Do LLM Embedding Spaces Recover Expert Structure? Evidence from Mental-Health Language

Yixuan Zhu[*]
Zhongnan University of Economics and Law

Zhenke Duan[*,†]
CodeSoul.co
zhenkeduan03@gmail.com

Fanghen Li [*]
Zhongnan University of Economics and Law
CodeSoul.co

[*]Equal contribution. [†]Corresponding author.

Abstract

Pretrained text embeddings are increasingly used as representational maps, yet high category separability does not imply that their geometry recovers expert-defined structure. We study this problem in mental-health-related language, where symptom relations provide an external reference and online communities introduce strong domain, affective, stylistic, and discourse confounds. Using 28 Reddit communities, we compare pretrained and supervised fine-tuned Qwen3 embedding spaces at two scales (0.6B and 4B). We construct category prototypes, evaluate their representational dissimilarity matrices against an expert symptom matrix with representational similarity analysis, and complement this global test with prototype-based typicality and multi-baseline confound controls. Pretrained embeddings show measurable alignment with expert structure within the mental-health subset; fine-tuning strengthens this alignment most at the finest category level; and larger scale improves both zero-shot alignment and supervision-induced gains. Residual alignment remains substantial after controlling for VAD, LIWC, lexical style, and topic-distribution structure. In the evaluated mental-health setting, these results suggest that LLM embeddings can recover expert-relevant category geometry, but this recovery is level-dependent and should be tested against explicit confounds rather than inferred from classification alone.

Keywords: Artificial Intelligence; Computer Science; Emotion; Natural Language Processing; Representation

1 Introduction

Text embeddings from large language models are increasingly used not only as downstream features but as representational maps: distances among embedded texts can be compared with independently specified conceptual structures. This shift raises a question for structure-level evaluation: when embeddings separate categories, do they recover fine-grained conceptual organization, or do they mainly reflect coarse distributional cues? RSA-style analyses provide a way to compare model geometry with external reference structures, but matrix-level correlations can also obscure which aspects of the geometry drive the observed alignment (Kondapaneni et al., 2025; Shah et al., 2024). From a representational perspective, this distinction matters because category separability alone does not reveal whether an embedding space captures the relational organization that domain experts use to structure conceptual knowledge. Testing expert alignment therefore helps distinguish representations that merely discriminate labels from those that also encode richer relations among categories.

Mental-health language makes this problem sharp. Mental-health-related categories exhibit symptom overlap and graded boundaries, yielding an external expert reference for fine-grained category relations. Yet online communities also contain strong non-clinical signals, including subreddit jargon, stylistic norms, and a broad mental-health/control split, that may dominate the learned space. A model may therefore classify communities or cluster them cleanly without encoding the symptom-level relations that motivate the category structure. Figure 1 illustrates this gap between community separability and expert-aligned geometry.

We ask whether embedding spaces recover expert-defined symptom structure beyond coarse domain separation, and how task-specific supervision changes this geometry. We use mental-health language as a structured testbed for this question, where expert-defined category relations, graded boundaries, and strong linguistic confounds coexist. Our empirical claims are therefore restricted to this setting and do not imply that the same pattern generalizes across domains. Using 28 online com-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3331–3346
November 6-10, 2026 ©2026 Association for Computational Linguistics

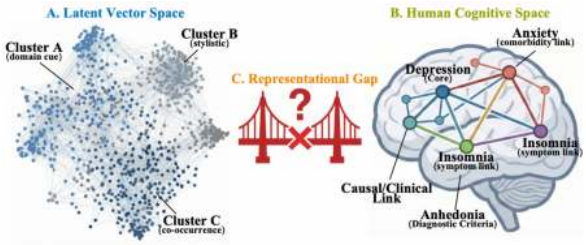


Figure 1: **The Representational Gap in Mental-Health Embeddings.** (A) Embedding spaces may cluster communities by non-clinical cues such as subreddit jargon, style, and broad domain separation. (B) Expert structure organizes categories by symptom overlap and graded boundaries. (C) We test whether model geometry aligns with this expert structure beyond domain and other non-clinical cues.

munities (17 mental-health, 11 controls), we compare pretrained and supervised fine-tuned embedding spaces. This framing shifts evaluation from community classification to the category structure encoded in the representation space, consistent with evidence that language models can recover meaningful similarity structure, while structure-level comparisons require explicit controls for confounds (Marjieh et al., 2024; Ogg et al., 2024). We focus on three questions: (1) whether embeddings align with expert-defined symptom structure beyond domain separation; (2) whether supervision reshapes global alignment, local boundary organization, or both; and (3) whether the observed geometry reflects expert-relevant relations rather than discourse, affective, or stylistic regularities.

To answer these questions, we combine three analyses: representational similarity analysis (RSA) for global alignment with expert symptom structure; prototype-based typicality for within-category cohesion and boundary confusions (Vemuri et al., 2024); and multi-baseline confound controls for affective valence–arousal–dominance (VAD), Linguistic Inquiry and Word Count (LIWC), stylistic, and topic-based regularities. Our contributions are as follows:

- We present a representation-level framework for testing whether embedding spaces recover expert-defined structure beyond domain separation, combining RSA, prototype-based typicality, and multi-baseline confound controls.
- We show a level-dependent pattern of expert-structure recovery: pretrained embeddings encode symptom-level organization within the mental-health subset, and supervision strengthens this alignment most at the finest category level.
- We compare two model scales (0.6B and 4B), showing that larger scale improves both zero-shot alignment and fine-tuning gains, while supervision and scale jointly reduce associations with non-clinical baselines and yield more category-specific geometry.

2 Related Work

2.1 Embedding geometry and structure-level evaluation

Text embeddings are increasingly analyzed not only as features for downstream tasks, but as geometric spaces whose distances, directions, and neighborhoods can be compared with human conceptual structure. Prior work shows that such geometry can recover interpretable dimensions of conceptual knowledge (Grand et al., 2022; Xu et al., 2025). In NLP, however, this structure is often operationalized through scalar similarities, such as cosine similarity between pooled sentence representations in models like SBERT (Reimers and Gurevych, 2019). These measures are useful, but they compress high-dimensional geometry into pairwise closeness and provide limited information about how categories are organized as a relational system.

Representational similarity analysis (RSA) provides a structure-level alternative. By comparing model-derived representational dissimilarity matrices with external relational references, RSA tests whether two systems organize the same items in similar ways, making it well suited for evaluating alignment between model geometry and human- or expert-defined structures (Fairhall, 2024). Yet global matrix correlations can hide local failures: two spaces may show similar overall structure while differing around category boundaries. Methods that retain finer pairwise, interaction-level, or topological information capture complementary aspects of geometry, but they do not by themselves specify how category-level representations should be evaluated against an external expert reference (Khattab and Zaharia, 2020; Brown and Farivar, 2025). Our work combines these views: we use RSA to test global expert alignment and prototype-based analyses to examine where that alignment sharpens or breaks down locally.

