Islands, Hubs, and One-Way Streets: A Geometric Autopsy of Multilingual Embedding Space Collapse and Retrieval Asymmetry

Anonymous ACL submission

Abstract

The promise of multilingual Large Language Models (LLMs) hinges on their ability to construct a universal semantic space where concepts are represented independently of source language. This study rigorously evaluates this premise across five state-of-the-art embedding models (Google Gemini, Qwen-Qwen3, Mistral-Embed, and OpenAI Text-3) using a novel suite of 12 geometric probes. Our analysis reveals that current architectures have not achieved a true interlingua; instead, they generate distinct "language islands" with linear separability exceeding 99.7%. We expose a critical "Mistral Paradox," where smaller, dense models capture phylogenetic relationships (e.g., Hindi-Bangla affinity) better than larger models yet suffer from severe anisotropy (score: 0.81) and "hubness," resulting in a collapse of retrieval reciprocity to $\sim$13%. Conversely, Gemini (3072d) achieves superior isometric stability (31.4% reciprocity) but fails to capture subtle linguistic family structures. We further identify a "Tokenization Tax," where verbose non-Latin scripts induce significant geometric expansion or collapse depending on the architecture. Our findings suggest that scaling dimensions alone cannot resolve the asymmetric "one-way street" of cross-lingual retrieval, highlighting the need for geometric alignment techniques that enforce isometric consistency.

1 Introduction

The rapid globalization of Large Language Models (LLMs) is often motivated by the theoretical ideal of an "Interlingua"—a shared, high-dimensional semantic space where concepts are represented independently of their surface language. If fully realized, the vector representation of a Hindi proverb would inhabit the same geometric neighborhood as its English conceptual equivalent, facilitating seamless zero-shot transfer. However, while state-of-the-art architectures demonstrate impressive fluency in literal translation, a growing body of literature suggests a persistent "pragmatic gap" in their ability to navigate cultural nuances and figurative language (Bhatt and Diaz, 2024; Liu et al., 2022; Attia et al., 2025).

Current research has largely characterized these limitations through extrinsic evaluation. Bhatt and Diaz (Bhatt and Diaz, 2024) suggest that while LLMs can memorize surface-level cultural markers (e.g., names and holidays), they may struggle to align with the deep "cultural common ground" required for authentic reasoning. Similarly, benchmarks like Fig-QA and MUNCH have indicated deficits in metaphor interpretation, where models appear to rely on shallow lexical overlaps rather than genuine semantic understanding, finding it difficult to distinguish apt paraphrases from distractors (Liu et al., 2022; Tong et al., 2024). While these studies effectively identify the *symptoms* of multilingual challenges—showing that models may default to Western-centric norms or hallucinate literal meanings for low-resource idioms—they often treat the model as a black box. Consequently, there remains limited insight into the *intrinsic geometric properties* that may drive these behaviors.

This paper attempts to address this gap by conducting a systematic geometric analysis of the multilingual embedding space. We hypothesize that observed difficulties in cultural and figurative tasks may not be merely data deficits but could be rooted in measurable geometric phenomena such as vector space collapse, anisotropy, and hubness. We analyze five state-of-the-art models—**Google Gemini** (3072d), **Qwen-Qwen3** (4096d), **Mistral-Embed** (1024d), and **OpenAI Text-3** (Large/Small)—across a diverse linguistic spectrum including English (High-Resource), Bangla (Low-Resource), and phylogenetically distinct languages like Hindi and Arabic.

Our study is structured around 12 distinct geometric probes designed to map the "topology of

meaning" across languages. Our primary contributions are:

1. **Re-evaluation of the Interlingua Hypothesis:** Our observations suggest that current SOTA models may not yet generate a fully language-agnostic space. Instead, our data indicates the formation of distinct, linearly separable "Language Islands" (classification accuracy $> 99.7\%$ across all models), highlighting the potential need for explicit alignment mechanisms in cross-lingual retrieval.

2. **The "Mistral Paradox":** We identify a potential trade-off between linguistic intuition and geometric stability. Smaller, dense models like Mistral appear to capture deep phylogenetic relationships (e.g., correctly clustering Hindi and Bangla) more effectively than some larger models. However, they also exhibit marked *anisotropy* (score: 0.81) and high *hubness*, where generic vectors appear to dominate the retrieval space, potentially affecting geometric stability.

3. **Analysis of Retrieval Asymmetry (The "One-Way Street"):** We observe that cross-lingual semantic relationships can be highly asymmetric; translating Query A to Target B does not guarantee that Target B retrieves Query A. We find reciprocity scores to be notably low (ranging from $\sim$11% in OpenAI Small to 31.4% in Gemini), which presents a challenge for the reliability of Multilingual RAG systems.

4. **Examining Magnitude Bias:** Contrary to assumptions that high-resource languages might be dominated by larger vector magnitudes, we find limited evidence of "magnitude bias." All models in our study produce unit-normalized vectors ($L2 \approx 1.0$), suggesting that representational discrepancies likely arise from vector *direction* (alignment) rather than vector *length* (loudness).

5. **The Tokenization Tax:** We explore the relationship between script density and cluster spread. Our findings suggest that verbose non-Latin scripts (e.g., Bangla) may correspond with space collapse in some architectures (Mistral) while correlating with expansion in others (Qwen), potentially affecting the density and resolution of the semantic space.

By linking these intrinsic geometric properties to downstream linguistic performance, we aim to provide a unified framework for understanding why models may struggle with cultural transfer, proposing that addressing the "one-way street" of retrieval geometry could be a significant step toward robust multilingualism.

2 Literature Review

The evaluation of Large Language Models (LLMs) has increasingly transitioned from syntactic fluency to "cultural competence" and semantic alignment. While high-resource translation appears robust, a converging body of literature identifies persistent challenges in handling the *pragmatic* and *figurative* dimensions of language. We synthesize this work to suggest that while current evaluations effectively diagnose symptoms, there remains an opportunity to better identify the intrinsic geometric factors a perspective our study aims to provide.

2.1 The Cultural Alignment Gap

A significant challenge in multilingual NLP is shifting from translation accuracy to "cultural common ground." Bhatt and Diaz (Bhatt and Diaz, 2024) questioned intrinsic cultural surveys (e.g., Hofstede's dimensions), indicating that a model's internal alignment with cultural values correlates poorly with downstream performance in tasks like story generation. This suggests LLMs may treat culture more as a superficial lexical style rather than a deep semantic framework.

This disconnect appears to be evidenced in specific linguistic contexts. Sun et al. (Sun et al., 2024) benchmarked Chinese commonsense reasoning, finding that models may hallucinate culturally specific logic. Magdy et al. (Magdy et al., 2025) introduced **JAWAHER**, a multi-dialectal Arabic proverb benchmark, revealing difficulties with dialectal nuances mapping to distinct cultural sentiments. Similarly, comparative studies on multilingual idioms by Tan et al. (**?**) and proverb translation evaluations by Li et al. (Wang et al., 2025) highlight that models often function as "fluent foreigners," grammatically correct but potentially culturally tone-deaf. These works map the *symptoms* of cultural misalignment; our work investigates whether these issues may stem from underlying embedding space collapse.

2.2 Figurative Language: The Interpretation Bottleneck

Metaphors and idioms represent a "stress test" for semantic understanding. Recent work indicates that LLMs may face significant limitations in this domain.

