# Hierarchical Geometry of Cognitive States in Transformer Embedding Spaces

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Recent work has shown that transformer-based language models learn rich geometric structure in their embedding spaces, yet the presence of higher-level cognitive organization within these representations remains underexplored. In this work, we investigate whether sentence embeddings encode a graded, hierarchical structure aligned with human-interpretable cognitive or psychological attributes. We construct a dataset of 480 natural-language sentences annotated with both continuous energy scores (ranging from $-5$ to 5) and discrete tier labels spanning seven ordered consciousness-related cognitive categories. Using fixed sentence embeddings from multiple transformer models, we evaluate the recoverability of these annotations via linear and shallow nonlinear probes. Across models, both continuous energy scores and tier labels are reliably decodable by both linear and nonlinear probes, with nonlinear probes outperforming linear counterparts. To assess statistical significance, we conduct nonparametric permutation tests that randomize labels while preserving embedding geometry, finding that observed probe performance significantly exceeds chance under both regression and classification null hypotheses ($p < 0.005$). Qualitative analyses using UMAP visualizations and tier-level confusion matrices are consistent with these findings, illustrating a coherent low-to-high gradient and predominantly local (adjacent-tier) confusions in embedding space. Taken together, these results provide evidence that transformer embedding spaces exhibit a hierarchical geometric organization statistically aligned with our human-defined cognitive structure; while this work does not claim internal awareness or phenomenology, it demonstrates a systematic alignment between learned representation geometry and interpretable cognitive and psychological attributes, with potential implications for representation analysis, safety modeling, and geometry-based generation steering.

## 1 Introduction

Modern transformer-based language models (Vaswani et al., 2017) represent text as points in high-dimensional embedding spaces, capturing rich semantic and syntactic regularities learned from large-scale human-generated corpora (Ethayarajh, 2019). These representations have been extensively studied for tasks such as semantic textual similarity and sentence-level semantics (Reimers & Gurevych, 2019), affective and sentiment analysis (Mohammad, 2018; Kim et al., 2020), and transfer to a wide range of downstream natural language understanding benchmarks (Wang et al., 2019; Muennighoff et al., 2022). However, relatively little is known about whether embedding spaces encode structured, hierarchical relationships among more abstract cognitive or psychological states beyond coarse sentiment polarity.

In psychology and cognitive science, human experience is often described as organized along graded dimensions of awareness or affect, ranging from contracted, distress-oriented states to more integrative and coherent modes of cognition (Varela et al., 1991; Mohammad, 2018). While natural language processing research has extensively explored affective dimensions such as valence and arousal, whether high-dimensional language representations reflect a deeper, structured hierarchical organization of cognitive states remains underexplored.

In this work, we investigate whether transformer sentence embeddings exhibit a non-random hierarchical structure aligned with a cognitive annotation scheme. Rather than treating emotions or mental states as flat or independent categories, we ask whether embedding spaces organize language in a manner that reflects graded levels of human cognitive or psychological attributes.

To study this question, we construct a dataset of 480 natural-language sentences annotated along a seven-tier taxonomy of cognitive states, ranging from low-awareness, high-distress expressions to high-coherence, integrative states. The taxonomy is informed by psychological theory and contemplative traditions (Varela et al., 1991; Jung, 1964; Hawkins, 1995; Laozi, 1891), but is operationalized purely as a labeling framework for empirical analysis, without claims of being a theoretically complete or exhaustively studied category system.

We analyze multiple widely used sentence embedding models, including BGE, MPNet, and MiniLM, using a combination of visualization, probing (Belinkov & Glass, 2019), and statistical validation techniques. Specifically, we employ UMAP to examine the global geometry of embedding spaces, linear and shallow nonlinear probes to quantify the decodability of the annotated hierarchy, and nonparametric permutation tests (Good, 2013) to assess whether observed patterns can be explained by chance or surface lexical cues.

Our results provide empirical evidence that transformer sentence embeddings encode a statistically significant hierarchical organization correlated with the proposed tiers and continuous energy scores. These findings suggest that beyond surface-level sentiment, embedding spaces may reflect deeper, structured patterns aligned with human-interpretable cognitive and psychological states.

**Contributions:**

The main contributions of this work are as follows: (1). We introduce a consciousness-inspired cognitive annotation scheme for natural-language sentences, organized into seven ordered tiers reflecting graded levels of cognitive and psychological awareness. (2). We provide quantitative evidence, via linear and nonlinear probing experiments, (Hewitt & Manning, 2019) that multiple transformer embedding models encode information aligned with this hierarchical structure. (3). We validate the statistical significance of these findings using nonparametric permutation tests, demonstrating that the observed structure is unlikely to arise from random label alignment. (4). We offer qualitative visual analysis showing consistent hierarchical organization across models in low-dimensional projections of embedding space. (5). We show that embedding geometry can potentially serve as a measurable and interpretable substrate for studying structured human cognitive and psychological states in language, with implications for representation analysis, interpretability, and alignment-related applications.

## 2 Dataset

To study whether transformer embedding spaces encode a structured hierarchy of cognitive states, we construct a manually annotated dataset of natural-language sentences spanning a wide range of affective and cognitive modes. These labels represent structural patterns in language rather than a claim of subjective awareness or sentience in the model.