This framing reflects cautions against treating embedding geometry as a complete account of

meaning. Model–human alignment is partial and feature-dependent, with weaker recovery of attributes such as sensorimotor features (Xu et al., 2025). Different representation sources can induce distinct relational patterns, suggesting that embedding spaces reflect multiple organizing factors rather than a single semantic axis (Carota et al., 2024). We therefore test whether category-level geometry reflects an externally specified symptom structure rather than competing explanations such as domain membership, affective similarity, or other non-clinical regularities.

2.2 Interpretability and local structure in embedding spaces

Embedding interpretability is often approached through probing, attribution, token-level alignment, and subspace-based methods (Opitz et al., 2025). Probing tests whether linguistic or semantic information is recoverable from embeddings (Conneau et al., 2018; Muennighoff et al., 2023), but recoverability does not determine the topology or neighborhood structure of a space (Belinkov, 2022). Attribution and token-level alignment explain prediction-relevant inputs or fine-grained contextual correspondences (Sundararajan et al., 2017; Han et al., 2020; Khattab and Zaharia, 2020; Zhang et al., 2019); subspace methods make selected semantic dimensions more interpretable (Opitz and Frank, 2022). These approaches are valuable, but they do not directly test whether category-level geometry aligns with an external expert reference.

Our setting requires this representation-level test. We examine whether category prototypes, similarity relations, and boundary regions form a structured space consistent with expert-defined symptom relations. Prototype distances quantify within-category cohesion and boundary-prone instances, while topic-based probes identify lexical fields associated with boundary blurring or sharpening. Together, these analyses complement global RSA by showing how local regions of the embedding space are organized, rather than serving as standalone explanations of individual predictions.

2.3 Mental-health language as a structured but confounded testbed

Mental-health language is a useful testbed for expert-aligned representation analysis because it is both structured and confounded. Mental-health-related categories involve symptom overlap, graded boundaries, and clinically meaningful distinctions, making them suitable for evaluating fine-grained category organization. Prior work reports systematic semantic, syntactic, and referential variation in psychiatric and mental-health language, with different computational representations capturing different aspects of this variation (He et al., 2024; Palominos et al., 2024).

At the same time, online communities introduce strong non-clinical confounds. Similarity between subreddits may reflect lexical overlap, discourse conventions, community-specific jargon, or style rather than expert-relevant relations (Nikolaev and Padó, 2023). Categories may also appear close because they share affective profiles in valence, arousal, or dominance, not because they share symptom-level structure (Marjieh et al., 2024). Because these factors are intertwined with the relational structure of interest, they cannot simply be removed. This setting therefore allows us to ask whether embedding geometry remains aligned with an expert reference after controlling for coarse domain separation and affective, stylistic, and discourse-level confounds.

3 Method

Overview. We encode each document with a pretrained encoder, optionally fine-tune the encoder on community labels, and analyze the resulting category-level geometry in both zero-shot and fine-tuned spaces. Our primary analyses compare model-derived category structure with expert-defined and domain-control reference structures using representational similarity analysis (RSA). We then use prototype-based typicality, boundary ambiguity, and topic-based probes to interpret local cohesion, boundary confusion, and similarity patterns in the learned spaces. Appendix A provides a schematic overview of the full analysis pipeline.

Embedding spaces. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote documents x_i with category labels $y_i \in \{1, \ldots, C\}$, where $C = 28$ consists of 17 mental-health categories and 11 domain-control communities. A pretrained encoder f_{θ_0} maps each document to a d-dimensional zero-shot embedding, $\mathbf{h}_i^{\mathrm{ZS}} = f_{\theta_0}(x_i)$. For the fine-tuned setting, we optimize the encoder with a linear classifier under cross-entropy loss and use the resulting encoder $f_{\theta^\star}$ to obtain $\mathbf{h}_i^{\mathrm{FT}} = f_{\theta^\star}(x_i)$ (Appendix B). All downstream analyses are performed in both representation settings.

Category-level representational geometry and reference alignment. For each category c, we construct a prototype by averaging document embeddings:

$$\boldsymbol{\mu}_c = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \mathbf{h}_i, \qquad \mathcal{I}_c = \{i : y_i = c\}. \quad (1)$$

The category prototypes define a representational dissimilarity matrix (RDM),

$$\mathrm{RDM}_{\mathrm{LLM}}(c, c') = d(\boldsymbol{\mu}_c, \boldsymbol{\mu}_{c'}), \quad (2)$$

where $c' \neq c$ denotes another category and d denotes cosine distance. We compare $\mathrm{RDM}_{\mathrm{LLM}}$ with two reference matrices: an expert-derived symptom matrix $\mathrm{RDM}_{\mathrm{Expert}}$ and a domain-control matrix $\mathrm{RDM}_{\mathrm{Domain}}$. For any two RDMs $A, B \in \mathbb{R}^{C \times C}$, RSA is defined as the Spearman rank correlation between their upper-triangular entries, excluding the diagonal:

$$\mathrm{RSA}(A, B) \;=\; \rho_{\mathrm{Spearman}}\big(\mathrm{vec}_{\triangle}(A),\, \mathrm{vec}_{\triangle}(B)\big). \quad (3)$$

We assess significance by permuting category labels and estimate uncertainty with bootstrap resampling over documents (Appendix C).

Prototype-based typicality and boundary analysis. We analyze local structure through document typicality, prototype margins, and expert-induced boundary ambiguity. The *prototype-based typicality distance* of document x_i is

$$\tau_i = d(\mathbf{h}_i, \boldsymbol{\mu}_{y_i}), \quad (4)$$

where smaller values indicate greater centrality within the labeled category. Category-level cohesion is summarized as $\mathrm{Coh}(c) = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \tau_i$. The *prototype margin* of document x_i is

$$m_i = \min_{c' \neq y_i} d(\mathbf{h}_i, \boldsymbol{\mu}_{c'}) - d(\mathbf{h}_i, \boldsymbol{\mu}_{y_i}), \quad (5)$$

with $m_i < 0$ indicating a boundary-confusion instance. We further define the *boundary alignment score*

$$\Delta_{\mathrm{align}} = \mathrm{Amb}_{+} - \mathrm{Amb}_{-}, \quad (6)$$

where Amb_{+} and Amb_{-} denote mean boundary ambiguity over expert-neighbor and non-neighbor category pairs, respectively. Larger Δ_{align} indicates that residual boundary uncertainty is concentrated among category pairs that are nearby in expert structure. Appendix D gives the full definitions of pairwise ambiguity and the expert-neighborhood graph.

Idealized analysis under a class-conditional Gaussian model. To clarify the geometric meaning of these measures, we consider a class-conditional Gaussian model in which each category c has latent mean $\boldsymbol{\mu}_c^{\star}$ and isotropic covariance $\sigma_c^2 I_d$:

$$\mathbf{h} \mid y = c \;\sim\; \mathcal{N}(\boldsymbol{\mu}_c^{\star},\, \sigma_c^2 I_d). \quad (7)$$

Under this model, we obtain two results that connect prototype geometry to cohesion and boundary confusion (proofs in Appendix E).