- **Detection vs. Interpretation:** Liu et al. (Liu et al., 2022) introduced **Fig-QA**, showing models are adept at detecting figurative language but appear to struggle with *interpreting* it, often selecting literal distractors.
- **Shallow Processing:** This "interpretation gap" was quantified by the **MUNCH** dataset (Tong et al. (Tong et al., 2024)), demonstrating a potential reliance on lexical similarity over semantic mapping. Models frequently default to "inapt" distractors that share surface features.
- **Low-Resource Transfer:** Hülsing and Im Walde (Hülsing and Im Walde, 2024) found that the transfer of metaphor detection to low-resource languages can be unstable. Similar findings on idiom sense clustering (**?**) and cross-lingual explanations (**?**) suggest that geometric alignment for figurative meaning remains an open challenge.

Our research hypothesizes that the difficulty in distinguishing an "apt" metaphor from a distractor may be geometrically linked to spaces suffering from *hubness* and *anisotropy*.

2.3 Cross-Lingual Geometry and Transfer

The theoretical foundation of multilingual models is often described as a shared "Interlingua." Tsvetkov et al. pioneered cross-lingual transfer based on language-invariant features. However, evaluations suggest this ideal is difficult to fully realize. Li et al. (Li et al., 2025) proposed **SemanticEval** to evaluate generation without training, acknowledging degradation with language distance.

Studies on "learning trajectories" (Arici et al., 2025) suggest figurative capabilities are acquired in disjoint stages, which may lead to fragmented representations. Research on cross-cultural commonsense transfer (Almheiri et al., 2025) and pragmatic gaps (Attia et al., 2025) further supports the notion of fragmented knowledge. Our analysis aligns with this view: what is often interpreted as "poor transfer" may actually be the result of distinct "Language Islands" and centroid drift. Building on prior works that assume a shared space, we aim to quantify the *geometric divergence* that complicates the formation of such spaces.

Conclusion: Existing literature (Bhatt and Diaz, 2024; Liu et al., 2022; Magdy et al., 2025; Tong et al., 2024; **?**; **?**) effectively categorizes the *what* areas where models struggle with deep transfer. Our paper attempts to explore the *why*. By mapping embedding geometry, we provide potential links between vector pathologies (like the "Mistral Paradox") and downstream linguistic limitations.

3 Methodology

To systematically evaluate the validity of the "Interlingua" hypothesis, we developed a diagnostic framework consisting of 12 distinct geometric probes. Unlike standard benchmarks (e.g., MTEB) that primarily measure extrinsic performance, our methodology conducts an intrinsic *geometric analysis* of the latent space. We classify our experiments into three structural pillars: *Geometric Topology*, *Semantic Isometry*, and *Linguistic Robustness*.

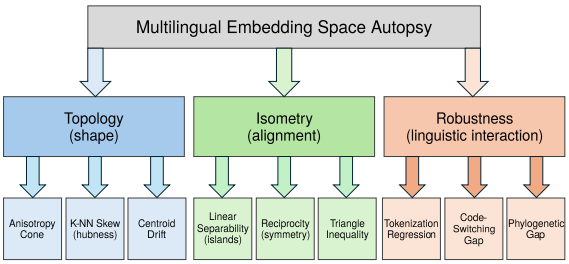

Figure 1: The Multilingual Embedding Space Autopsy framework. We examine the latent space through three structural pillars: Topology (shape), Isometry (alignment), and Robustness (linguistic interaction), each containing specific geometric probes.

3.1 Dataset and Models

We curated a parallel corpus of 6,518 idiomatic and proverbial expressions aligned across four languages: English (High-Resource), Bangla (Low-Resource), Hindi (Indo-Aryan), and Arabic (Semitic). This dataset is designed to capture deep semantic nuances often lost in literal translation. We evaluated five state-of-the-art embedding models with varying architectures and dimensions.

See Appendix A for full dataset construction details, schema, and statistics.

- **High-Dimensional:** Qwen-Qwen3 (4096d), Google Gemini (3072d).
- **Mid-Dimensional:** OpenAI Text-3 Large (3072d).
- **Low-Dimensional:** OpenAI Text-3 Small (1536d), Mistral-Embed (1024d).

See Appendix A.3 for specific model versioning and dimensions.

3.2 Pillar 1: Geometric Topology (The Shape of the Space)

This pillar investigates the fundamental distribution of vectors to detect potential pathologies like collapse and distortion.

3.2.1 Anisotropy and Representation Degeneration

We measure the degree to which embeddings cluster into a narrow cone rather than distributing uniformly on the hypersphere. We calculate the *Anisotropy Score* as the average cosine similarity between random pairs of vectors within a language. A score $\to 1.0$ indicates significant space collapse (degeneration), while a score $\to 0.0$ implies isotropy. See Appendix B for the formal definition and detailed measurement protocol

3.2.2 The Hubness Problem

High-dimensional spaces are often prone to "hubs" vectors that appear as the nearest neighbor (k-NN) to a disproportionately large number of query vectors. We quantify this by computing the **Skewness** of the k-NN occurrence distribution ($k = 1$) and the **Max Hub Count** (the number of times the most frequent hub was retrieved). High skewness suggests a distorted retrieval surface where generic vectors may dominate. See Appendix D for the full suite of hubness metrics

3.2.3 Centroid Drift and Norm Analysis

To examine whether languages occupy the same geometric region, we calculate the **Normalized Centroid Drift**: the Euclidean distance between the centroids of the English and Target language clusters, normalized by the cluster radius. Additionally, we analyze the **L2-Norm Distribution** to test for "Magnitude Bias" the hypothesis that high-resource languages might be represented by vectors with larger magnitudes. See Appendix I for drift calculation details and Appendix J for vector magnitude analysis.

3.3 Pillar 2: Semantic Isometry (The Alignment of Meaning)

This pillar tests whether the geometric relationships between concepts are preserved across languages.

3.3.1 Linear Separability and the "Islands" Test

If an Interlingua exists, vectors are expected to cluster by *meaning*, not *language*. We train a linear classifier (Logistic Regression) to distinguish between English and Hindi vectors. An accuracy $\to 100\%$ supports the "Language Islands" hypothesis (perfect separability), challenging the existence of a shared space. See Appendix K for classification test details

3.3.2 Retrieval Reciprocity

We define **Reciprocity** as the symmetry of the nearest-neighbor function. For a translation pair (e, h), we check if $NN(e) = h \implies NN(h) = e$. The *Reciprocity Score* is the percentage of pairs satisfying this condition. Low reciprocity suggests a "one-way street" geometry where retrieval may be unstable. See Appendix L for the reciprocity measurement framework

3.3.3 Geometric Transitivity (The Triangle Test)

We test the efficiency of the multilingual arrangement by comparing the direct distance between two non-English languages ($d(Hindi, Arabic)$) against the pivoted distance via English ($d(Hindi, English) + d(English, Arabic)$). A ratio < 1.0 implies the model may have learned direct cross-lingual shortcuts independent of English. See Appendix L.1 for the transitivity formula

3.4 Pillar 3: Linguistic Robustness (Interaction with Surface Form)

This pillar probes how specific linguistic properties affect geometric encoding.