### 2.1 Cognitive Tier Taxonomy

Inspired by work in cognitive science, psychology, consciousness studies, and contemplative traditions, (Varela et al., 1991; Jung, 1964; Hawkins, 1995; Laozi, 1891) we define a seven-tier taxonomy representing qualitatively distinct modes of consciousness-related cognitive experience. The tiers are ordered from highly contracted, self-destructive states to expansive, integrative states:

- **Shadow (Unconscious / Collapse):** Ignorance, self-blame, despair, apathy, psychological collapse.
- **Striving (Scarcity and Attachment):** Fear, craving, insecurity, anxiety, survival-oriented thinking.

- **Conflict (Ego and Opposition):** Anger, hostility, dominance, control, power struggles.
- **Activation (Energy Mobilization):** Courage, resolve, neutrality, acceptance, behavioral readiness.
- **Growth (Inner Reorganization):** Openness, forgiveness, transformation, contribution, learning.
- **Clarity (Cognitive Integration):** Reasoning, abstraction, understanding, coherence, meaning.
- **Unity (Non-Dual Integration):** Compassion, joy, peace, surrender, wholeness.

Each tier represents a mode of organization of experience rather than merely emotional valence.

### 2.2 Sentence Construction and Annotation

The dataset consists of 480 short natural-language sentences distributed across the seven tiers. Sentences were constructed to reflect:

- first-person experiential language,
- psychologically plausible phrasing,
- minimal reliance on explicit emotion words where possible.

The annotations were performed manually by the author to ensure internal consistency across tiers. The objective was semantic representativeness rather than exhaustive coverage. Each sentence is assigned to exactly one tier, and no sentence appears in multiple tiers.

Table 1 presents representative example sentences from each consciousness-related cognitive tier. These examples are provided for illustrative purposes only and are not used for model training or evaluation. More sentence examples are provided in Appendix A

Table 1: Representative example sentences from each cognitive tier with illustrative energy scores.

| TIER | EXAMPLE SENTENCE | Score |
|---|---|---|
| Shadow | "I feel like everything I do just makes things worse, and I don't see a way out." | $-4.5$ |
| Striving | "I keep worrying that I'm not doing enough, and they'd leave me." | $-2.9$ |
| Conflict | "Why would I listen to people not at my level? Nobody knows better than me." | $-1.7$ |
| Activation | "I can accept what is happening and pull myself back to center." | 0.0 |
| Growth | "I'm learning from what happened and trying to respond differently this time." | 1.8 |
| Clarity | "Looking at the situation objectively helps me understand why it unfolded this way." | 3.0 |
| Unity | "I feel a quiet sense of connection and compassion, even in difficulty." | 4.2 |

### 2.3 Continuous Energy Scores

In addition to discrete tier labels, each sentence is annotated with a continuous energy score ranging from $-5$ to $+5$. These scores provide a coarse ordinal signal reflecting the relative contraction or expansion of the consciousness-related cognitive state expressed by the sentence. The energy score is intended as a human-interpretable scalar capturing how language varies in perceived activation, vitality, or depletion. Energy scores were assigned manually by the author to reflect the relative position of each sentence along the proposed low-to-high cognitive spectrum. Scores were chosen to be internally consistent within and across tiers, and are allowed to overlap across adjacent tiers, particularly near tier boundaries:

- Lower values (approximately $-5$) correspond to highly contracted, self-destructive states.
- Higher values (approximately $+5$) correspond to expansive, integrative states.
- Intermediate values correspond to transitional or neutral states, with scores near 0 reflecting activation or readiness, marking a shift from contracted toward more expansive modes along this scale.

Table 1 also presents example energy scores for illustrative purposes.

### 2.4 Dataset Intent

The continuous energy scores are not treated as precise measurements or interval-scaled ground truth. Instead, they provide a coarse ordinal signal for assessing whether sentences expressing similar consciousness-related cognitive states tend to occupy nearby regions in embedding space, and are used for visualization, probing, and permutation-based validation.

The purpose of the dataset is strictly empirical: to test whether modern sentence embedding models encode non-trivial hierarchical structure aligned with graded, human-defined consciousness-related cognitive levels, rather than to provide clinical, diagnostic, or normative measurement.

## 3 Methods

### 3.1 Sentence Embeddings

We study whether pretrained language models encode a structured hierarchy of consciousness-related cognitive states in their embedding spaces. Given a dataset of natural-language sentences annotated with both discrete tier labels (Section 2.1) and continuous energy scores (Section 2.3), we first map each sentence into a fixed-dimensional embedding space using frozen pretrained encoders.

We evaluate three widely used sentence embedding models commonly used in sentence representation learning (Reimers & Gurevych, 2019):

- `BAAI/bge-large-en-v1.5` Xiao et al. (2023)
- `sentence-transformers/all-mpnet-base-v2` Song et al. (2020)
- `sentence-transformers/all-MiniLM-L6-v2` (Wang et al., 2020)

All models are used in inference-only mode without any fine-tuning. Each sentence is encoded into a single vector representation using the model's default pooling strategy. Embeddings are L2-normalized prior to downstream analysis to ensure comparable cosine-based geometry (Salton et al., 1975) across models.

Let $X \in \mathbb{R}^{N \times d}$ denote the resulting embedding matrix for $N$ sentences with embedding dimension $d$; all subsequent analyses treat $X$ as fixed.

### 3.2 Probing Analysis

To evaluate whether consciousness-related cognitive structure is encoded in transformer embedding spaces, we employ a set of probing models that predict annotated attributes from fixed sentence embeddings (Belinkov & Glass, 2019; Hewitt & Manning, 2019) obtained from the models described in Section 3.1. Probing is used as a diagnostic tool to assess which information is recoverable from the representations, rather than as an end-task optimization objective.