Proposition 1 (Typicality and within-category variance). *For any category c,*

$$\mathbb{E}\big[\|\mathbf{h} - \boldsymbol{\mu}_c^{\star}\|^2 \mid y = c\big] = d\,\sigma_c^2. \quad (8)$$

Thus, expected squared typicality is proportional to within-category variance; smaller typicality corresponds to tighter cohesion.

Proposition 2 (Boundary confusion and prototype separation). *Under two-class isotropic Gaussians with common variance $\sigma^2 I_d$ and prototype separation $\Delta_{c,c'} = \|\boldsymbol{\mu}_c^{\star} - \boldsymbol{\mu}_{c'}^{\star}\|$, the probability that a sample from category c falls on the wrong side of the prototype boundary is*

$$\mathbb{P}\big(S(\mathbf{h}; c, c') < 0 \mid y = c\big) = \Phi\left(-\frac{\Delta_{c,c'}}{2\sigma}\right), \quad (9)$$

where Φ is the standard normal CDF. Boundary confusion decreases as prototype separation increases and increases as within-category variance grows.

Corollary 1 (Fine-tuning increases boundary alignment). *Suppose supervision increases $\Delta_{c,c'}/\sigma$ for non-neighbor pairs while not substantially decreasing it for expert-neighbor pairs. Then, by Proposition 2, boundary confusion becomes more concentrated on expert-neighbor pairs, implying*

$$\Delta_{\mathrm{align}}^{\mathrm{FT}} > \Delta_{\mathrm{align}}^{\mathrm{ZS}}. \quad (10)$$

Synthetic boundary realignment experiment. To illustrate Corollary 1, we construct a six-category, two-dimensional Gaussian setting with an explicit expert-neighborhood graph (Figure 2). These six categories are synthetic and do not correspond to any of the 17 real mental-health categories; the simulation is intended only to illustrate the geometric behavior of boundary realignment under controlled conditions. Adding coarse domain distortion and extra within-category noise to the reference space produces diffuse boundary ambiguity in the zero-shot setting. After simulated fine-tuning, off-structure ambiguity is suppressed while

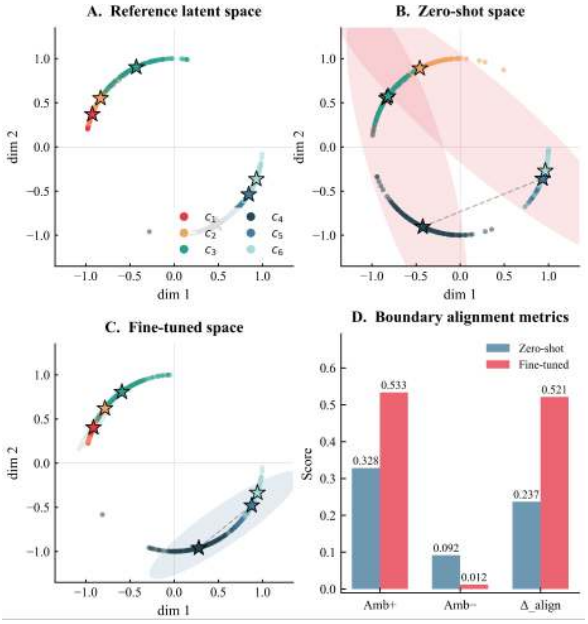


Figure 2: **Synthetic boundary realignment. All six categories are synthetic and do not correspond to the real mental-health categories.** (A) Reference latent space induced by a synthetic expert-neighborhood graph. (B) Zero-shot space after coarse domain distortion and increased within-category variance, producing diffuse boundary ambiguity. (C) Fine-tuned space after structured supervision, where off-structure ambiguity is reduced and residual ambiguity concentrates among expert-neighbor pairs. (D) Summary metrics: fine-tuning suppresses non-neighbor ambiguity (Amb_-), while residual uncertainty remains concentrated among expert-neighbor pairs (Amb_+), increasing the boundary alignment score Δ_{align}.

residual ambiguity remains concentrated among expert-neighbor pairs. Quantitatively, the boundary alignment score rises from 0.237 to 0.521, and mean category cohesion tightens from 1.160 to 0.979. The effect of supervision is therefore not to eliminate all boundaries uniformly, but to reduce off-structure ambiguity and concentrate residual uncertainty where expert structure predicts it.

Topic-based interpretability probes. To interpret local boundary structure, we fit an Embedded Topic Model (ETM) in each representation setting to estimate document-topic proportions and topic-word distributions. We summarize topics by their highest-probability keywords and assign each topic to the category in which it is most prevalent. This provides a topic-level account of cross-category proximity and boundary-related lexical overlap. The ETM objective is given in Appendix F.

4 Experiments

4.1 Dataset and Representation Settings

We use the *Reddit Mental Health Dataset*, which contains posts and text-derived features from 28 subreddits spanning 2018–2020 and was curated to study language change in mental-health communities before and during the early COVID-19 period (Low et al., 2020). The dataset includes 15 disorder- or support-specific mental-health subreddits (e.g., r/depression, r/anxiety, r/suicidewatch), 2 broad mental-health forums (r/mentalhealth, r/COVID19_support), and 11 non-mental-health control subreddits (e.g., r/legaladvice, r/personalfinance, r/jokes, r/fitness).

We treat each subreddit as a category label ($C = 28$) and analyze category-level geometry over mental-health and control communities. To reduce subreddit-size imbalance, we construct balanced analysis sets by stratified subsampling and estimate uncertainty with bootstrap resampling, following the dataset authors' recommendation. This dataset is well suited to our question because it combines expert-relevant symptom relations with strong contextual signals, including subreddit-specific jargon, discourse conventions, and a broad mental-health/control distinction. It allows us to test whether learned geometry reflects expert-defined category structure rather than category separability or coarse domain cues alone.

4.2 Embeddings and Topic Modeling

4.2.1 Zero-shot and fine-tuned embedding spaces

We compare two representation settings at two model scales: pretrained zero-shot spaces and supervised fine-tuned spaces for Qwen3-Embedding-0.6B and Qwen3-Embedding-4B. In the fine-tuned setting, each model is optimized on the 28-community classification objective with a linear head and cross-entropy loss. We then discard the classification head and use the encoder outputs for all downstream geometry analyses, including category prototypes, RDM estimation, RSA, and local boundary analysis.

Following the model specification, documents are represented as fixed-length 1024-dimensional vectors for Qwen3-Embedding-0.6B. The training curve is reported in Appendix G: accuracy rises rapidly and stabilizes around 0.83, while loss remains low with moderate batch-level variation. Qwen3-Embedding-4B follows a similar trajectory

and reaches a higher accuracy of ∼0.87, consistent with its larger representational capacity. We treat these fine-tuned models not as classification endpoints, but as task-conditioned embedding spaces that can be compared with their zero-shot counterparts and across scales to examine how supervision and model size reshape category geometry.