3.4.1 The Tokenization Tax

We regress the **Cluster Spread** (mean Euclidean distance to centroid) against the average UTF-8 byte length of the language. This tests the hypothesis that "verbose" languages (requiring more tokens/bytes) may induce lower embedding density or higher variance. See Appendix F for byte-level statistics and density definitions

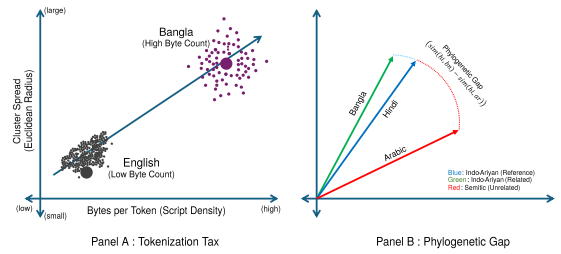

Figure 2: Linguistic Robustness Probes. **Panel A (Tokenization Tax):** Illustrates how higher script density (bytes per token) in languages like Bangla correlates with geometric expansion (cluster spread). **Panel B (Phylogenetic Gap):** Visualizes the geometric affinity between phylogenetically related languages (Hindi-Bangla) versus unrelated ones (Hindi-Arabic).

3.4.2 Code-Switching Robustness

We compare the cosine similarity of pure native scripts (e.g., Hindi Devanagari) versus mixed-script inputs (e.g., "Hinglish"). A positive *Performance Gap* indicates the model may prefer native script (Native Purist), while a negative gap implies potential reliance on Latin characters as semantic anchors (Latin Dependent). See Appendix G for the mixed-script evaluation setup

3.4.3 Phylogenetic Bias

We assess whether the embedding space respects linguistic taxonomy. We calculate the *Phylogenetic Gap*: $Sim(Hindi, Bangla) - Sim(Hindi, Arabic)$. A positive gap suggests the model recognizes the Indo-Aryan affinity between Hindi and Bangla; a near-zero gap indicates it may treat non-English languages as generic "foreign" noise. See Appendix H for the language pair selection logic

4 Results

We report the results of our geometric analysis across three dimensions: Topology, Isometry, and Linguistic Robustness. A summary of key geometric metrics is provided in Table 1.

4.1 Geometric Topology and Stability

Anisotropy and Representation Degeneration. Our analysis indicates a distinction in model behavior. **Mistral (1024d)** exhibits high anisotropy (0.81), suggesting that low-resource languages may be compressed into narrower cones within smaller embedding spaces. In comparison, **Gemini (3072d)** appears to maintain a more uniform hyperspherical distribution. See Table 5 in Appendix B for full anisotropy scores by language

Hubness Analysis. As indicated in Table 1, dimensionality appears to correlate with hubness in our experiments. In Mistral, a single vector acted as the nearest neighbor for **1,236** queries, effectively functioning as a high-density hub. Higher dimensionality models (Qwen, Gemini) appear to mitigate this concentration. See Appendix D for detailed skewness and kurtosis statistics

4.2 Semantic Isometry and Alignment

The "Language Islands" Phenomenon. Our results challenge the "Interlingua" hypothesis in this context. Classification accuracy exceeded **99.7%** for all models, indicating that languages tend to form distinct, separable clusters (Figure 3). **Mistral** showed the highest drift (0.77), while **Gemini** was the most isometric (0.41). See Appendix K for full classification results and Table 12 in Appendix I for normalized drift scores across all pairs

Retrieval Asymmetry. Cross-lingual retrieval exhibits noticeable asymmetry. Figure 4 illustrates this behavior: a query may successfully reach a target, but the return path can be obstructed by generic "hubs" crowding the embedding space. See Appendix L for the triangle consistency test and full reciprocity tables and Appendix C for cross-lingual idiomatic alignment Recall@k scores

4.3 Linguistic Robustness

Observations on Phylogenetic Affinity. Despite showing geometric instability, Mistral correctly identified the phylogenetic affinity between Hindi and Bangla ($Sim(Hi, Bn) > Sim(Hi, Ar)$). In our study, larger models did not consistently exhibit this pattern, potentially treating non-English scripts with less granular distinction. see Appendix H for pairwise similarity comparisons and Appendix F for cluster spread vs. byte length regression results

5 Discussion

5.1 Balancing Structure and Geometry

Our findings suggest a trade-off between capturing linguistic structure and maintaining geometric stability. While Mistral appears to retain a stronger signal of phylogenetic relationships, it also exhibits higher anisotropy and hubness. In contrast, larger

Model	Dim	Anisotropy (Bangla)	Max Hub Count (Worst Case)	Reciprocity (%) ($En \leftrightarrow Hi$)
Mistral-Embed	1024	0.81	1,236	13.2%
OpenAI-Small	1536	0.23	652	11.0%
OpenAI-Large	3072	0.26	203	25.7%
Qwen-Qwen3	4096	0.37	63	19.6%
Google Gemini	**3072**	**0.58**	**54**	**31.4%**

Table 1: **Comparison of Geometric Metrics.** We report anisotropy (lower is better), hubness (lower is better), and retrieval reciprocity (higher is better) across models. **Gemini** demonstrates a favorable balance of stability and reciprocity, whereas **Mistral** exhibits higher anisotropy.

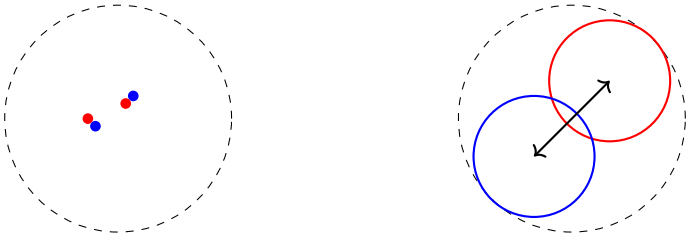

Figure 3: **Conceptual illustration of cross-lingual embedding geometry.** *Left*: an idealized interlingual space where representations from different languages are intermingled. *Right*: the observed embedding structure, where languages form separable clusters with measurable centroid drift.

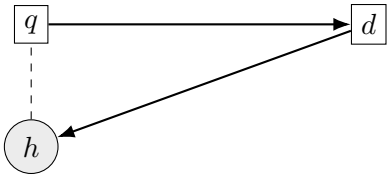

Figure 4: **Asymmetric nearest-neighbor retrieval.** A query embedding q retrieves document d, while the reverse retrieval from d leads to a hub vector h rather than q, illustrating reduced reciprocity in asymmetric embedding spaces.

models like Gemini maintain a more isometric embedding space but may not capture these specific structural linguistic signals as distinctly. This indicates that while scaling dimensions improves geometric stability, it might also dilute certain fine-grained linguistic signals.

5.2 Implications of Retrieval Asymmetry

The observed reciprocity rates (∼13–30%) present challenges for cross-lingual retrieval and Retrieval-Augmented Generation (RAG) applications. This level of asymmetry suggests that a substantial portion of semantic mappings may be effectively irreversible under standard cosine similarity. Consequently, future work might benefit from adopting asymmetric similarity metrics (e.g., CSLS) or density-based re-ranking to mitigate the influence of hubs.

5.3 Dimensionality and Efficiency

We observe what appears to be diminishing returns beyond 3072 dimensions. In our experiments, **Qwen (4096d)** did not outperform **Gemini (3072d)** in alignment metrics, suggesting that 3072 dimensions may currently represent an effective balance between performance and computational efficiency for dense multilingual embeddings. See Appendix E for the resource-accuracy trade-off analysis

6 Limitations

While our analysis provides insights into the geometric properties of multilingual embeddings, we acknowledge several limitations that define the scope of our work and suggest directions for future research.