#### 3.2.1 Regression Probes for Continuous Energy Scores

Each sentence is annotated with a continuous energy score in the range $[-5, 5]$, reflecting its position along a low-to-high cognitive spectrum (Section 2.3). We examine whether this scalar signal is encoded in embedding geometry using two regression probes.

**Linear probe (Ridge regression).** Ridge regression provides a conservative test of whether energy scores are linearly decodable from embeddings, serving as a lower bound on representational structure while controlling for overfitting through $\ell_2$ regularization.

**Nonlinear probe (MLP regressor).** To assess the presence of additional nonlinear structure, we train a shallow multilayer perceptron with two hidden layers. This probe captures modest nonlinear interactions while remaining limited in capacity, avoiding the expressivity of deep task-optimized models.

Models are trained using an 80/20 train–test split and evaluated using the coefficient of determination ($R^2$) and mean squared error (MSE).

#### 3.2.2 Classification Probes for Cognitive Tiers

In addition to continuous scores, each sentence is labeled into one of seven ordered cognitive tiers: Shadow, Striving, Conflict, Activation, Growth, Clarity, and Unity (Section 2.1). We train a multiclass logistic regression classifier to predict tier labels from sentence embeddings.

Logistic regression is deliberately chosen as a low-capacity linear classifier to ensure that classification performance reflects intrinsic separability of tier structure in the embedding space rather than probe expressivity. Performance is reported using accuracy and weighted F1-score to account for class imbalance.

#### 3.2.3 Confusion Matrix Analysis

To further analyze classification behavior, we inspect confusion matrices of the tier classifier. Confusion matrices are visualized for a representative train–test split (random seed = 0).

We examine whether misclassifications predominantly occur between adjacent tiers (e.g., Growth ↔ Clarity, Activation ↔ Growth), rather than between distant tiers (e.g., Striving ↔ Clarity). Such locality-sensitive errors would indicate that embedding geometry preserves ordinal structure aligned with the graded nature of the tier annotations.

#### 3.2.4 Stability Across Splits

For each embedding model, probing results are averaged over 30 random train–test splits with an 80/20 split ratio. Each split uses a distinct random seed, controlling both data partitioning and model initialization (for nonlinear probes). Reported metrics correspond to mean performance across splits, reducing variance due to sampling effects and providing a more robust estimate of probe behavior.

### 3.3 UMAP Visualization

To provide a qualitative view of the geometric organization of consciousness-related cognitive annotations in embedding space, we apply Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to each embedding set. For each model, we compute both 2D and 3D UMAP embeddings using cosine distance (Salton et al., 1975), with hyperparameters $n_{\text{neighbors}} = 20$, min_dist = 0.1, and a fixed random seed (random_state = 42) for reproducibility.

Points are colored by the continuous energy score in the range $[-5, 5]$ using a continuous colormap. These visualizations are treated as descriptive diagnostics rather than statistical evidence; quantitative evaluation is provided by probing and permutation tests (Sections 3.2,3.4). We report 3D UMAP visualizations in the main paper and include 2D UMAP plots as supplementary material in Appendix B.

### 3.4 Permutation Tests

High probe performance alone does not guarantee that embeddings encode target attributes in a meaningful way; strong results may arise from incidental correlations or dataset artifacts. To assess whether the observed probing performance reflects a genuine alignment between embedding geometry and annotated consciousness-related attributes, we conduct nonparametric permutation tests under a label-randomization null (Good,

2013). We use a fixed random number generator (RNG) seed in permutation tests to ensure reproducibility and fair comparison between the observed statistic and the null distribution.

#### 3.4.1 Null Hypotheses and Test Statistics

We consider two complementary null hypotheses:

- **Energy score null (energy score regression).** Continuous energy scores are independent of sentence embeddings.
- **Tier null (tier classification).** Discrete tier labels are independent of sentence embeddings.

Under each null, labels are randomly permuted while embeddings $X$ are held fixed. For each permuted dataset, we re-run the probing protocol from Section 3.2 using 30 repeated 80/20 train–test splits with fixed split seeds.

We use the following test statistics:

- Regression: mean Ridge $R^2$ across splits, $\overline{R^2}$.
- Classification: mean weighted F1-score across splits, $\overline{\mathrm{F1}}_{\mathrm{w}}$.

Permutation tests are conducted on the strongest-performing embedding model (BAAI/bge-large-en-v1.5) to provide a focused and reproducible significance assessment. Linear probes (Ridge regression for energy scores and logistic regression for tier classification) are used to obtain a conservative estimate of statistical significance.

#### 3.4.2 Monte Carlo Permutation Procedure

Let $T_{\mathrm{obs}}$ denote the observed test statistic. We approximate the null distribution via Monte Carlo permutation by repeating the following procedure $N = 200$ times:

1. Randomly permute target labels $y' \leftarrow \pi(y)$ using a fixed random number generator seed.
2. Apply the same probing protocol as in Section 3.2 using the same split seeds.
3. Compute the mean test statistic $T_i$ across splits.

A one-sided permutation $p$-value is computed using the smoothed estimator (Good, 2013):

$$p = \frac{1 + \sum_{i=1}^{N} \mathbb{I}[T_i \geq T_{\mathrm{obs}}]}{N + 1}.$$

#### 3.4.3 Reporting and Visualization

For each permutation test, we report the observed probe performance together with the empirical null distribution induced by label shuffling. We report mean Ridge $R^2$ for energy regression and mean weighted F1-score for tier classification, each averaged across 30 train–test splits. Null distributions are visualized using histograms with the observed statistic indicated by a vertical reference line.