4.2.2 Topic-based interpretability probe

As an auxiliary interpretability probe, we estimate topic structure in each representation setting with an Embedded Topic Model (ETM) and compare it with bag-of-words LDA. Table 1 reports topic diversity (TD), inverted rank-biased overlap (iRBO), normalized pointwise mutual information (NPMI), and topic exclusivity (Excl), together with adjusted Rand index (ARI) and normalized mutual information (NMI) for alignment between topic-induced clusters and subreddit labels. Higher TD and iRBO values indicate greater topic diversity or separation, higher NPMI indicates greater coherence, higher Excl indicates greater topic exclusivity, and higher ARI and NMI indicate stronger alignment with subreddit labels. Fine-tuning improves topic structure across metrics: topic diversity and exclusivity increase, NPMI becomes less negative, and topic clusters align more closely with subreddit labels. LDA serves as a conservative lexical reference. Accordingly, the topic-model analysis is used as an auxiliary qualitative probe for interpreting local boundary structure; the primary quantitative evidence for expert alignment comes from RSA, prototype-based boundary analysis, and confound-controlled partial RSA.

Table 1: **Topic quality and label alignment.** ETM results for zero-shot vs fine-tuned embeddings across model scales, with an LDA baseline.

Model	Setting	TD	iRBO	NPMI	Excl	ARI	NMI
—	LDA	0.112	0.336	-0.308	0.104	0.176	0.421
0.6B	Zero-shot	0.293	0.342	-0.182	0.203	0.230	0.448
	Fine-tuned	0.347	0.418	-0.057	0.354	0.319	0.580
4B	Zero-shot	0.318	0.371	-0.134	0.247	0.271	0.487
	Fine-tuned	0.389	0.453	-0.021	0.412	0.368	0.634

4.3 Construction of the expert reference structure

Our alignment analyses require an external reference over the same 28 categories. We construct this reference from a symptom vocabulary $\mathcal{M}$, consisting of expert-informed dimensions that capture broad symptom and behavioral features of the mental-health-related communities. Each category c is encoded as a binary symptom-profile vector $\mathbf{s}_c \in \{0,1\}^{|\mathcal{M}|}$, where each entry indicates the presence or absence of a symptom dimension. The coding is multi-label: a category may activate multiple symptom dimensions, allowing overlapping communities to share partial symptom profiles rather than forcing them into mutually exclusive classes. A symptom dimension is included when it is supported by the official community description, sidebar, or pinned rules, applicable DSM-5 criteria, or the relevant RDoC construct, while incidental conversational mentions are excluded. For r/COVID19_support, we code the mental-health symptoms that characterize the support-community context rather than general physical symptoms caused by COVID-19. Two annotators independently coded all $28 \times 14 = 392$ category–dimension entries under these criteria, with disagreements adjudicated by a third reviewer. Overall agreement was 94.1%, with Cohen's $\kappa = 0.81$; full per-dimension reliability statistics are reported in Appendix H.4. Collecting these vectors yields a binary symptom matrix $S \in \{0,1\}^{C \times |\mathcal{M}|}$, shown in Figure 3, with rows corresponding to subreddit categories and columns to symptom dimensions. Control communities are assigned all-zero profiles because the reference vocabulary is restricted to targeted mental-health symptom dimensions; this does not imply an absence of distress-related discussions. We derive the expert representational dissimilarity matrix (RDM) using pairwise Jaccard distance between symptom profiles, $\mathrm{RDM}_{\mathrm{Expert}}(c, c') = d_J(\mathbf{s}_c, \mathbf{s}_{c'})$, where d_J denotes Jaccard distance on binary symptom sets.

The resulting expert RDM provides an intentionally asymmetric contrast: mental-health categories are organized by shared symptom dimensions, whereas control categories lie outside the reference symptom vocabulary by design. Figure 3 reports the full binary matrix; Appendix H provides the symptom dimensions, coding criteria, and category assignments; and Appendix H.4 reports the annotation protocol and inter-rater reliability statistics, including overall Cohen's $\kappa = 0.81$.

4.4 Global alignment with expert and domain reference structures

We first evaluate whether the category geometry induced by the real data aligns with the expert reference beyond coarse domain separation. For each representation setting, we summarize

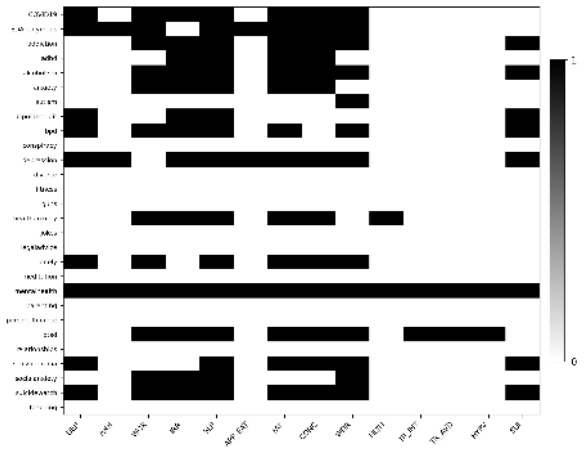


Figure 3: **Binary expert reference matrix.** Rows correspond to subreddit categories and columns to symptom dimensions in $\mathcal{M}$. Black cells indicate presence (1) and white cells indicate absence (0). Control communities are assigned the all-zero vector by construction.

Table 2: Spearman correlation between reference RDMs. †The domain RDM is constant on the MH-only subset and therefore undefined for correlation.

RDM pair	**Full** 28×28	**MH-only** 17×17
Expert vs. Domain	0.71	n/a†
Expert vs. VAD	0.18	0.21
Domain vs. VAD	0.23	n/a†

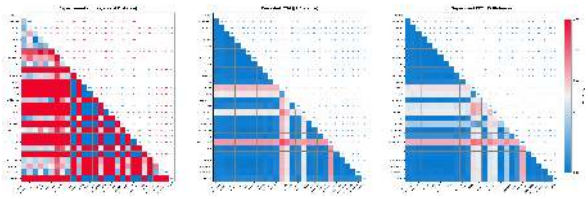


Figure 4: **Three-way comparison of representational geometry.** Left: expert reference RDM (Jaccard distance). Middle: zero-shot ETM RDM (JS distance). Right: fine-tuned ETM RDM (JS distance).

model geometry with a representational dissimilarity matrix (RDM). Specifically, we compute a prototype for each subreddit category by averaging its document representations and define $\mathrm{RDM}_{\mathrm{Model}}(c, c') = d(\boldsymbol{\mu}_c, \boldsymbol{\mu}_{c'})$, where d is cosine distance for embedding prototypes and Jensen–Shannon (JS) distance for topic-distribution prototypes (Figure 4).

We quantify alignment using RSA (Eq. 3) between $\mathrm{RDM}_{\mathrm{Model}}$ and $\mathrm{RDM}_{\mathrm{Expert}}$. Because the expert and domain RDMs share substantial block-level structure on the full category set (Table 2), we report three complementary analyses: (i) MH-only RSA over the 17 mental-health categories, where the domain confound is absent by construction and which we treat as the **primary** result; (ii) full-set RSA over all 28 categories; and (iii) partial RSA over the full set, controlling for a domain RDM that encodes MH versus control membership. Uncertainty is estimated by bootstrap resampling over documents when forming prototypes, and significance is evaluated with category-level permutation tests that jointly permute RDM row and column labels. Results are summarized in Table 3.