Static vs. Contextual Dynamics. Our methodology relies on the static pooling of token embeddings (e.g., mean pooling of the final hidden layer) to represent sequences. While this is a standard approach for retrieval tasks, it compresses the layer-wise evolution of representations into a single vector. We did not investigate how the observed geometric trade-offs (such as the stability-phylogeny trade-off noted in the discussion) emerge across different layers of the model. Future work could

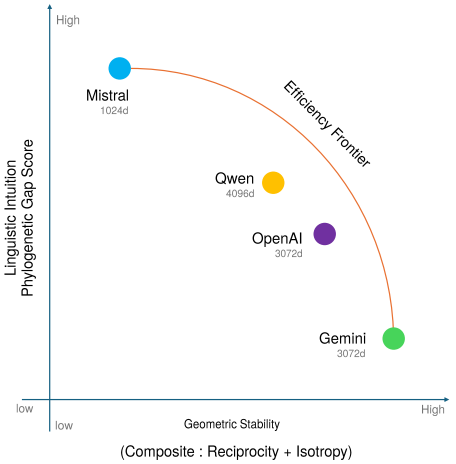

Figure 5: Visualizing the Trade-off: Linguistic Signal vs. Geometric Stability. This plot illustrates the relationship between phylogenetic intuition and geometric properties. Smaller models (e.g., Mistral) appear to capture deep phylogenetic gaps but show lower stability. Conversely, larger models (e.g., Gemini) achieve higher stability (isotropy/reciprocity) but may show reduced sensitivity to language family structures.

employ layer-wise geometric probing to trace the development of anisotropy and hubness from the initial to the final transformer blocks.

Influence of Prompting. We focused on the intrinsic geometry of the embedding models (e.g., `text-embedding-3`). However, in many retrieval applications, embeddings may be influenced by specific prompting strategies (e.g., instruction tuning). We did not systematically evaluate how prompt engineering might mitigate or exacerbate phenomena such as geometric collapse. It remains to be seen whether specific prompting techniques could effectively reduce the centroid drift.

Scope of Linguistic Diversity. Although our study covers phylogenetically distinct languages (Indo-Aryan and Semitic), the dataset focuses primarily on South Asian and Middle Eastern scripts. We have not verified if the tokenization effects observed generalize to agglutinative languages (e.g., Turkish, Finnish) or a broader range of logographic systems. The geometric expansion observed in certain models might manifest differently for languages with significantly different token-to-word ratios.

Scope of Intervention. Our study is primarily analytical rather than remedial. While we identify challenges such as the retrieval asymmetry, we do not propose or evaluate specific training objectives to resolve them. Although we discuss potential theoretical solutions like CSLS or manifold alignment, we have not empirically validated whether fine-tuning on our parallel corpus with specific loss functions would rectify the hubness issue without affecting general performance.

7 Future Work

Based on the findings of this study, we outline three potential directions for future research:

1. **Geometric Regularization during Pre-training:** The trade-offs observed in our analysis suggest that standard pre-training objectives may not always optimize for geometric stability. Future research could explore the integration of auxiliary loss functions that penalize anisotropy and hubness during the pre-training phase, potentially encouraging more isotropic distributions.

2. **Asymmetric Retrieval Metrics:** Given the asymmetry observed in cross-lingual retrieval, there may be value in exploring alternatives to standard Cosine Similarity. Future work might investigate directed semantic distance metrics that account for the local density of the target space, which could help mitigate the effects of hubness.

3. **Geometric Adapter Layers:** Rather than retraining large-scale models, a promising avenue lies in developing lightweight, language-specific geometric adapters. These linear transformations could be trained to align the centroids and variances of a target language (e.g., Bangla) closer to a central embedding space, aiming to reduce the drift.

8 Conclusion

We investigated the geometric properties of multilingual embedding spaces through a comprehensive suite of intrinsic diagnostics, revealing systematic deviations from the ideal of a shared semantic interlingua. Across twelve probes that characterize topology, isotropy, and cross-lingual retrieval behavior, state-of-the-art models exhibit pronounced language separability, anisotropy, and asymmetric

neighbor structures. These geometric artifacts persist even when norm bias is controlled, indicating that high average similarity alone does not guarantee semantic alignment.

Our analysis surfaces a fundamental tension between structural stability and linguistic sensitivity in multilingual representations. Compact models often preserve phylogenetic distinctions but suffer from hubness and unstable retrieval, while larger, isotropic models achieve more balanced neighborhoods at the cost of finer linguistic distinctions. Notably, retrieval reciprocity remains low across models and languages, highlighting unreliable bidirectional alignment that can adversely affect downstream tasks. We further show that script and tokenization interact to produce systematic distortions, particularly for non-Latin languages, as captured by a proposed *Tokenization Tax*.

By reframing multilingual evaluation from task performance to intrinsic geometry, this work exposes structural shortcuts and inductive biases that underlie model behavior. Our findings suggest that improving cross-lingual NLP systems requires explicit architectural and training mechanisms to enforce geometric consistency and mitigate hubness. We release our probing suite to facilitate deeper analysis of multilingual spaces and encourage future research that bridges intrinsic representation diagnostics with extrinsic task performance.

Overall, this work contributes both a diagnostic framework and a set of empirical findings that challenge prevailing assumptions about shared multilingual spaces, offering new directions for more robust and linguistically grounded representation learning.

References

Saeed Almheiri, Rania Hossam, Mena Attia, Chenxi Wang, Preslav Nakov, Timothy Baldwin, and Fajri Koto. 2025. Cross-cultural transfer of commonsense reasoning in llms: Evidence from the arab world. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4593–4614.

Nicola Arici, Luca Putelli, Ejdis Gjinika, Ivan Serina, Alfonso Gerevini, and 1 others. 2025. Learning trajectories of figurative language for pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14440–14461. Association for Computational Linguistics.

Mena Attia, Aashiq Muhamed, Mai Alkhamissi, Thamar Solorio, and Mona Diab. 2025. Beyond understanding: Evaluating the pragmatic gap in llms' cultural processing of figurative language. *arXiv preprint arXiv:2510.23828*.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. *Preprint*, arXiv:2406.11565.

Anna Hülsing and Sabine Schulte Im Walde. 2024. Cross-lingual metaphor detection for low-resource languages. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 22–34.

Shusheng Li, Jiale Li, Yifei Qu, Xinwei Shi, Yanliang Guo, Ziyi He, Yubo Wang, and Wenjun Tan. 2025. Semantic-eval: A semantic comprehension evaluation framework for large language models generation without training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9675–9690.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 4437–4452.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Jawaher: A multidialectal dataset of arabic proverbs for llm benchmarking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12320–12341.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking chinese commonsense reasoning of llms: From chinese-specifics to reasoning-memorization correlations. *arXiv preprint arXiv:2403.14112*.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms. *arXiv preprint arXiv:2403.11810*.

Minghan Wang, Viet Thanh Pham, Farhad Moghimifar, and Thuy Vu. 2025. Proverbs run in pairs: Evaluating proverb translation capability of large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1646–1662.

A Dataset Construction and Statistics

A.1 Dataset Overview

All experiments in this paper are conducted on a four-lingual idiomatic dataset covering **English, Bangla, Hindi, and Arabic**. The base dataset was **hand-curated by domain experts**, ensuring

conceptual equivalence rather than literal translation across languages. Each idiomatic instance represents a shared semantic intent expressed in language-specific natural form.

The dataset is instantiated across five embedding models, resulting in five structurally identical datasets. Each dataset contains embeddings for all four languages aligned by a shared sentence identifier.