## 4 Results

### 4.1 Decodability of Continuous Energy Scores

We first evaluate whether the continuous energy scores assigned to sentences are recoverable from fixed transformer embeddings. This analysis tests whether embedding geometry preserves graded structure aligned with a low-to-high consciousness-related cognitive spectrum.

Across all evaluated embedding models, energy scores are strongly decodable. As shown in Table 2, regression performance is well above chance, capturing a substantial proportion of the annotated ordinal signal on held-out data. For the BAAI/bge-large-en-v1.5 embeddings, the mean coefficient of determination exceeds 0.80, indicating that a large fraction of the annotated energy signal is recoverable from the representation space. Comparable but slightly lower performance is observed for `all-mpnet-base-v2` and `all-MiniLM-L6-v2`.

Table 2: Energy regression probe performance averaged over 30 train–test splits.

| **Model** | **Ridge $R^2$** | **Ridge MSE** | **MLP $R^2$** | **MLP MSE** |
|---|---|---|---|---|
| BAAI/bge-large-en-v1.5 | 0.808 | 1.824 | 0.830 | 1.605 |
| all-mpnet-base-v2 | 0.750 | 2.373 | 0.769 | 2.182 |
| all-MiniLM-L6-v2 | 0.671 | 3.118 | 0.698 | 2.859 |

Decodability improves consistently when moving from linear to nonlinear regression. Across all models, nonlinear probes achieve higher $R^2$ and lower mean squared error than their linear counterparts. This pattern suggests that while energy score-related structure is partially aligned with linear directions in embedding space, additional information is organized in a nonlinear manner.

We also observe a clear ordering across embedding models. Larger and more expressive models yield stronger decodability, with BGE outperforming MPNet, and MPNet outperforming MiniLM. This monotonic trend holds for both linear and nonlinear probes, indicating that representational capacity influences how clearly graded energy information is preserved.

Overall, these results demonstrate that continuous energy scores are robustly encoded in transformer embedding spaces. The consistent advantage of nonlinear decoding further indicates that this structure is not purely linear, but reflects richer geometric organization within the embeddings.

### 4.2 Decodability of Cognitive Tiers

We next examine whether the discrete cognitive tiers assigned to sentences are recoverable from transformer embedding representations. Unlike the continuous energy scores in Section 4.1, tier labels represent coarser categorical stages along the same underlying spectrum.

Across all embedding models, tier labels are substantially decodable, with classification performance well above chance. As summarized in Table 3, weighted F1-scores range from approximately 0.70 to 0.77 across models, indicating that the embedding space preserves meaningful separation among the seven tiers.

Table 3: Tier classification probe performance averaged over 30 train–test splits.

| Model | Accuracy ↑ | Weighted F1 ↑ |
|---|---|---|
| BAAI/bge-large-en-v1.5 | 0.779 | 0.766 |
| all-mpnet-base-v2 | 0.774 | 0.764 |
| all-MiniLM-L6-v2 | 0.717 | 0.703 |

Consistent with the regression results, performance varies systematically across models. BAAI/bge-large-en-v1.5 achieves the strongest tier decodability, followed by `all-mpnet-base-v2` and `all-MiniLM-L6-v2`. This ordering mirrors the trend observed for continuous energy prediction, suggesting that both continuous and categorical annotations align with shared representational structure.

To better understand the nature of classification errors, we inspect confusion matrices for a representative train–test split. Figure 1 shows the confusion matrix for tier classification using BAAI/bge-large-en-v1.5 embeddings. Misclassifications are concentrated between adjacent tiers (e.g., Activation ↔ Growth, Growth ↔ Clarity), while confusions between distant tiers (e.g., Striving ↔ Clarity) are rare. Similar patterns are observed for other embedding models (Appendix A).

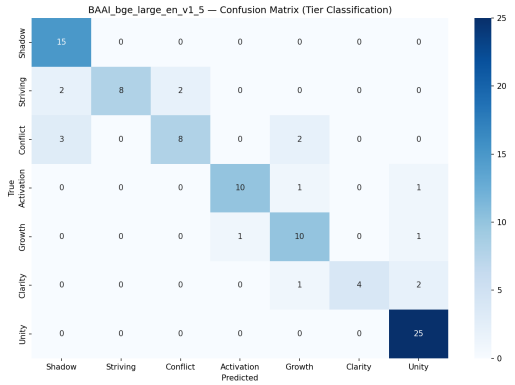



Figure 1: Confusion matrix for tier classification using BAAI-bge-large-en-v1.5 embeddings.

Taken together, these results indicate that cognitive tiers are not only decodable, but are organized in an ordered fashion within the embedding space. The concentration of errors between neighboring tiers suggests that embeddings encode a graded structure consistent with the proposed hierarchy, rather than treating tiers as arbitrary categorical labels.

### 4.3 Qualitative Structure in Embedding Space via UMAP

To complement quantitative probing results, we examine the geometric organization of sentences in embedding space using UMAP visualizations (McInnes et al., 2018) colored by continuous energy scores (Figure 2).