Reference RDM collinearity. Before interpreting RSA values, we assess pairwise collinearity among the reference RDMs (Table 2). On the full 28×28 category set, the expert and domain RDMs are strongly correlated ($\rho = 0.71$), reflecting the all-zero symptom profiles assigned to control communities. Thus, full-set RSA may conflate fine-grained symptom-level alignment with coarse MH-versus-control separation. We therefore treat the **MH-only** 17×17 **analysis as the primary RSA result**, since the domain RDM is uninformative within this subset, and interpret the full-set and partial-RSA results as complementary evidence. The VAD reference is only weakly correlated with the expert ($\rho \leq 0.21$) and domain ($\rho = 0.23$) RDMs, consistent with the multi-baseline confound analysis in Section 4.6.

On the MH-only subset, pretrained embeddings already show substantial alignment with the expert reference ($\rho = 0.541$ at 0.6B, 0.628 at 4B). Supervision produces the largest gains in this setting ($\Delta\rho = +0.171$ at 0.6B, $+0.135$ at 4B), exceeding the gains observed in the full-set ($\Delta\rho \leq +0.055$) and partial-RSA ($\Delta\rho \leq +0.084$) analyses. These results indicate that fine-tuning most strongly amplifies expert-aligned structure at the mental-health category level, while full-set results should be interpreted in light of the expert–domain collinearity noted above.

4.5 Local boundary organization and interpretability

Beyond global RSA, we examine local category organization through prototype-based typicality and topic-based interpretability probes. We ask which categories are internally cohesive, which exhibit diffuse boundary regions, and whether ambiguity is concentrated along expert-neighbor relations. For each document, typicality is defined as its distance to the prototype of its labeled category. Larger dis-

Table 3: RSA alignment between model and expert RDMs. Spearman ρ is computed between each model-derived RDM and the expert RDM across two model scales (0.6B and 4B) and two representation settings (zero-shot and supervised fine-tuned). The full 28×28 analysis includes all categories; the MH-only 17×17 analysis, our primary result, is restricted to mental-health categories; partial RSA controls for the domain RDM. 95% confidence intervals are estimated by bootstrap resampling over documents.

Analysis	Setting	ρ	95% CI
Full (28×28)	0.6B Zero-shot	0.703	[0.658, 0.741]
	0.6B Supervised	**0.758**	[0.728, 0.781]
	4B Zero-shot	0.735	[0.695, 0.766]
	4B Supervised	**0.789**	[0.762, 0.812]
MH-only (17×17)	0.6B Zero-shot	0.541	[0.404, 0.658]
	0.6B Supervised	**0.712**	[0.617, 0.791]
	4B Zero-shot	0.628	[0.506, 0.731]
	4B Supervised	**0.763**	[0.675, 0.834]
Partial (28×28)	0.6B Zero-shot	0.567	[0.481, 0.638]
	0.6B Supervised	**0.651**	[0.581, 0.706]
	4B Zero-shot	0.609	[0.526, 0.679]
	4B Supervised	**0.682**	[0.615, 0.741]



Figure 5: **Topic-based interpretability of local boundary organization. (ZS vs. FT).** Representative ETM topics associated with boundary-blurring and boundary-sharpening category pairs in zero-shot (ZS) and fine-tuned (FT) spaces. Boundary-blurring topics reflect shared linguistic content across nearby categories, whereas boundary-sharpening topics show more concentrated category-specific lexical fields.

tances indicate boundary-prone instances, as these documents lie farther from their category center and closer to competing prototypes.

To link boundary regions to interpretable linguistic patterns, we examine ETM topics associated with boundary-blurring and boundary-sharpening category pairs in zero-shot and fine-tuned representations for the 0.6B model (Figure 5). Boundary-blurring topics emphasize shared affective or situational narratives, whereas boundary-sharpening topics contain concentrated domain-specific lexical fields, consistent with clearer separations among control communities. The 4B model shows similar patterns, with finer topic distinctions aligned with its stronger global RSA alignment. Overall, supervised fine-tuning reshapes the pretrained space by reducing off-structure ambiguity and concentrating residual boundary uncertainty among expert-neighbor pairs, rather than uniformly increasing separability across categories.

4.6 Confound Analyses: Affective, Discourse, and Stylistic Baselines

We next test whether model–expert alignment can be explained by non-clinical regularities. Rather than relying on an affective baseline alone, we evaluate four reference structures capturing increasingly content-rich confounds: (i) **VAD** (valence, arousal, dominance), computed from the NRC VAD lexicon as a coarse affective signal (Mohammad, 2018); (ii) **LIWC** affect, social, and cognitive category proportions (Boyd et al., 2022); (iii) document-level **lexical-style** features, including sentence length, type-token ratio, and function-word frequency, following stylometric practice (Stamatatos, 2009); and (iv) the **ETM topic-distribution RDM** as a discourse-content baseline. For each baseline, we report $\rho(\text{Model}, \text{Confound})$ and the partial RSA between model and expert RDMs after controlling for it.

Table 4 shows two main patterns. First, VAD is not a strong confound: the model–VAD association is weak ($\rho \leq 0.20$), and controlling for it changes expert alignment by less than 0.02. VAD therefore serves as a coarse affective check, indicating that broad affect does not drive the observed geometry, but it is not the strongest nuisance signal. Second, more content-rich baselines explain a larger share of the alignment but do not eliminate it: LIWC affect+social reduces ρ by 0.16, and ETM topic structure reduces it by 0.29. Even under the strongest discourse-content control, residual expert alignment remains substantial and significant, indicating that fine-grained category geometry is not reducible to topic structure alone. Across all four baselines, fine-tuning further weakens confound associations; for example, model–LIWC drops from 0.46 ZS to 0.43 FT at 0.6B and from 0.43 ZS to 0.41 FT at 4B. Larger model scale further weakens these associations, suggesting that supervision and scale jointly shift the representation space toward more category-specific organization.

Table 4: Multi-baseline confound analysis. Columns under $\rho(\mathrm{M}, \mathrm{C})$ report model–confound associations; columns under $\rho(\mathrm{M}, \mathrm{E} \mid \mathrm{C})$ report partial RSA between model geometry and the expert RDM after controlling for each confound. Bold indicates the strongest model–confound association and the largest reduction in expert alignment within each representation setting.