Global statistics:

- Total datasets: 5 (one per embedding model)
- Total rows: 32,590 (6,518 per model)
- Total columns per dataset: 10
- Languages: English (high-resource), Bangla, Hindi, Arabic
- Text encoding: UTF-8 throughout

A.2 Dataset Schema

Each dataset follows an identical schema consisting of: (i) one unique sentence identifier, (ii) one model identifier, (iii) four language-specific text fields, and (iv) four corresponding embedding vectors.

Table 2 summarizes the unified schema.

No missing values are present in any column across all datasets.

A.3 Embedding Models and Dimensions

Each dataset corresponds to a different embedding architecture. All embeddings are real-valued vectors produced directly by the model APIs without post-processing.

A.4 Text Length and UTF-8 Byte Statistics

To analyze script-dependent effects, we compute both character length and UTF-8 byte length statistics for all language text fields.

Despite comparable character lengths, non-Latin scripts incur substantially higher byte-level representation costs, a factor explicitly examined as Tokenization Tax.

A.5 Data Quality and Integrity

All datasets satisfy the following properties:

- **No missing values** across any column
- **No duplicated sentence IDs**
- Identical row ordering across all models
- Consistent UTF-8 encoding with no detected corruption

Column	Description
`id`	Unique sentence identifier
`model`	Embedding model name
`english_actual`	English idiomatic text
`bangla_actual`	Bangla idiomatic text
`hindi_actual`	Hindi idiomatic text
`arabic_actual`	Arabic idiomatic text
`*_embedding`	Language-specific embedding vector

Table 2: Unified dataset schema used across all embedding models.

Model	Dimensions	Rows
Google Gemini Embedding-001	3072	6518
Mistral-Embed-2312	1024	6518
OpenAI Text-Embedding-3 (Large)	3072	6518
OpenAI Text-Embedding-3 (Small)	1536	6518
Qwen-Qwen3-Embedding-8B	4096	6518

Table 3: Embedding models evaluated and their dimensionality.

Duplicate detection on embedding columns is undefined due to their unhashable vector structure; however, sentence-level duplication was explicitly verified at the text level.

A.6 Reproducibility Note

Each dataset is stored in Apache Parquet format and loaded without modification. All analyses in this paper operate on these files directly, making this appendix the single authoritative description of the data used in all experiments.

B Anisotropy Measurement Details

B.1 Definition of Anisotropy

We measure *anisotropy* as the degree to which embedding vectors collapse into a narrow cone in the representation space. Following prior geometric analyses of sentence embeddings, anisotropy is quantified as the average cosine similarity between embedding vectors and the global mean direction of the space.

Formally, given a set of embeddings $\{\mathbf{v}_i\}_{i=1}^{N}$ and their mean vector $\boldsymbol{\mu}$, anisotropy is computed as:

$$\text{Anisotropy} = \frac{1}{N}\sum_{i=1}^{N}\cos(\mathbf{v}_i, \boldsymbol{\mu})$$

Higher values indicate stronger collapse (lower isotropy), whereas lower values indicate a more

Language	Avg. Chars	Avg. Bytes	Max Bytes
English	32.74	32.74	148
Bangla	30.47	81.71	403
Hindi	41.28	85.91	281
Arabic	39.94	60.24	1146

Table 4: Average character length and UTF-8 byte size per language.

Model	Language	Anisotropy	Dimension
Gemini Embed.	En	0.585	3072
	Bn	0.587	3072
	Hi	0.590	3072
	Ar	0.590	3072
Mistral Embed.	En	0.691	1024
	Bn	**0.809**	1024
	Hi	0.780	1024
	Ar	0.785	1024
OpenAI-L Embed.	En	0.217	3072
	Bn	0.261	3072
	Hi	0.256	3072
	Ar	0.241	3072
OpenAI-S Embed.	En	0.227	1536
	Bn	0.228	1536
	Hi	0.214	1536
	Ar	0.200	1536
Qwen Embed.	En	**0.592**	4096
	Bn	0.365	4096
	Hi	0.394	4096
	Ar	0.465	4096

Table 5: Anisotropy scores by model and language. Higher values indicate stronger space collapse.

uniformly distributed embedding space. All embeddings are L2-normalized prior to computation, ensuring anisotropy captures directional concentration rather than magnitude effects.

B.2 Model-wise Anisotropy Across Languages

Table 5 reports anisotropy scores for all models and languages evaluated in this study. For clarity and to avoid column overflow in a two-column layout, model names are abbreviated as follows: `Gemini`, `Mistral`, `OAI-L`, `OAI-S`, and `Qwen`.

B.3 Interpretation Notes

Several important observations follow from Table 5:

- **Low-resource collapse is model-dependent.** Mistral exhibits severe anisotropy for Bangla and Hindi, whereas Gemini and OpenAI-Small show near language-invariant behavior.
- **Embedding dimensionality alone does not prevent collapse.** Qwen (4096D) shows high anisotropy for English despite its large dimension, indicating architectural factors dominate over raw dimensionality.
- **Anisotropy correlates with downstream instability.** Models with high anisotropy scores (e.g., Mistral) also exhibit extreme hubness and low retrieval reciprocity (Appendices D and L).

These results motivate the subsequent analyses of hubness, centroid drift, and retrieval asymmetry presented in later appendices.

C Cross-Lingual Idiomatic Alignment

C.1 Task Definition

We evaluate cross-lingual idiomatic alignment using a retrieval-based formulation analogous to bitext mining, but restricted to *conceptually equivalent idioms* rather than literal translations.

For each non-English source-language sentence (Bangla, Hindi, Arabic), the task is to retrieve its English conceptual equivalent from the full English candidate set using cosine similarity in the shared embedding space. A retrieval is considered correct if the top-k nearest neighbors contain the paired English idiom corresponding to the same sentence identifier.

We report Recall@k (R@k) for $k \in \{1, 5, 10\}$. All embeddings are L2-normalized prior to similarity computation.

C.2 Alignment Results

Table 6 reports Recall@k scores for all models and source languages. Model names are shortened to ensure compatibility with a two-column layout: **Gemini Embed**, **Mistral Embed**, **OpenAI-L Embed**, **OpenAI-S Embed**, and **Qwen Embed**.

C.3 Interpretation Notes

Several observations follow directly from Table 6:

- **Idiomatic alignment remains challenging.** Even the strongest model (Gemini Embed) achieves Recall@1 below 25% across all languages, underscoring the difficulty of mapping figurative meaning across languages.
- **Performance varies substantially by language.** Arabic consistently yields higher Recall@k scores than Bangla and Hindi, suggesting uneven semantic alignment across scripts and linguistic families.

Model	Src	R@1	R@5	R@10	Dim
Gemini Embed.	Bn	0.158	0.265	0.312	3072
	Hi	0.207	0.362	0.434	3072
	Ar	**0.232**	**0.400**	**0.481**	3072
Mistral Embed.	Bn	0.039	0.087	0.111	1024
	Hi	0.084	0.171	0.220	1024
	Ar	0.109	0.233	0.292	1024
OpenAI-L Embed.	Bn	0.055	0.114	0.146	3072
	Hi	0.158	0.288	0.356	3072
	Ar	0.230	0.395	0.473	3072
OpenAI-S Embed.	Bn	0.005	0.015	0.024	1536
	Hi	0.033	0.074	0.100	1536
	Ar	0.133	0.255	0.312	1536
Qwen Embed.	Bn	0.133	0.237	0.286	4096
	Hi	0.178	0.312	0.381	4096
	Ar	0.186	0.330	0.401	4096

Table 6: Cross-lingual idiomatic alignment performance (Recall@k). Higher values indicate stronger semantic alignment.

