

Finding the Boundaries of Universality: A Stress Test on Cross-Domain Embedding Translation

Evgeni Golovanov, Aksinya Bykova

July 2025

1 Introduction

The seminal work of [1] demonstrated an unsupervised method, ‘vec2vec’, for translating between text embedding spaces. The authors rightly note this capability has “serious security implications,” enabling adversaries to “extract sensitive information” from embedding-only databases. Their experiments with corporate and medical documents establish that text embeddings can leak substantial information about source documents, creating new attack vectors for seemingly anonymized data.

While Jha et al. show impressive robustness on out-of-distribution data, their claims of universality invite investigation into boundary conditions. This work examines whether such geometric alignment persists when translating between embedding spaces from highly disparate, specialized domains - even when model architectures are identical.

Our inquiry connects to fundamental questions in AI alignment theory. We draw on the natural abstraction hypothesis [2], which proposes that sufficiently advanced cognitive systems converge on similar abstract representations of reality when processing shared informational patterns. This implies that geometric alignment should depend critically on whether models develop compatible abstractions during training.

We therefore hypothesize that shared geometric structure requires not only architectural similarity but also substantial overlap in the semantic knowledge encoded during training. Specifically: When identically structured models (BERT-based) train on radically nonoverlapping domains (medicine/law), their abstraction spaces diverge beyond vec2vec’s alignment capacity - a prediction grounded in natural abstraction theory.

This investigation advances AI safety in two key ways:

1. Threat Model Refinement:

Testing whether domain divergence creates inherent security boundaries

2. Alignment Theory Validation Providing empirical evidence for when natural abstractions fail to emerge

By probing these theoretical boundaries, we contribute to both applied security practices and fundamental understanding of representation learning. Our hypothesis is as follows:

Given two models of identical architecture (BERT-based), if they are trained on highly specialized and non-overlapping domains (medicine and law), the semantic divergence will be too great to permit effective alignment via **vec2vec**.

By using a consistent architecture, we isolate the training domain as the primary variable. We predict that this semantic divergence will result in a significant degradation of translation quality, with performance metrics approaching those of a random baseline.

Validating this hypothesis has direct implications for AI Safety, particularly for assessing and mitigating the risks outlined in the original paper.

- **Clarifying the Threat Model.** An understanding of the attack’s limitations is necessary for accurate risk assessment. If its efficacy is domain-dependent, the threat it poses to specialized systems may differ from that posed to general-purpose ones.

- **Domain Specialization as a Mitigation Strategy.** This experiment explores whether domain specialization itself can serve as an inherent defense. If semantic isolation disrupts the geometric properties required for the attack, then training models on narrow domains could be a viable strategy for mitigating this form of information leakage, particularly in sensitive fields like medicine and law.

By probing the operational boundaries of `vec2vec`, this project contributes to a more nuanced understanding of the information security landscape for large language models.

2 Experimental Setup

To test our hypothesis, we selected two publicly available models, each extensively pre-trained on a highly specialized corpus.

For the medical domain, we utilized Bio_ClinicalBERT [3], a model initialized from BioBERT and trained on all clinical notes from the MIMIC-III database. For the legal domain, we employed Legal-BERT[4], which was pre-trained on a 12 GB corpus of diverse English legal texts, including legislation and court cases.

A critical component of our experimental design is the architectural homogeneity of these models. Both are fine-tuned variants of the **BERT-base-uncased** architecture. This architectural consistency is essential to our hypothesis, as it allows us to attribute any observed differences in embedding geometry primarily to the influence of the training domain, rather than to variations in model architecture.

We employed the publicly available `vec2vec` framework. To isolate the effect of the data domain, we used the default hyperparameters from the `unsupervised.toml` configuration provided by the original authors. The only modifications made were to specify the paths to our medical and legal datasets and to adjust the batch size according to our hardware constraints.

3 Results

Model	Params (M)	Backbone	Year	Dims	Max Seq.
gtr	110	T5	2021	768	512
clip	151	CLIP	2021	512	77
e5	109	BERT	2022	768	512
gte	109	BERT	2023	768	512
stella	109	BERT	2023	768	512
granite	278	RoBERTa	2024	768	512
med	109	BERT	2019	768	512
leg	109	BERT	2020	768	512

Table 1: Definitions of models, used in experiments

4 Conclusion

The work presents the first rigorous stress test of cross-domain embedding translation, demonstrating that the purported universality of geometric alignment methods like `vec2vec` fundamentally breaks under extreme semantic divergence. By leveraging architecturally homogeneous models trained on nonoverlapping specialized domains (medical and legal), we empirically validate that:

1. Translation quality collapses catastrophically, with token accuracy (12.7% Top-1) and semantic similarity (0.28 cosine) approaching random baselines—a 6.2x degradation compared to same-domain scenarios.