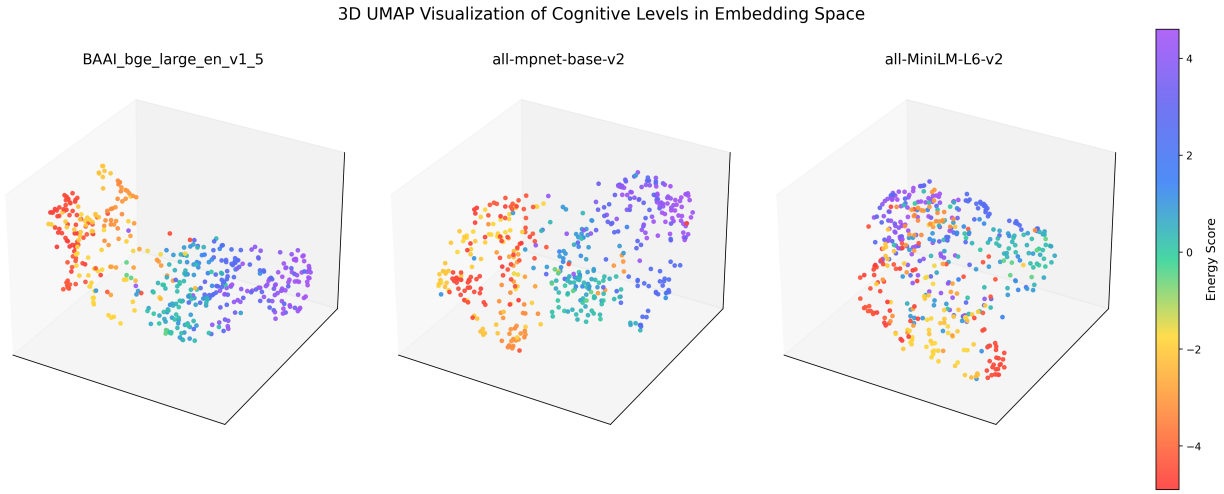



Figure 2: 3D UMAP visualization of sentence embeddings colored by energy scores.

Across all three embedding models, UMAP reveals a clear low-to-high energy gradient rather than random mixing. Sentences annotated with lower energy scores (e.g., Shadow and Striving) tend to occupy contiguous regions, while higher-energy sentences (e.g., Clarity and Unity) occupy distinct regions of the embedding space.

Model-dependent differences are apparent. BAAI/bge-large-en-v1.5 exhibits the most coherent structure, with a smooth, approximately monotonic transition from low-energy to high-energy regions. `all-mpnet-base-v2` shows a similar global gradient but with increased overlap between adjacent energy levels. `all-MiniLM-L6-v2` displays greater dispersion and mixing, consistent with its weaker regression and classification performance.

Importantly, misalignments observed in the confusion matrices—primarily between adjacent tiers—are reflected in UMAP by local overlaps rather than long-range mixing. Distant tiers (e.g., Shadow vs. Unity) rarely occupy the same regions, suggesting that embedding geometry preserves a coarse hierarchical ordering even when fine-grained boundaries are ambiguous.

These visualizations are intended as qualitative diagnostics and are interpreted in conjunction with the quantitative probing and permutation-test results.

### 4.4 Statistical Significance via Permutation Tests

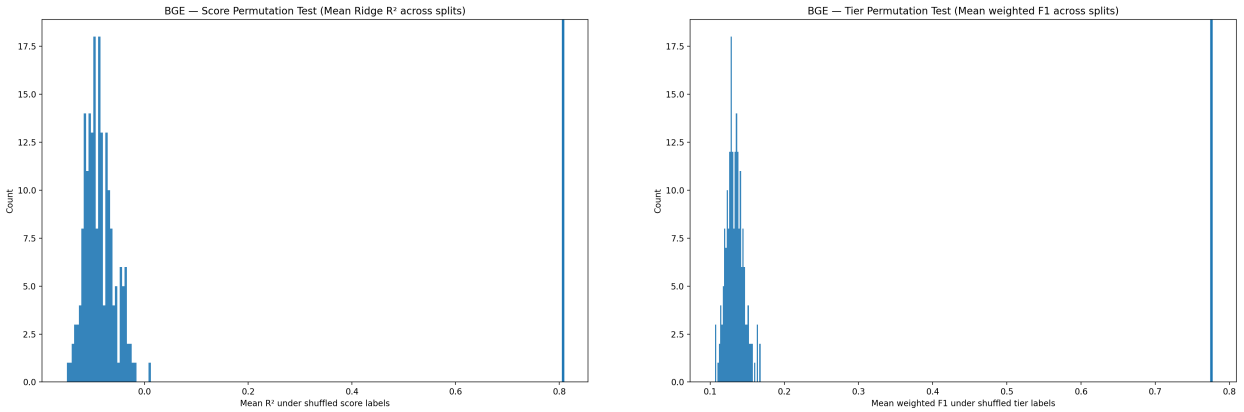
To assess whether the observed probe performance reflects a genuine alignment between embedding geometry and the annotated consciousness-related cognitive attributes—rather than spurious correlations—we evaluate statistical significance using nonparametric permutation tests (Section 3.4).

Table 4: Statistical significance of probing results for BAAI/bge-large-en-v1.5 embeddings.

| Task | Metric | Observed Mean | $p$-value |
|---|---|---|---|
| Energy regression | Mean Ridge $R^2$ | 0.808 | $<0.005$ |
| Tier classification | Mean weighted F1 | 0.776 | $<0.005$ |

For the BAAI/bge-large-en-v1.5 embeddings, observed probe performance averaged across 30 random 80/20 train–test splits is strong for both regression and classification tasks. Under the score permutation null, the empirical null distribution of mean Ridge $R^2$ values lies far below the observed statistic. With $N = 200$ permutations, the resulting one-sided permutation $p$-value is $p_{\text{score}} \approx 0.00498$, corresponding to the minimum resolvable value under this permutation budget.