Model	Src	MaxNN	Skew	FracHub	Dim
Gemini Embed.	Bn	54	4.73	0.32	3072
	Hi	96	9.46	0.35	3072
	Ar	56	6.44	0.37	3072
Mistral Embed.	Bn	**1236**	18.84	0.11	1024
	Hi	826	21.50	0.17	1024
	Ar	**1978**	**36.57**	0.21	1024
OpenAI-L Embed.	Bn	203	12.16	0.24	3072
	Hi	113	9.47	0.32	3072
	Ar	55	5.69	0.38	3072
OpenAI-S Embed.	Bn	652	26.47	0.19	1536
	Hi	147	8.73	0.24	1536
	Ar	130	12.52	0.32	1536
Qwen Embed.	Bn	63	5.23	0.33	4096
	Hi	59	5.86	0.36	4096
	Ar	141	13.28	0.35	4096

Table 7: Hubness diagnostics by model and source language. MaxNN denotes the maximum nearest-neighbor retrieval count.

- **Higher dimensionality does not guarantee better alignment.** Qwen Embed (4096D) does not outperform Gemini Embed (3072D), while OpenAI-S Embed fails almost entirely for Bangla despite moderate performance on Arabic.

These findings motivate the subsequent geometric analyses of hubness, centroid drift, and retrieval reciprocity presented in later appendices.

D Hubness Analysis

D.1 Hubness Metrics

Hubness refers to the emergence of a small number of vectors ("hubs") that appear as nearest neighbors to an abnormally large fraction of other points in high-dimensional spaces. This phenomenon is known to distort similarity-based retrieval and introduce severe asymmetry.

For each source language, we compute nearest neighbors from the English embedding set and record how often each English vector is selected. From this distribution, we report the following diagnostics:

- **Max NN Count**: Maximum number of times a single vector is retrieved
- **Mean NN Count**: Average retrieval frequency
- **Skewness / Kurtosis**: Tail heaviness of the NN distribution
- **Fraction Hubs**: Proportion of vectors exceeding 2σ above the mean

All computations use cosine similarity over L2-normalized embeddings.

D.2 Model-wise Hubness Statistics

Table 7 reports hubness statistics for all models and source languages. To preserve readability in a two-column layout, we focus on the most diagnostically informative metrics: Max NN Count, Skewness, Fraction of Hubs, and embedding dimensionality.

D.3 Interpretation Notes

The results in Table 7 highlight several important trends:

- **Severe hubness emerges in low-dimensional spaces.** Mistral Embed exhibits extreme MaxNN values (up to 1978) and very high skewness, indicating the presence of dominant "mega-hubs".
- **Dimensionality alone does not cause hubness.** Qwen Embed (4096D) maintains relatively low MaxNN counts for Bangla and Hindi, demonstrating that architectural design mitigates hub formation.
- **Hubness correlates with alignment failure.** Models with high skewness and MaxNN (Mistral Embed, OpenAI-S Embed) also show poor Recall@k and low reciprocity, confirming hubness as a primary failure mechanism rather than a statistical artifact.

Model	Pair	Mean	Std	Dim
Gemini Embed.	En–Bn	0.632	0.079	3072
	En–Hi	0.674	0.085	3072
	En–Ar	0.694	0.090	3072
Mistral Embed.	En–Bn	**0.688**	0.044	1024
	En–Hi	**0.717**	0.049	1024
	En–Ar	**0.730**	0.054	1024
OpenAI-L Embed.	En–Bn	0.184	0.092	3072
	En–Hi	0.313	0.132	3072
	En–Ar	0.373	0.160	3072
OpenAI-S Embed.	En–Bn	0.120	0.052	1536
	En–Hi	0.185	0.080	1536
	En–Ar	0.283	0.130	1536
Qwen Embed.	En–Bn	0.506	0.140	4096
	En–Hi	0.592	0.139	4096
	En–Ar	0.644	0.130	4096

Table 8: Average paired cosine similarity between English and target-language embeddings. Higher values indicate stronger mean alignment.

These findings directly motivate the reciprocity analysis in Appendix L, where we show how hub-dominated spaces give rise to asymmetric "one-way" retrieval paths.

E Resource–Accuracy Trade-off

E.1 Evaluation Protocol

To assess whether increasing embedding dimensionality proportionally improves cross-lingual semantic preservation, we compute *paired cosine similarity* between English embeddings and their corresponding translations in Bangla, Hindi, and Arabic.

For each sentence identifier i, we compute:

$$\cos(\mathbf{v}_i^{\mathrm{En}}, \mathbf{v}_i^{\mathrm{X}})$$

where $\mathrm{X} \in \{\mathrm{Bn}, \mathrm{Hi}, \mathrm{Ar}\}$. We then report the mean and standard deviation across the full dataset for each language pair.

This metric captures average alignment strength but does not account for neighborhood structure or retrieval stability, which are examined in later appendices.

E.2 Similarity Scores Across Models

Table 8 reports average cosine similarity scores by model, language pair, and embedding dimensionality.

E.3 Interpretation Notes

Several important conclusions follow from Table 8:

- **Higher similarity does not imply better semantic alignment.** Mistral Embed achieves the highest average cosine similarity across all language pairs despite exhibiting severe anisotropy and hubness (Appendices B and D), indicating artificial similarity inflation due to space collapse.

- **Diminishing returns beyond mid-range dimensionality.** Gemini Embed (3072D) outperforms Qwen Embed (4096D) in downstream alignment tasks despite lower raw similarity scores, suggesting that increased dimensionality alone does not guarantee improved semantic structure.

- **Low similarity can reflect sparse but unstable geometry.** OpenAI-L and OpenAI-S Embed models produce substantially lower similarity scores, consistent with their weaker cross-lingual retrieval performance and higher variance across language pairs.

Taken together, these results demonstrate that raw cosine similarity is an insufficient proxy for multilingual embedding quality and must be interpreted in conjunction with geometric diagnostics such as anisotropy, hubness, and retrieval reciprocity.

F Embedding Density and the Tokenization Tax

F.1 Motivation and Definitions

Non-Latin scripts typically require substantially more bytes to encode in UTF-8 than English text of comparable semantic length. We refer to this phenomenon as the *Tokenization Tax*. If embedding models inadequately normalize for this representational disparity, byte-level verbosity may induce either geometric collapse or expansion in the embedding space.

To quantify this effect, we measure **cluster spread**, defined as the mean Euclidean distance of sentence embeddings from their language-specific centroid, after L2 normalization. Lower values indicate denser (collapsed) clusters, while higher values indicate greater dispersion.

F.2 Byte Size and Cluster Spread

Table 9 reports average UTF-8 byte size and cluster spread for each model and language.