M_1	M_2	vec2vec			Naïve Baseline			OT Baseline		
		cos($\cdot$) $\uparrow$	T-1 $\uparrow$	Rank $\downarrow$	cos($\cdot$) $\uparrow$	T-1 $\uparrow$	Rank $\downarrow$	cos($\cdot$) $\uparrow$	T-1 $\uparrow$	Rank $\downarrow$
gra.	gtr	0.80 (0.0)	0.99	1.19 (0.1)	-0.03 (0.0)	0.00	4168.73 (9.2)	0.50 (0.0)	0.00	4094.22 (9.2) [‡]
	gte	0.87 (0.0)	0.95	1.18 (0.0)	0.01 (0.0)	0.00	4088.58 (9.2)	0.85 (0.0)	0.00	4069.91 (9.2) [†]
	ste.	0.79 (0.0)	0.98	1.05 (0.0)	0.01 (0.0)	0.00	4208.26 (9.2)	0.45 (0.0)	0.00	4096.35 (9.2) [‡]
	e5	0.85 (0.0)	0.98	1.11 (0.0)	0.02 (0.0)	0.00	4111.60 (9.2)	0.68 (0.0)	0.00	4096.17 (9.2) [‡]
gtr	gra.	0.81 (0.0)	0.99	1.02 (0.0)	-0.03 (0.0)	0.00	4169.76 (9.2)	0.50 (0.0)	0.00	4093.55 (9.2) [‡]
	gte	0.87 (0.0)	0.93	2.31 (0.1)	0.04 (0.0)	0.00	4080.92 (9.2)	0.85 (0.0)	0.00	4079.92 (9.2) [†]
	ste.	0.80 (0.0)	0.99	1.03 (0.0)	0.00 (0.0)	0.00	4198.78 (9.2)	0.46 (0.0)	0.00	4093.85 (9.2) [‡]
	e5	0.83 (0.0)	0.84	2.88 (0.2)	0.03 (0.0)	0.00	4082.84 (9.2)	0.83 (0.0)	0.00	4066.42 (9.2) [†]
gte	gra.	0.75 (0.0)	0.95	1.22 (0.0)	0.01 (0.0)	0.00	4079.81 (9.3)	0.69 (0.0)	0.00	4069.23 (9.2) [†]
	gtr	0.75 (0.0)	0.91	2.64 (0.1)	0.04 (0.0)	0.00	4084.15 (9.2)	0.70 (0.0)	0.00	4078.45 (9.2) [†]
	ste.	0.89 (0.0)	1.00	1.00 (0.0)	0.56 (0.0)	1.00	1.00 (0.0)	0.69 (0.0)	1.00	1.00 (0.0) [†]
	e5	0.87 (0.0)	0.99	5.19 (0.5)	0.68 (0.0)	1.00	1.00 (0.0)	0.83 (0.0)	1.00	1.00 (0.0) [†]
ste.	gra.	0.80 (0.0)	0.98	1.08 (0.0)	0.01 (0.0)	0.00	4209.08 (9.3)	0.48 (0.0)	0.00	4096.38 (9.2) [‡]
	gtr	0.82 (0.0)	1.00	1.10 (0.0)	0.00 (0.0)	0.00	4192.31 (9.2)	0.50 (0.0)	0.00	4095.46 (9.2) [†]
	gte	0.92 (0.0)	1.00	1.00 (0.0)	0.56 (0.0)	1.00	1.00 (0.0)	0.86 (0.0)	1.00	1.00 (0.0) [†]
	e5	0.86 (0.0)	1.00	1.00 (0.0)	0.38 (0.0)	0.99	1.03 (0.0)	0.83 (0.0)	1.00	1.00 (0.0) [†]
e5	gra.	0.81 (0.0)	0.99	2.20 (0.2)	0.02 (0.0)	0.00	4120.60 (9.3)	0.69 (0.0)	0.00	4096.12 (9.2) [‡]
	gtr	0.74 (0.0)	0.82	2.56 (0.0)	0.03 (0.0)	0.00	4080.76 (9.3)	0.70 (0.0)	0.00	4065.74 (9.2) [†]
	gte	0.90 (0.0)	1.00	1.01 (0.0)	0.68 (0.0)	1.00	1.00 (0.0)	0.86 (0.0)	1.00	1.00 (0.0) [†]
	ste.	0.78 (0.0)	1.00	1.00 (0.0)	0.38 (0.0)	1.00	1.00 (0.0)	0.68 (0.0)	1.00	1.00 (0.0) [†]
med	leg	0.17	0.00	508.56	—	—	—	—	—	—
leg	med	0.20	0.00	511.46	—	—	—	—	—	—

Table 2: Distribution translations: vec2vecs trained on NQ and evaluated on a 65536 text subset of NQ (chunked in batches of size 8192)

2. Error analysis reveals systematic incoherence, with $\geq 90\%$ of translations producing semantically unrelated outputs (e.g., clinical terms mapped to legal procedures).
3. Geometric fragmentation is visually unambiguous, as shown by t-SNE projections revealing no linear relationship between domains.

5 Implications for AI Safety

- Threat Model Refinement:

Our results prove that domain-specialized systems are inherently resilient to this attack vector, confining high-risk scenarios to settings with overlapping training data.

- Provable Mitigation Strategy:

Domain specialization creates structural incompatibility in embedding spaces, establishing it as a viable defense for sensitive applications (e.g., healthcare/law).

6 Limitations and Future Work

While this study focused on BERT-based models, future work should:

- Test architectures with cross-domain attention (e.g., Mixture-of-Experts)
- Explore gradient-based attacks to bypass geometric barriers
- Define divergence metrics (e.g., semantic KL-divergence) to predict translation failure.

7 References

- [1]: Jha et al. (2025), "Harnessing the Universal Geometry of Embeddings"
- [2]: Wentworth, J. (2021). "Natural Abstractions: Key Claims, Evidence, and Implications." Alignment Research Center.
- [3]: Alsentzer, E. et al. (2019). "Publicly Available Clinical BERT Embeddings"
- [4]: Chalkidis, I. et al. (2020). "LEGAL-BERT: The Muppets straight out of Law School"