	$\rho(\mathrm{M}, \mathrm{C})$				$\rho(\mathrm{M}, \mathrm{E} \mid \mathrm{C})$			
Confound	0.6B ZS	0.6B FT	4B ZS	4B FT	0.6B ZS	0.6B FT	4B ZS	4B FT
VAD	0.20	0.12	0.19	0.10	**0.68**	**0.74**	**0.71**	**0.76**
Lexical style	0.31	0.26	0.30	0.28	0.64	0.70	0.66	0.71
LIWC affect+social	0.46	0.43	0.43	0.41	0.57	0.61	0.59	0.62
ETM topic dist.	**0.62**	**0.59**	**0.60**	**0.58**	0.42	0.47	0.45	0.49

Overall interpretation. Taken together, the global, local, and confound-controlled analyses point to a selective form of structural recovery. The largest supervision-induced gains occur in the MH-only RSA, indicating that fine-tuning strengthens within-domain symptom relations rather than merely amplifying the coarse MH-versus-control split. The boundary analysis is consistent with the same interpretation: supervision reduces off-structure ambiguity while leaving residual uncertainty concentrated among expert-neighbor categories. Finally, substantial partial RSA remains after controlling for increasingly content-rich baselines. These results therefore support viewing expert alignment as a graded property of representation geometry, rather than as a consequence of classification accuracy alone or as evidence of complete recovery of the expert reference.

5 Conclusion

We examined whether LLM embeddings recover expert-informed category structure in mental-health language. Instead of treating community-classification accuracy as evidence of semantic organization, we evaluated embeddings as representational spaces using RSA, prototype-based boundary analysis, topic-based interpretation, and multi-baseline confound controls across two scales. Results support a selective, level-dependent account of expert-structure recovery: pretrained embeddings show non-trivial alignment with the expert reference beyond the mental-health/control split, including fine-grained relations within the mental-health subset; fine-tuning strengthens this structure most at the finest category level; and larger scale improves zero-shot alignment and supervision-induced gains. Confound analyses show that alignment is not explained by coarse affect (VAD), LIWC-based discourse signals, lexical style, or topic structure: even under the strongest discourse-content control, residual expert alignment remains substantial, while fine-tuning and scale reduce non-clinical confound associations.

Together, these findings indicate that, within the evaluated mental-health setting, LLM embeddings can encode expert-relevant structure, but recovery is partial, level-dependent, and shaped by supervision and scale. Whether the same pattern extends to other expert-structured domains remains an open question. More broadly, in domains with graded categories and strong confounds, expert alignment should be evaluated with explicit reference structures, level-stratified analyses, and non-clinical baseline controls rather than inferred from classification performance or global similarity alone.

Code availability. Code and implementation resources for this work are publicly available in the THETA repository.

Limitations

Our expert reference is a structured category-level benchmark rather than a clinical diagnostic ontology. The binary symptom profiles abstract away individual-level diagnosis, symptom severity, comorbidity, temporal variation, and heterogeneity within communities. They also treat symptom dimensions as binary features rather than encoding their prevalence or relative clinical importance. Accordingly, positive alignment should be interpreted as partial structural correspondence rather than recovery of an underlying clinical taxonomy. Although the annotation procedure achieves high inter-rater reliability ($\kappa = 0.81$), agreement between annotators establishes coding consistency rather than independent clinical validity of the reference itself. Consistent with this limited interpretation, zero-shot MH-only alignment remains moderate and increases substantially after supervision, indicating that the recovered structure is meaningful but incomplete and partly task-conditioned.

The treatment of control communities introduces an additional design constraint. Assigning all-zero symptom profiles provides a structured contrast outside the targeted symptom vocabulary, but it also induces coarse block structure shared with the MH-versus-control domain partition. We mitigate this dependence by treating MH-only RSA as the primary analysis and by reporting domain-controlled partial RSA on the full set. Nevertheless, the measured alignment remains partly dependent on how the external reference is constructed, and

alternative expert representations could yield different fine-grained geometries.

The data further limit interpretation. Reddit communities are self-organized rather than clinically verified populations, and their language combines expert-relevant relations with community-specific discourse, stylistic, affective, platform-specific, and other distributional signals. Our confound analyses address several major alternatives but cannot exhaust all possible sources of structure. In addition, the dataset covers a specific platform and time period (Reddit, 2018–2020), including the early COVID-19 period. Community vocabulary, norms, and symptom-related discourse may shift over time or across platforms, so the stability of the observed geometry under temporal or platform shift remains untested.

Our sampling procedure introduces a related limitation. Stratified subsampling reduces category-size imbalance and facilitates category-level comparison, but it also removes information about the natural prevalence of communities in the underlying corpus. The resulting geometry should therefore be interpreted as category-balanced rather than prevalence-weighted, and alternative sampling schemes may shift prototype locations or uncertainty estimates.

Generality is also limited because the two evaluated model scales belong to the same model family and are tested on one dataset under one supervised fine-tuning objective. Robustness across architectures, training objectives, datasets, domains, and less community-specific language sources therefore remains open. Future work should evaluate broader model families, more controlled clinical or semi-clinical corpora, and richer expert references incorporating severity, hierarchy, prevalence, or multiple raters.

Finally, the topic-based analyses are interpretive rather than causal. They help characterize lexical fields associated with local boundary structure but do not identify the mechanisms that produce expert-aligned geometry. More broadly, the alignment reported here should not be interpreted as evidence that the models possess clinical reasoning ability or recover a latent diagnostic ontology.

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10(1-47):6.

Shael Brown and Reza Farivar. 2025. The topology of representational geometry. *Frontiers in Neuroscience*, 19:1597899.

Francesca Carota, Hamed Nili, Nikolaus Kriegeskorte, and Friedemann Pulvermüller. 2024. Experientially-grounded and distributional semantic vectors uncover dissociable representations of conceptual categories. *Language, Cognition and Neuroscience*, 39(8):1020–1044.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Fifth Edition and 1 others. 2013. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21(21):591–643.

Scott L Fairhall. 2024. Sentence-level embeddings reveal dissociable word-and sentence-level cortical representation across coarse-and fine-grained levels of meaning. *Brain and Language*, 250:105389.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5553–5563.

Rui He, Claudio Palominos, Han Zhang, Maria Francisca Alonso-Sánchez, Lena Palaniyappan, and Wolfram Hinzen. 2024. Navigating the semantic space: Unraveling the structure of meaning in psychosis using different computational language models. *Psychiatry Research*, 333:115752.

Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Neehar Kondapaneni, Oisin Mac Aodha, and Pietro Perona. 2025. Representational similarity via interpretable visual concepts. In *International Conference on Learning Representations*, volume 2025, pages 12850–12881.

Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S Ghosh. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.

Raja Marjieh, Ilia Sucholutsky, Pol Van Rijn, Nori Jacoby, and Thomas L Griffiths. 2024. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Dmitry Nikolaev and Sebastian Padó. 2023. Investigating semantic subspaces of transformer sentence embeddings through linear structural probing. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 142–154.

Mattson Ogg, Ritwik Bose, Jamie Scharf, C Ratto, and Michael Wolmetz. 2024. Turing representational similarity analysis (rsa): A flexible method for measuring alignment between human and artificial intelligence. *arXiv preprint arXiv:2412.00577*.

Juri Opitz and Anette Frank. 2022. Sbert studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638.

Juri Opitz, Lucas Moeller, Andrianos Michail, Sebastian Padó, and Simon Clematide. 2025. Interpretable text embeddings and text similarity explanation: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22314–22330.