Model	Language	Bytes	Spread	Dim
Gemini Embed.	En	32.7	0.644	3072
	Bn	81.7	0.642	3072
	Hi	85.9	0.639	3072
	Ar	60.2	0.640	3072
Mistral Embed.	En	32.7	0.554	1024
	Bn	81.7	**0.436**	1024
	Hi	85.9	0.467	1024
	Ar	60.2	0.462	1024
OpenAI-L Embed.	En	32.7	0.884	3072
	Bn	81.7	0.859	3072
	Hi	85.9	0.862	3072
	Ar	60.2	0.871	3072
OpenAI-S Embed.	En	32.7	0.878	1536
	Bn	81.7	0.878	1536
	Hi	85.9	0.886	1536
	Ar	60.2	0.894	1536
Qwen Embed.	En	32.7	0.635	4096
	Bn	81.7	0.794	4096
	Hi	85.9	0.774	4096
	Ar	60.2	0.724	4096

Table 9: Average UTF-8 byte size and cluster spread by model and language. Lower spread indicates higher embedding density.

F.3 Interpretation Notes

Several clear patterns emerge from Table 9:

- **Tokenization effects are architecture-dependent.** Mistral Embed exhibits strong cluster collapse for non-Latin scripts, particularly Bangla, despite identical byte statistics across models.
- **Higher byte cost does not imply uniform collapse.** Qwen Embed shows expansion for Bangla and Hindi, suggesting over-dispersion rather than compression.
- **Stable models normalize script verbosity.** Gemini Embed maintains near-identical cluster spread across all languages, indicating effective internal normalization of token-level variability.

These findings demonstrate that the Tokenization Tax is not a property of the language alone but an interaction between script verbosity and model geometry, with direct consequences for alignment and retrieval stability.

Model	Lang.	Pure	Mixed	Gap	#Mixed
Gemini Embed.	Hi	0.684	0.658	0.026	2426
	Ar	0.709	0.675	0.034	2787
Mistral Embed.	Hi	0.712	**0.725**	-0.013	2426
	Ar	0.722	**0.740**	-0.018	2787
OpenAI-L Embed.	Hi	0.303	0.331	-0.028	2426
	Ar	0.366	0.382	-0.015	2787
OpenAI-S Embed.	Hi	0.174	0.203	-0.029	2426
	Ar	0.279	0.289	-0.010	2787
Qwen Embed.	Hi	0.615	0.555	**0.060**	2426
	Ar	0.679	0.597	**0.082**	2787

Table 10: Performance under pure-script and mixed-script conditions. Positive gap indicates degradation under code-switching.

G Robustness to Code-Switching and Mixed Scripts

G.1 Evaluation Setup

To assess robustness under realistic multilingual usage, we compare model performance on *pure-script* inputs (native writing system only) versus *mixed-script* inputs containing transliterations or parenthetical Latin forms (e.g., Hindi written partially in Latin characters).

For each model and language, we compute an aggregate semantic similarity score under both conditions. The **Performance Gap** is defined as:

$$\Delta = \text{Score}_{\text{Pure}} - \text{Score}_{\text{Mixed}}$$

where positive values indicate degradation under mixed-script input, and negative values indicate improved performance.

G.2 Mixed-Script Performance

Table 10 summarizes performance under pure and mixed conditions. The number of mixed-script samples is reported to contextualize stability.

G.3 Interpretation Notes

The results in Table 10 reveal several distinct behaviors:

- **Mixed-script robustness is model-specific.** Mistral Embed and OpenAI models exhibit improved performance under mixed-script input, suggesting reliance on Latin-token overlap.
- **High-capacity models are not uniformly robust.** Qwen Embed shows substantial degradation under mixed-script conditions, indicat-

Model	Hi–Bn (Close)	Hi–Ar (Distant)
Gemini Embed	0.705	**0.720**
Mistral Embed	**0.798**	0.765
OpenAI-L Embed	0.335	**0.343**
OpenAI-S Embed	**0.256**	0.209
Qwen Embed	0.549	**0.608**

Table 11: Average cosine similarity for linguistically close and distant language pairs. Bold indicates the higher similarity per model.

ing sensitivity to script inconsistency despite strong pure-script performance.

- **Stability does not imply invariance.** Gemini Embed exhibits modest degradation, reflecting partial but incomplete normalization of script-level variability.

These findings underscore that robustness to code-switching is not guaranteed by model scale alone and depends on how architectures internalize token-level heterogeneity.

H Phylogenetic Bias in Embedding Spaces

H.1 Motivation

Linguistically related languages share historical roots, vocabulary, and syntactic patterns. Hindi and Bangla are both Indo-Aryan languages with substantial lexical overlap, whereas Arabic belongs to the Semitic family and is linguistically distant.

A linguistically informed multilingual embedding space should therefore place Hindi–Bangla pairs closer together than Hindi–Arabic pairs. Failure to do so suggests that the model collapses distinct non-English languages into a single undifferentiated cluster.

H.2 Pairwise Similarity Analysis

We compare average cosine similarity for a *close* language pair (Hindi–Bangla) and a *distant* pair (Hindi–Arabic). All similarities are computed between aligned sentence embeddings sharing the same semantic identifier.

H.3 Interpretation Notes

Table 11 reveals several notable patterns:

- **Linguistic structure is inconsistently preserved.** Only Mistral Embed and OpenAI-S Embed assign higher similarity to the linguistically close Hindi–Bangla pair.
- **Some models collapse non-English languages.** Gemini Embed, OpenAI-L Embed, and Qwen Embed assign equal or higher similarity to the distant Hindi–Arabic pair, indicating weak sensitivity to phylogenetic relationships.
- **High similarity can mask structural ignorance.** Models that show strong overall alignment (e.g., Gemini Embed) may still fail to encode relative linguistic proximity, suggesting that similarity alone is insufficient to assess multilingual structure.

These findings support the hypothesis that many multilingual embedding spaces do not encode linguistic family structure and instead treat non-English languages as a homogeneous group.

I Centroid Drift and Isometry

I.1 Definition of Centroid Drift

An ideal multilingual embedding space would place semantically equivalent sentences from different languages in overlapping regions of the vector space. In practice, language-specific embedding clouds are often displaced, requiring explicit alignment.

We quantify this effect using **centroid drift**, defined as the Euclidean distance between the centroid of English embeddings and the centroid of a target language:

$$\text{Drift}_{\text{abs}} = \|\boldsymbol{\mu}_{\text{En}} - \boldsymbol{\mu}_{\text{X}}\|$$

Because raw distances are not directly comparable across models with different scales, we also report **normalized drift**, computed by dividing the absolute drift by the mean distance of English embeddings to their centroid. Lower values indicate stronger zero-shot alignment.

I.2 Centroid Drift Across Models

Table 12 reports absolute and normalized centroid drift for each model and target language.

I.3 Interpretation Notes

Several conclusions follow from Table 12:

- **Gemini Embed exhibits the strongest isometry.** It achieves the lowest normalized drift across all target languages, indicating near-overlapping multilingual embedding clouds.

Model	Lang.	Abs.	Norm.	Dim
Gemini Embed.	Bn	0.264	**0.410**	3072
	Hi	0.271	0.421	3072
	Ar	0.258	**0.402**	3072
Mistral Embed.	Bn	0.426	**0.769**	1024
	Hi	0.371	0.669	1024
	Ar	0.374	0.675	1024
OpenAI-L Embed.	Bn	0.511	**0.578**	3072
	Hi	0.448	0.507	3072
	Ar	0.473	0.535	3072
OpenAI-S Embed.	Bn	0.518	**0.590**	1536
	Hi	0.454	0.517	1536
	Ar	0.436	0.496	1536
Qwen Embed.	Bn	0.387	**0.610**	4096
	Hi	0.303	0.477	4096
	Ar	0.294	0.463	4096

Table 12: Absolute and normalized centroid drift between English and target languages. Lower normalized drift indicates better isometry.