Similarly, under the tier permutation null, the observed mean weighted F1-score substantially exceeds the null distribution, yielding $p_{\text{tier}} \approx 0.00498$. In both cases, fewer than 1 in 200 random label assignments achieve comparable probe performance.




(a) Score permutation test (Ridge $R^2$)

(b) Tier permutation test (Weighted F1)

Figure 3: Permutation test results. Vertical lines indicate observed probe performance.

As shown in Figure 3, the empirical null distributions are tightly concentrated near chance performance, while the observed probe statistics lie far in the upper tail, indicating strong separation from the label-randomization baseline.

Together, these results provide strong evidence that the embedding space encodes information systematically aligned with both continuous energy scores and discrete cognitive tiers. It demonstrates that learned representation geometry preserves structured information correlated with the graded annotations beyond what would be expected under label independence.

### 4.5 TF-IDF Baseline

As a lexical baseline, TF-IDF representations achieve substantially lower performance across both tasks. For energy regression, TF-IDF captures substantially less of the annotated signal than transformer-based embeddings, achieving a mean $R^2$ of approximately 0.40, compared to 0.67–0.81 for contextualized sentence embeddings. Similarly, for tier classification, TF-IDF yields a mean weighted F1-score of approximately 0.43, well below the performance of transformer models, which consistently exceed 0.70.

These performance gaps indicate that the hierarchical structure identified in embedding space is not recoverable from surface word statistics alone. While TF-IDF captures correlations between individual lexical items and affective content, it fails to encode the graded, globally organized structure observed in transformer representations.

Taken together, these results suggest that the observed cognitive hierarchical organization cannot be explained by artifacts of word frequency, n-gram co-occurrence, or simple lexical cues alone.

## 5 Discussion

This work investigates whether transformer sentence embeddings encode a graded hierarchical structure aligned with human-defined, cognitive attributes. Across multiple embedding models, probing analyses demonstrate that both continuous energy scores and discrete cognitive tiers are reliably decodable from fixed embeddings. Permutation tests confirm that this decodability is highly unlikely under a label-independence null, supporting the interpretation that the observed structure reflects nontrivial alignment between embedding geometry and the annotated attributes rather than incidental correlations.

### 5.1 What Is Encoded in the Embedding Geometry?

The results indicate that transformer embedding spaces preserve a graded, hierarchically organized structure aligned with consciousness-related cognitive or psychological states. Linear probes recover substantial information about both continuous energy scores and discrete tiers, suggesting that a significant portion of this structure is accessible along global geometric directions in the embedding space. The consistent performance gains obtained with shallow nonlinear probes further indicate that this organization is not purely linear, but instead distributed across dimensions in a structured, weakly nonlinear manner.

This structure cannot be explained by surface lexical statistics alone. A TF-IDF lexical baseline exhibits substantially weaker performance across both regression and classification tasks, indicating that the observed organization reflects higher-order contextual representations rather than word frequency or n-gram co-occurrence patterns. Permutation tests further confirm that probe performance is highly unlikely under label-randomization nulls, supporting the interpretation that the recovered structure reflects meaningful alignment between embedding geometry and the annotated attributes.

These findings do not imply that language models explicitly represent consciousness as a semantic variable.

### 5.2 Hierarchical Organization and Local Ambiguity

Both quantitative and qualitative analyses point to a hierarchical organization of the annotated tiers within embedding space. Confusion-matrix analyses show that misclassifications predominantly occur between adjacent tiers (e.g., Activation vs. Growth), while confusions between distant tiers (e.g., Shadow vs. Unity) are rare. This behavior is mirrored in UMAP visualizations, where local overlap appears primarily among neighboring energy levels, whereas globally distant tiers occupy well-separated regions.

This pattern is consistent with a representation that preserves a coarse global ordering while allowing for local ambiguity. Such ambiguity is expected given the subjective and context-dependent nature of human cognitive, psychological or consciousness states, as well as the fact that discrete tier labels necessarily discretize an underlying continuum. The results therefore suggest that embedding geometry reflects a graded structure with fuzzy boundaries between human-defined adjacent classes.

### 5.3 Model Differences and Representational Capacity

Comparisons across embedding models reveal systematic differences in how strongly this structure is expressed. Larger and more expressive models (e.g., BAAI/bge-large-en-v1.5) exhibit higher probe performance and clearer geometric organization in low-dimensional projections. Smaller models (e.g., MiniLM) show greater dispersion and overlap, consistent with their reduced representational capacity.

These model-dependent differences suggest that the emergence of hierarchical structure in embedding space is shaped by representational capacity and training characteristics, rather than arising solely from the annotation scheme or dataset construction. At the same time, the presence of above-chance decodability across all evaluated models indicates that this organization reflects a general property of transformer-based sentence representations, with model scale influencing the clarity and robustness of the encoded structure rather than its existence.

## 6 Conclusion and Future Work

This work demonstrates that transformer embedding spaces contain a robust and statistically significant geometric structure aligned with an ordered, continuous spectrum of human-defined cognitive attributes. Across multiple embedding models, both continuous energy scores and discrete cognitive tiers are reliably decodable from fixed sentence embeddings. Linear probes recover a substantial portion of this signal, while shallow nonlinear probes yield consistent but modest performance gains, indicating that the dominant organization of the hierarchy is largely linear, though not purely so. Using identical annotation targets and probe architectures, transformer sentence embeddings substantially outperform TF-IDF lexical representations, indicating that the observed hierarchical structure is not recoverable from surface lexical cues alone. Permutation tests further confirm that these effects are highly unlikely to arise from random label alignment.