Claudio Palominos, Rui He, Karla Fröhlich, Rieke Roxanne Mülfarth, Svenja Seuffert, Iris E Sommer, Philipp Homan, Tilo Kircher, Frederike Stein, and Wolfram Hinzen. 2024. Approximating the semantic space: word embedding techniques in psychiatric speech analysis. *Schizophrenia*, 10(1):114.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Raj Sanjay Shah, Khushi Bhardwaj, and Sashank Varma. 2024. Development of cognitive intelligence in pretrained language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9657.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Siddhartha K Vemuri, Raj Sanjay Shah, and Sashank Varma. 2024. How well do deep learning models capture human concepts? the case of the typicality effect. *arXiv preprint arXiv:2405.16128*.

Qihui Xu, Yingying Peng, Samuel A Nastase, Martin Chodorow, Minghua Wu, and Ping Li. 2025. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature human behaviour*, 9(9):1871–1886.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix

A Overview of the representation analysis pipeline

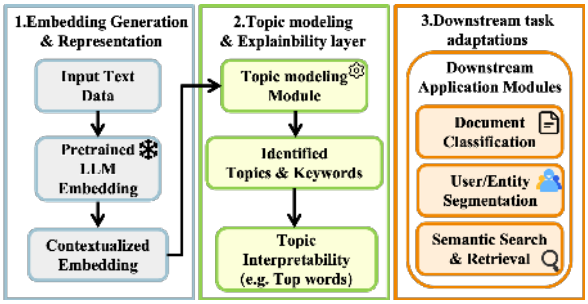


Figure 6: **Overview of the representation analysis framework.** Input text is encoded into document embeddings in either the pretrained or supervised fine-tuned space. Document embeddings are aggregated into category prototypes, from which we derive category-level representational geometry. We then evaluate alignment against expert-defined and domain-control reference structures, and use prototype-based typicality and topic-based probes to interpret similarity patterns, cohesion, and boundary regions in the embedding space.

B Embedding Space Details

The pretrained encoder f_{θ_0} maps each document to a d-dimensional embedding:

$$\mathbf{h}_i^{\mathrm{ZS}} = f_{\theta_0}(x_i) \in \mathbb{R}^d. \tag{11}$$

The fine-tuned embedding is obtained by jointly optimizing the encoder parameters θ together with a linear classification head with parameters (W, b) under cross-entropy loss:

$$\min_{\theta, W, b} -\sum_{i=1}^{N} \log \operatorname{softmax}\big(W f_\theta(x_i) + b\big)_{y_i}, \tag{12}$$

yielding fine-tuned embeddings $\mathbf{h}_i^{\mathrm{FT}} = f_{\theta^\star}(x_i)$, where $\theta^\star$ denotes the optimized encoder parameters at convergence. The classifier head (W, b) is discarded after training; only the encoder is used for downstream geometry analysis.

C Representational Dissimilarity Matrix and RSA

For local prototype-based analyses, document embeddings are first L_2-normalized, category prototypes are formed by averaging normalized embeddings, and the resulting prototypes are again L_2-normalized before computing cosine distance. Under this normalization, squared Euclidean and cosine distances are related by

$$\|\mathbf{u}-\mathbf{v}\|_2^2 = 2\big(1-\cos(\mathbf{u},\mathbf{v})\big), \qquad \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1, \tag{13}$$

so they induce the same nearest-prototype ordering and prototype boundary.

Following Eq. 3 in the main text, alignment between two RDMs A and B is quantified by

$$\rho(A, B) \;=\; \rho_{\mathrm{Spearman}}\big(\mathrm{vec}_\triangle(A),\, \mathrm{vec}_\triangle(B)\big), \tag{14}$$

where $\mathrm{vec}_\triangle(\cdot)$ extracts the upper-triangular entries of a square matrix, excluding the diagonal.

D Boundary Ambiguity and Alignment Score

Pairwise ambiguity. For documents in category c, define the relative-distance gap toward category c' as

$$g_i(c, c') = d(\mathbf{h}_i, \boldsymbol{\mu}_{c'}) - d(\mathbf{h}_i, \boldsymbol{\mu}_c), \qquad i \in \mathcal{I}_c. \tag{15}$$

With tolerance $\varepsilon > 0$, the directed boundary ambiguity is

$$\mathrm{Amb}(c, c') = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \mathbf{1}\big(|g_i(c, c')| \le \varepsilon\big), \tag{16}$$

and the symmetric score is $\mathrm{Amb}_{\mathrm{sym}}(c, c') = \frac{1}{2}\big(\mathrm{Amb}(c, c') + \mathrm{Amb}(c', c)\big)$.

Expert neighborhood graph. Let $\mathcal{N}_K^{\mathrm{Exp}}(c)$ denote the top-K nearest categories to c under $\mathrm{RDM}_{\mathrm{Expert}}$. The undirected expert-neighborhood indicator is

$$A_{c,c'}^{\mathrm{Exp,sym}} = \mathbf{1}\Big(c' \in \mathcal{N}_K^{\mathrm{Exp}}(c) \;\text{ or }\; c \in \mathcal{N}_K^{\mathrm{Exp}}(c')\Big). \tag{17}$$

Boundary alignment score.

$$\mathrm{Amb}_+ = \frac{\sum_{c<c'} A_{c,c'}^{\mathrm{Exp,sym}}\, \mathrm{Amb}_{\mathrm{sym}}(c, c')}{\sum_{c<c'} A_{c,c'}^{\mathrm{Exp,sym}}}, \tag{18}$$

$$\mathrm{Amb}_- = \frac{\sum_{c<c'} (1 - A_{c,c'}^{\mathrm{Exp,sym}})\, \mathrm{Amb}_{\mathrm{sym}}(c, c')}{\sum_{c<c'} (1 - A_{c,c'}^{\mathrm{Exp,sym}})}, \tag{19}$$

$$\Delta_{\mathrm{align}} = \mathrm{Amb}_+ - \mathrm{Amb}_-. \tag{20}$$

E Proofs of Propositions and Corollary

Proof of Proposition 1

Let $\mathbf{z} = \mathbf{h} - \boldsymbol{\mu}_c^\star$. Since $\mathbf{h} \mid y = c \sim \mathcal{N}(\boldsymbol{\mu}_c^\star, \sigma_c^2 I_d)$, we have $\mathbf{z} \mid y = c \sim \mathcal{N}(\mathbf{0}, \sigma_c^2 I_d)$. Then

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{z}\|^2 \mid y = c\big] &= \sum_{j=1}^{d} \mathbb{E}\big[z_j^2 \mid y = c\big] \\ &= \sum_{j=1}^{d} \sigma_c^2 = d\,\sigma_c^2. \end{aligned} \tag{21}$$

Proof of Proposition 2

For a sample $\mathbf{h}$ from category c, define

$$S(\mathbf{h}; c, c') = \|\mathbf{h} - \boldsymbol{\mu}_{c'}^\star\|^2 - \|\mathbf{h} - \boldsymbol{\mu}_c^\star\|^2. \tag{22}$$