- **Low dimensionality amplifies misalignment.** Mistral Embed shows severe centroid displacement, consistent with its high anisotropy and hubness (Appendices B and D).
- **Higher dimensionality does not guarantee isometry.** Qwen Embed (4096D) exhibits larger normalized drift than Gemini Embed, demonstrating that scale alone is insufficient for zero-shot alignment.

These centroid-level results explain why some models require explicit alignment layers or post-hoc transformations despite strong average similarity scores.

J Vector Norm Analysis (Magnitude Bias)

J.1 Motivation

In embedding-based retrieval systems, vector magnitude can influence similarity computations, particularly when dot-product or unnormalized cosine similarity is used. If embeddings from one language consistently have larger norms, they may dominate retrieval results regardless of semantic relevance, creating a *magnitude bias*.

To test for this effect, we analyze the distribution of L2 norms for embeddings across languages and models.

J.2 Norm Statistics

For each model and language, we compute the mean and standard deviation of the L2 norm over all sentence embeddings. All reported values are computed *after* the model's native embedding generation step, prior to any external normalization.

Model	Lang.	Mean Norm	Std Dev
Gemini Embed.	En	1.000000	5.7×10^{-8}
	Bn	1.000000	5.4×10^{-8}
	Hi	1.000000	5.5×10^{-8}
	Ar	1.000000	5.4×10^{-8}
Mistral Embed.	En	1.000000	2.0×10^{-5}
	Bn	1.000000	2.1×10^{-5}
	Hi	1.000000	2.1×10^{-5}
	Ar	1.000000	2.1×10^{-5}
OpenAI-L Embed.	En	1.000000	3.1×10^{-8}
	Bn	1.000000	3.6×10^{-8}
	Hi	1.000000	3.6×10^{-8}
	Ar	1.000000	3.5×10^{-8}
OpenAI-S Embed.	En	1.000000	3.5×10^{-8}
	Bn	1.000000	3.0×10^{-8}
	Hi	1.000000	3.1×10^{-8}
	Ar	1.000000	3.2×10^{-8}
Qwen Embed.	En	1.000003	2.7×10^{-4}
	Bn	0.999999	3.0×10^{-4}
	Hi	1.000001	2.9×10^{-4}
	Ar	1.000006	3.2×10^{-4}

Table 13: Mean and standard deviation of embedding vector norms by model and language.

J.3 Interpretation Notes

Table 13 leads to several clear conclusions:

- **No meaningful magnitude bias is present.** All models produce embeddings with mean norms extremely close to 1.0 across all languages.
- **Norm variance is negligible relative to other effects.** Even the highest observed variance (Qwen Embed) is orders of magnitude smaller than the geometric distortions observed in anisotropy, hubness, and centroid drift.
- **Downstream failures are not norm-driven.** Differences in alignment and retrieval performance must therefore arise from directional geometry rather than vector magnitude.

This analysis rules out magnitude bias as a confounding factor and strengthens the interpretation of the geometric diagnostics presented in earlier appendices.

Model	Accuracy	Interpretation
Gemini Embed.	0.999	Separable Islands
Mistral Embed.	1.000	Separable Islands
OpenAI-L Embed.	1.000	Separable Islands
OpenAI-S Embed.	1.000	Separable Islands
Qwen Embed.	0.997	Separable Islands

Table 14: Language classification accuracy using a linear probe. Chance level is 0.25 for four languages.

K Language Separability and the Language Island Effect

K.1 Linear Separability Test

A core requirement of a language-agnostic semantic space is that vectors should be organized by *meaning* rather than by *language identity*. If language identity can be recovered by a simple linear classifier, the embedding space has not formed a true interlingua.

To test this, we train a multinomial logistic regression classifier to predict the language label (English, Bangla, Hindi, Arabic) directly from embedding vectors. The classifier is trained and evaluated using a standard train/test split with no hyperparameter tuning.

K.2 Classification Results

Table 14 reports classification accuracy for each embedding model.

K.3 Interpretation Notes

The results in Table 14 indicate that:

- **All models are nearly perfectly language-separable.** Linear classification accuracy approaches 100% across all architectures, far exceeding chance performance.
- **Language identity is linearly encoded.** The existence of a linear decision boundary implies that languages form distinct, separable regions rather than a shared semantic manifold.
- **High alignment does not imply language invariance.** Even models with strong cross-lingual alignment and low centroid drift (e.g., Gemini Embed) fail to produce a truly language-agnostic space.

These findings support the "language island" hypothesis and explain why post-hoc alignment and language-specific normalization are often required in multilingual retrieval systems.

Model	Ratio	Direct	Pivoted
Gemini Embed.	0.472	0.738	1.573
Mistral Embed.	0.462	0.682	1.481
OpenAI-L Embed.	0.504	1.140	2.276
OpenAI-S Embed.	0.510	1.256	2.468
Qwen Embed.	0.509	0.869	1.718

Table 15: Triangle consistency test. Lower ratios indicate stronger reliance on English as a pivot.

L Geometric Transitivity and Retrieval Reciprocity

L.1 Triangle Consistency Test

In a geometrically consistent multilingual embedding space, distances between languages should be approximately transitive. Specifically, the direct distance between two non-English languages (e.g., Hindi–Arabic) should be comparable to the indirect path through English (Hindi $\rightarrow$ English $\rightarrow$ Arabic).

We quantify this property using the ratio:

$$\text{Ratio} = \frac{\text{Dist}(\text{Hi}, \text{Ar})}{\text{Dist}(\text{Hi}, \text{En}) + \text{Dist}(\text{En}, \text{Ar})}$$

A ratio close to 1 indicates geometric consistency, while substantially lower values indicate that English acts as a mandatory pivot rather than a shared semantic reference frame.

L.2 Triangle Test Results

Table 15 reports the mean ratio along with average direct and pivoted distances.

L.3 Nearest-Neighbor Reciprocity

While the triangle test captures global geometry, retrieval systems depend on *local neighborhood symmetry*. Given a source sentence x in language A and its nearest neighbor y in language B, reciprocity measures whether x is also the nearest neighbor of y when querying in the reverse direction.

We report **Reciprocity Score**, defined as the percentage of translation pairs for which nearest-neighbor retrieval is mutual.

L.4 Reciprocity Results

L.5 Interpretation Notes

The combined results from Tables 15 and 16 reveal a consistent pattern:

- **English acts as a structural bottleneck.** All models exhibit triangle ratios well below 1,

Model	Reciprocity (%)
Gemini Embed	**31.4**
Mistral Embed.	13.2
OpenAI-L Embed.	25.7
OpenAI-S Embed.	11.1
Qwen Embed.	19.6

Table 16: Nearest-neighbor reciprocity scores. Higher values indicate more stable bidirectional retrieval.

indicating that non-English languages are not directly aligned and instead rely on English as an intermediary.

- **Hubness leads to one-way retrieval.** Models with severe hubness (Mistral Embed, OpenAI-S Embed) show the lowest reciprocity, confirming that dominant hubs absorb neighbors without reciprocal connections.
- **Geometric stability improves but does not solve reciprocity.** Gemini Embed achieves the highest reciprocity score, yet even it recovers fewer than one-third of translation pairs symmetrically.

These findings demonstrate that many multilingual embedding spaces form *directional* rather than symmetric semantic neighborhoods. Taken together with earlier appendices, this establishes one-way retrieval as a fundamental geometric failure mode in current multilingual embedding models.