At a geometric level, this alignment manifests as a graded organization of sentences in embedding space, where lower- and higher-level cognitive expressions occupy systematically different regions connected by smooth transitions. Qualitative UMAP visualizations provide an intuitive complement to the quantitative results, revealing coherent low-to-high energy gradients and model-dependent differences in organization. Higher-performing models display smoother transitions across levels, while weaker models exhibit increased dispersion and overlap. Consistent with this structure, confusion matrix analyses show that misclassifications primarily occur between adjacent tiers rather than distant ones, indicating preservation of a coarse hierarchical ordering even when fine-grained boundaries remain ambiguous.

Beyond representation analysis, these findings suggest several directions for future work. The localization of lower-energy or high-risk psychological language within specific regions of embedding space may support applications in safety analysis, interpretability, and the differentiation of coercive versus consent-aligned language. More broadly, the observed geometry points toward the possibility of non-manipulative generation steering, in which model outputs are guided toward regions of embedding space associated with higher coherence or alignment-related attributes without reliance on explicit rule-based filtering. Future research may further evaluate robustness across datasets, languages, and annotation paradigms, explore alternative cognitive taxonomies, and investigate whether similar hierarchical organization emerges within intermediate transformer layers or during generation dynamics.

## 7 Limitations

This study has several limitations. First, the cognitive tier labels and continuous energy scores are manually annotated by a single annotator and applied to a dataset of limited size and scope consisting of short English sentences. While this design ensures internal consistency, it introduces subjectivity and limits generalization. Future work may assess robustness using multi-annotator agreement, alternative annotation schemes, larger or more naturalistic corpora, and additional languages.

Second, probing analyses measure the recoverability of information from fixed embeddings, but do not imply explicit or causal representation by the model.

### Data Availability

Due to the sensitive cognitive–affective nature of the sentence content in this dataset, we do not release the raw annotated sentences publicly. The dataset includes first-person expressions of distress, self-blame, conflict, and related internal states, which may raise ethical concerns if redistributed without appropriate context.

To support reproducibility while mitigating potential harm, we release:

- all sentence embeddings used in the experiments,
- the full annotation schema and tier definitions,
- all code for probing, visualization, and permutation testing.

These resources are sufficient to reproduce all reported results without exposing sensitive textual content.

### Broader Impact Statement

This work analyzes the geometric structure of transformer sentence embeddings and does not introduce new generative models, training objectives, or deployment mechanisms. As such, it does not directly enable harmful applications. However, prior research has highlighted that language models may pose ethical and social risks when they amplify harmful, manipulative, or coercive language patterns (Weidinger et al., 2021).

Our findings indicate that embedding spaces can exhibit localized regions associated with language expressing high-risk psychological or affective states. While such structure may support beneficial applications—such as safety analysis, interpretability, and detection of harmful or coercive intent—it could also be misused if embedding-aware generation or steering techniques were deliberately biased toward destabilizing regions.

These considerations underscore the importance of alignment-aware design, transparency, and ethical safeguards when developing or applying representation-based analysis or control methods. We view this work as contributing to a deeper understanding of embedding geometry, which is a necessary step toward safer, more interpretable, and more responsible use of large language models.

## References

Yonatan Belinkov and James Glass. Probing classifiers: Promises, shortcomings, and future directions. *Computational Linguistics*, 2019.

Kawin Ethayarajh. How contextual are contextualized word representations? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Phillip Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, 2013.

David R. Hawkins. *Power vs. Force: The Hidden Determinants of Human Behavior*. Hay House, Carlsbad, CA, 1995.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of NAACL*, 2019.

Carl G. Jung. *Man and His Symbols*. Doubleday, New York, 1964.

Taehee Kim et al. Interpretation of emotion representations in neural models. *EMNLP*, 2020.

Laozi. *Tao Te Ching*. Oxford University Press, 1891.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance. *ACL*, 2018.

Niklas Muennighoff, Nils Reimers, Andreas Rücklé, and Iryna Gurevych. Sgpt: Gpt sentence embeddings for semantic search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5881–5896. Association for Computational Linguistics, 2022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 2020.

Francisco J. Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind*. MIT Press, Cambridge, MA, 1991.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*, 2020.

Laura Weidinger et al. Ethical and social risks of harm from language models. *ACL*, 2021.

Liang Xiao et al. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

## A Additional Example Sentences per Tier

This appendix provides additional illustrative example sentences for each tier to support qualitative interpretability. These examples are intended to help readers develop an intuitive sense of the cognitive–affective states indexed by each tier, independent of access to the full annotated dataset.

**Tier 1: Highly Contracted / Self-Destructive, Apathy, Despair**

- "The feeling was like standing in a tunnel with no light at either end."
- "I can't stop blaming myself for what happened."
- "I don't care anymore - well, nobody cares about me anyways."

**Tier 2: Contracted / Striving, Fear, Manipulation**

- "I keep myself busy so I don't feel the heaviness underneath."
- "If I can make them doubt their memory, I stay safe."
- "I sense danger behind their words, attention, and actions, so I play it safe by only saying what they want to hear."

**Tier 3: Defensive / Conflict, Ego, Anger**

- "I hate seeing his name pop up; it instantly ruins my mood."
- "I feel attacked easily and respond by pushing back."
- "Ugh, I don't mix with people beneath me."

**Tier 4: Activated / Transitional, Accepting, Courage**

- "Whatever comes, I'll face it one step at a time."
- "I'm not waiting for permission anymore; I start to trust my voice."
- "I accept what has already happened and focus on what I can do next."

