59.29* |
| | iBERT-v2 | 25.9 ± 0.76 | 6.7 ± 0.65 | 57.77 |
| | iBERT-v3-1 | *28.1* ± 0.59 | 7.4 ± 0.42 | *59.47* |
| | iBERT-v3-10 | **28.5** ± 0.58 | 7.3 ± 0.41 | **59.87** |

**STEL:** In all cases iBERT outperforms corresponding BERT baselines

→ iBERT-v3-1 achieves **38.3%** STEL vs. **31.6%** for BERT

— a gain of **+6.7 points** (SD-ONLY, 128 tokens)

**SoC:** Results are mixed

→ iBERT-v3-1: **92.8%** vs BERT 91.5% (SD-ONLY, 128 tokens)

**PAN:** iBERT outperforms its equivalent BERT baselines

→ iBERT-v3-10: **59.87%** vs BERT 58.20% (WG-ONLY, 128 tokens)

# Interpretability: Emergent Sense Specialization

*Do individual senses capture coherent stylistic themes?*

**A Two-Step Method: Probing → Ablation**

# Interpretability: Emergent Sense Specialization

*Do individual senses capture coherent stylistic themes?*

**A Two-Step Method: Probing → Ablation**

**1)** **Probing:** Which sense $\ell$ encodes which style axis?

# Interpretability: Emergent Sense Specialization

*Do individual senses capture coherent stylistic themes?*

**A Two-Step Method: Probing → Ablation**

**1) Probing:** Which sense ℓ encodes which style axis?

| STYLE | $\ell = 0$ | $\ell = 1$ | $\ell = 2$ | $\ell = 3$ | $\ell = 4$ | $\ell = 5$ | $\ell = 6$ | $\ell = 7$ | BEST |
|---|---|---|---|---|---|---|---|---|---|
| Active / Passive | 0.6222 | 0.6222 | 0.5278 | 0.6278 | 0.6111 | 0.6444 | 0.5944 | 0.6333 | 5 |
| Affective process / Perceptual process | 0.6389 | 0.5444 | 0.6500 | 0.5889 | 0.6222 | 0.5833 | 0.6444 | 0.6278 | 2 |
| Affective processes / Cognitive processes | 0.6722 | 0.5833 | 0.6833 | 0.6111 | 0.6833 | 0.6389 | 0.7000 | 0.6722 | 6 |
| All Lower Case / Proper Capitalization | 0.9778 | 0.7611 | 0.9667 | 0.8444 | 0.9667 | 0.8778 | 0.9833 | 0.9833 | 7 |
| All Upper Case / Proper Capitalization | 0.9278 | 0.9444 | 0.9278 | 0.9333 | 0.9333 | 0.9389 | 0.9167 | 0.9222 | 1 |
| Certain / Uncertain | 0.6444 | 0.6056 | 0.6167 | 0.6444 | 0.6389 | 0.6500 | 0.6833 | 0.6944 | 7 |
| Cognitive process / Perceptual process | 0.6444 | 0.6167 | 0.5500 | 0.6500 | 0.6500 | 0.6500 | 0.6444 | 0.6667 | 7 |
| Complex / Simple | 0.5944 | 0.5444 | 0.6278 | 0.5611 | 0.6000 | 0.5556 | 0.6333 | 0.5944 | 6 |
| Fluent sentence / Disfluent sentence | 0.6278 | 0.5889 | 0.5111 | 0.6167 | 0.6111 | 0.6167 | 0.6444 | 0.6000 | 6 |
| Formal / Informal | 0.7833 | 0.7389 | 0.8167 | 0.7667 | 0.8167 | 0.7667 | 0.7667 | 0.7778 | 2 |
| Long average word length / Short average word length | 0.8389 | 0.8444 | 0.7944 | 0.8444 | 0.8389 | 0.8333 | 0.8278 | 0.8389 | 1 |
| Offensive / Non-Offensive | 0.9111 | 0.7389 | 0.9444 | 0.7611 | 0.9056 | 0.7722 | 0.9111 | 0.9000 | 2 |
| Polite / Impolite | 0.6389 | 0.4167 | 0.7333 | 0.4833 | 0.6667 | 0.4889 | 0.6444 | 0.5833 | 2 |
| Positive / Negative | 0.5944 | 0.5111 | 0.6944 | 0.5222 | 0.5944 | 0.5111 | 0.6056 | 0.5556 | 2 |

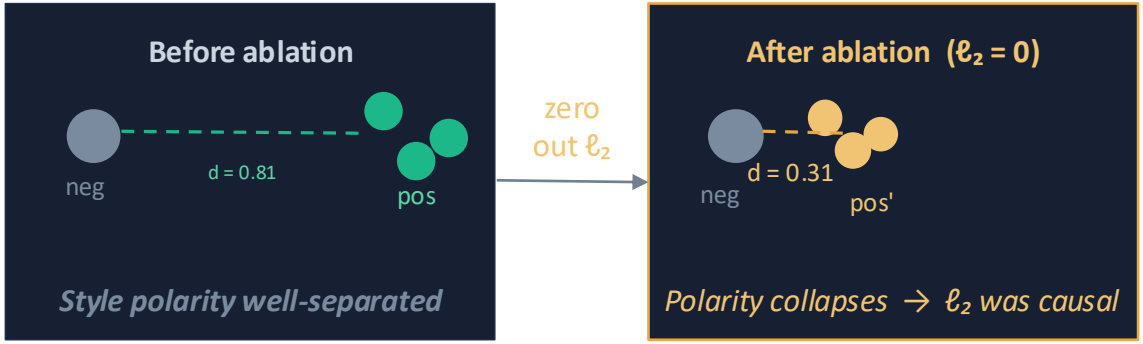
# Interpretability: Emergent Sense Specialization

*Do individual senses capture coherent stylistic themes?*

**A Two-Step Method: Probing → Ablation**

1) **Probing:** Which sense ℓ encodes which style axis?