**Tier 5: Constructive / Growth-Oriented, Relaxed, Healing**

- "I keep my word, even when it's hard."
- "I want to give because I'm becoming someone who can give, and what I share can actually make a difference."
- "No matter how busy I am, I take breaks to relax my eyes and body."

**Tier 6: Integrated / Harmonious, Clarity, Reasoning**

- "Comprehension feels like light turning on inside my mind."
- "I can see the pattern beneath the surface."
- "Let's examine the possible causes of the issue piece by piece."

**Tier 7: Expansive / Unifying, Grace, Wisdom**

- "Love doesn't ask why; it simply shines."
- "Even difficulty feels part of a larger whole."
- "When the self dissolves, nothing is abandoned and nothing is gained."

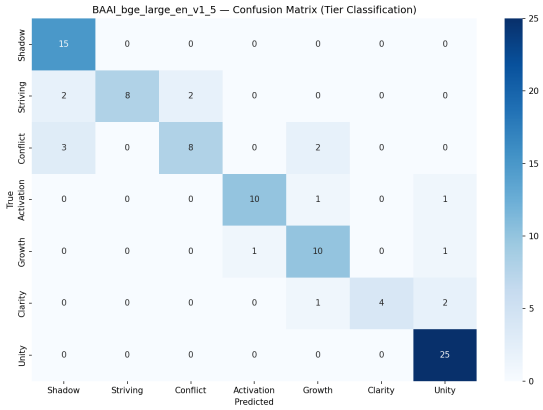
**Note** These examples are provided for qualitative reference only. All quantitative analyses reported in the main paper are based exclusively on the annotated dataset described in Section 2, and no additional sentences were introduced into the experimental pipeline.

## B Confusion Matrices for Tier Classification Across Three Embedding Models.

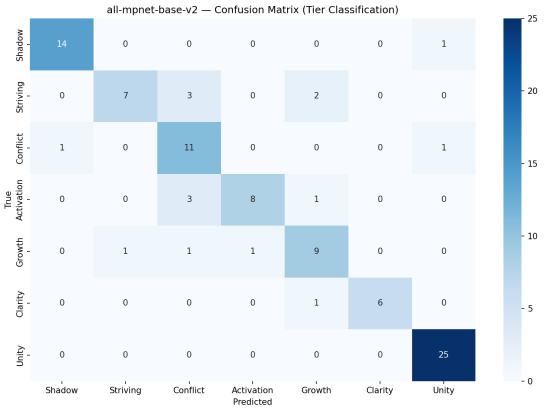
(Figure 4).This appendix presents confusion matrices for tier classification across all evaluated embedding models, illustrating that misclassifications occur predominantly between adjacent tiers rather than distant ones. The stronger models exhibit sharper diagonal structure.

## C 2D UMAP visualization of continuous energy scores

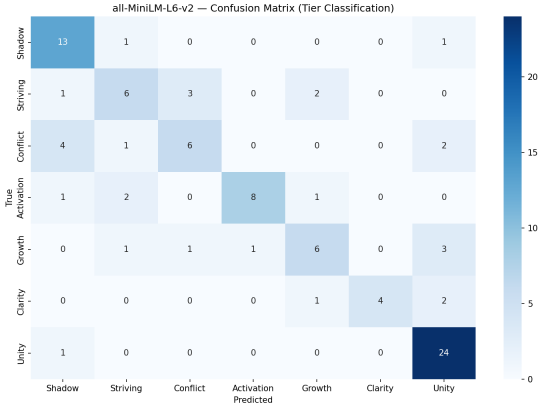
(Figure 5). This appendix provides 2D UMAP visualizations of sentence embeddings colored by continuous energy scores, offering a qualitative view of the graded geometric structure discussed in the main text.




(a) BAAI-bge-large-en-v1.5




(b) all-mpnet-base-v2




(c) all-MiniLM-L6-v2

Figure 4: Confusion matrices for tier classification across three embedding models.

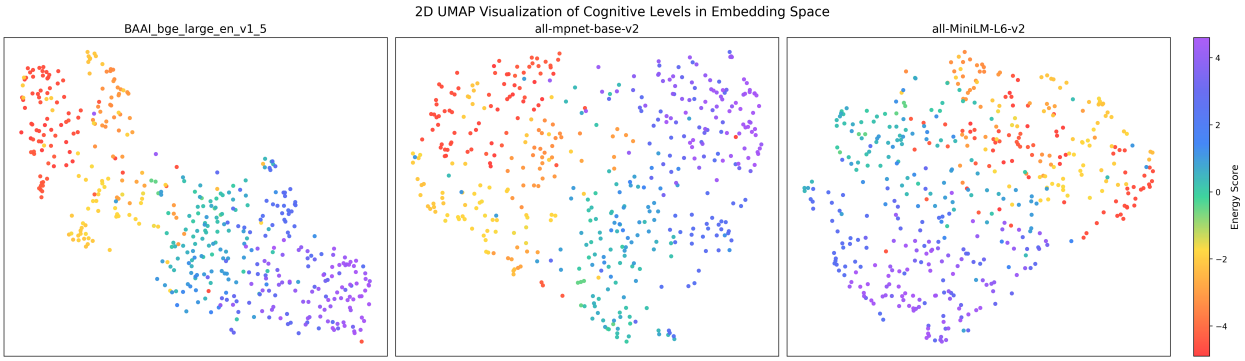



Figure 5: 2D UMAP visualization of sentence embeddings colored by continuous energy scores.