2) **Ablation:** Is the sense causally responsible (not just correlated)?



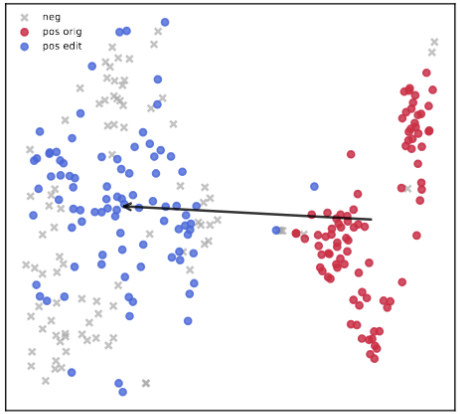
# Interpretability: Emergent Sense Specialization

*Do individual senses capture coherent stylistic themes?* Yes — without explicit annotation.

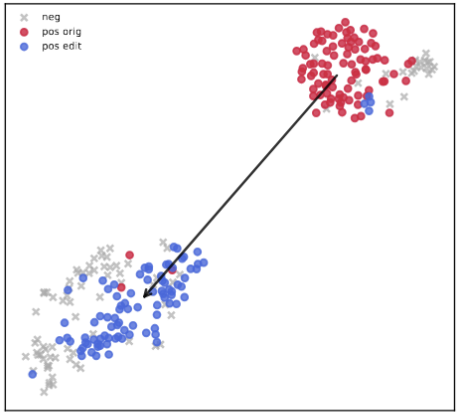
| Sense | Top-Aligned Style Axes | Emergent Theme |
|---|---|---|
| $\ell = 0$ | With Emojis / No Emojis<br>Frequent/Infrequent Conjunctions<br>Personal Pronouns | **Surface-level markers** |
| $\ell = 1$ | All Upper Case / Proper Capitalization<br>Text Emojis · Long/Short Word Length<br>Number Substitution | **Orthographic & visual style** |
| $\ell = 2$ | Humorous / Non-Humorous<br>Sarcastic / Non-Sarcastic<br>Metaphoric · Offensive | **Affect & expressive tone** |
| $\ell = 5$ | Active / Passive<br>Contracted / Non-Contracted | **Syntactic voice & register** |
| $\ell = 6$ | Frequent/Infrequent Pronouns<br>More/Less Frequent Verbs | **Pronoun & verbal focus** |
| $\ell = 7$ | With/Without Determiners<br>Certain / Uncertain | **Grammatical commitment** |

# Interpretability: Targeted Embedding Editing

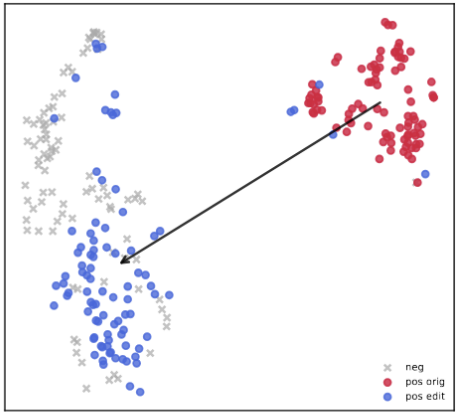
- Ablate a target sense (set its gain to zero) → positive samples collapse toward negative centroid
- Non-target senses left unchanged → unrelated style contrasts remain stable
- Three examples from iBERT-v3-10:




(a) Text Emojis / No Emojis ($\ell = 1$, $\Delta$Dist: 84 %)




(b) With/Without Nominalizations ($\ell$=4, $\Delta$Dist: 67%)




(c) All Lowercase / Proper Caps ($\ell = 7$, $\Delta$Dist: 70%)

***Non-target senses ablated → positive/negative clusters remain well-separated, confirming locality.***

# Conclusion

### Interpretability without performance cost

iBERT matches or exceeds dense baselines on STEL, SoC, and PAN while being inherently decomposable. Interpretability and performance are not in tension here.

### Structure + supervision are both necessary

Data alone (SD) is insufficient for disentanglement. Meaningful editablity and disentanglement emerges with architecture (iBERT) and data (SD). Under proxy signals (WG-only), sense specialization weakens.

### Localized edits, targeted control

Ablating a single sense (e.g., $\ell$=1 for emoji, $\ell$=4 for nominalization) shifts only the intended stylistic axis, leaving unrelated styles intact. Infeasible in dense encoders.

### Applications beyond style

iBERT is a general encoder: sense ablation for classifier debiasing, sense-conditioned retrieval, latent data augmentation, embedding-level control in RAG pipelines.

# Conclusion

## Interpretability without performance cost

iBERT matches or exceeds dense baselines on STEL, SoC, and PAN while being inherently decomposable. Interpretability and performance are no

## Structure + supervision are both necessary

Data alone (SD) is insufficient for disentanglement. Meaningful editablity and disentanglement emerges with BERT) and data (SD). Under proxy signals se specialization weakens.

**Thanks for listening**

## Localized edits, targeted control

Ablating a single sense (e.g., ℓ=1 for em nominalization) shifts only the intended stylistic axis, leaving unrelated styles intact. Infeasible in dense encoders.

## eyond style

eral encoder: sense ablation for classifier debiasing, sense-conditioned retrieval, latent data augmentation, embedding-level control in RAG pipelines.