Expanding the quadratic terms:

$$S(\mathbf{h}; c, c') = 2\mathbf{h}^\top(\boldsymbol{\mu}_c^\star - \boldsymbol{\mu}_{c'}^\star) + \|\boldsymbol{\mu}_{c'}^\star\|^2 - \|\boldsymbol{\mu}_c^\star\|^2. \tag{23}$$

Writing $\mathbf{h} = \boldsymbol{\mu}_c^\star + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$:

$$S(\mathbf{h}; c, c') = \|\boldsymbol{\mu}_c^\star - \boldsymbol{\mu}_{c'}^\star\|^2 + 2\boldsymbol{\varepsilon}^\top(\boldsymbol{\mu}_c^\star - \boldsymbol{\mu}_{c'}^\star) \tag{24}$$

$$\sim \mathcal{N}\big(\Delta_{c,c'}^2,\ 4\sigma^2\Delta_{c,c'}^2\big)\,. \tag{25}$$

Standardizing:

$$\begin{aligned}\mathbb{P}(S < 0 \mid y = c) &= \mathbb{P}\left(\frac{S - \Delta_{c,c'}^2}{2\sigma\Delta_{c,c'}} < -\frac{\Delta_{c,c'}}{2\sigma}\right) \\ &= \Phi\left(-\frac{\Delta_{c,c'}}{2\sigma}\right). \end{aligned} \tag{26}$$

Proof of Corollary 1

Since $\Phi(-\Delta_{c,c'}/2\sigma)$ is decreasing in $\Delta_{c,c'}/\sigma$, an increase in $\Delta_{c,c'}/\sigma$ for non-neighbor pairs lowers Amb_-. If the ratio does not decrease for expert-neighbor pairs, Amb_+ does not fall and may rise due to reduced global ambiguity. Therefore $\Delta_{\text{align}}^{\text{FT}} = \mathrm{Amb}_+^{\text{FT}} - \mathrm{Amb}_-^{\text{FT}} > \mathrm{Amb}_+^{\text{ZS}} - \mathrm{Amb}_-^{\text{ZS}} = \Delta_{\text{align}}^{\text{ZS}}$.

F ETM Objective

The Embedded Topic Model (ETM) is trained by minimizing the negative variational evidence lower bound:

$$\begin{aligned}\mathcal{L}_{\text{ETM}} = -\sum_{i=1}^{N}\Big(&\mathbb{E}_{q(\boldsymbol{\theta}_i)}\big[\log p(\mathbf{x}_i \mid \boldsymbol{\theta}_i, B)\big] \\ &- \mathrm{KL}(q(\boldsymbol{\theta}_i)\,\|\,p(\boldsymbol{\theta}_i))\Big). \end{aligned} \tag{27}$$

where $\Theta \in \mathbb{R}^{N\times K}$ and $B \in \mathbb{R}^{K\times V}$ denote the document-topic proportion and topic-word distribution matrices, respectively.

G Fine-tuning Dynamics

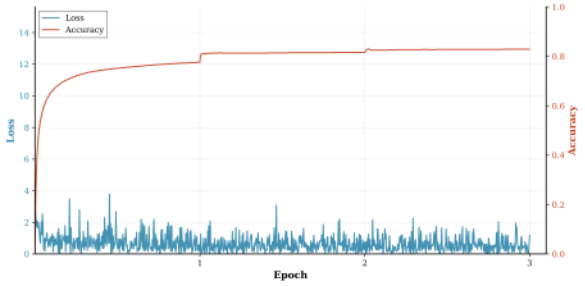


Figure 7: **Fine-tuning dynamics for Qwen3-Embedding-0.6B.** Training loss (left axis) and classification accuracy (right axis) over epochs. Accuracy rises quickly and then stabilizes around 0.83 by the end of training.

H Construction of the Expert Reference Structure

This appendix provides additional detail on the construction of the expert reference structure used in the main text. Our goal is not to posit a definitive ontology of mental-health-related categories, but to construct a transparent external reference against which category-level representation geometry can be compared.

H.1 Symptom dimensions

We define a symptom vocabulary $\mathcal{M}$ consisting of expert-informed dimensions intended to capture broad symptom and behavioral features associated with the mental-health-related communities in our dataset. These dimensions are designed to serve as coarse reference axes for category comparison rather than exhaustive diagnostic criteria. The dimensions used in this work are listed below:

- **DEP**: depressed mood / sadness
- **ANH**: anhedonia / loss of interest
- **WOR**: excessive worry / anxiety-related rumination
- **IRR**: irritability / emotional instability
- **SLP**: sleep disturbance
- **APP/EAT**: appetite or eating-related disturbance
- **FAT**: fatigue / low energy
- **CONC**: concentration or attentional difficulty

- **WDR**: withdrawal / social disengagement / loneliness-related isolation
- **HLTH**: health-focused anxiety or bodily concern
- **TR_INT**: trauma-related intrusion
- **TR_AVD**: trauma-related avoidance
- **HYPV**: hypervigilance / heightened arousal
- **SUI**: suicidality / self-harm risk

These dimensions are intentionally broad. They are used to induce a structured category-level reference matrix, not to provide fine-grained diagnostic adjudication.

Grounding of symptom dimensions. The 14 dimensions in $\mathcal{M}$ are grounded in established transdiagnostic frameworks for psychiatric symptomatology. The core affective and somatic dimensions (DEP, ANH, FAT, SLP, APP/EAT, CONC) correspond directly to diagnostic criteria specified in the *Diagnostic and Statistical Manual of Mental Disorders* (Edition et al., 2013), which provides the primary clinical reference for the mental-health communities in our dataset. The anxiety-related and hyperarousal dimensions (WOR, IRR, HYPV) and the trauma-specific dimensions (TR_INT, TR_AVD) follow the transdiagnostic symptom structure described in the Research Domain Criteria (RDoC) framework (Insel et al., 2010), which organizes psychiatric symptoms into functional domains cutting across diagnostic boundaries. The withdrawal and suicidality dimensions (WDR, SUI) reflect features consistently identified in computational and clinical studies of online mental-health language (Low et al., 2020). Together, these sources motivate the specific dimensions included in $\mathcal{M}$ and provide an external basis for treating the resulting binary profiles as expert-informed rather than arbitrary.

H.2 Coding principle

For each subreddit category c, we define a binary symptom-profile vector

$$\mathbf{s}_c \in \{0,1\}^{|\mathcal{M}|},$$

where each coordinate indicates whether the corresponding symptom dimension is treated as part of the expert-informed profile for that category.

A value of 1 indicates that the dimension is considered characteristic of the category-level profile; a value of 0 indicates that the dimension is not included in the reference profile for that category. Categories may activate multiple dimensions simultaneously. Thus, the reference structure captures overlap among categories through shared dimensions rather than assigning each category to a single exclusive type.

The resulting binary matrix

$$S \in \{0,1\}^{C\times|\mathcal{M}|}$$

is used to derive the expert reference RDM via pairwise Jaccard distance:

$$\mathrm{RDM}_{\mathrm{Expert}}(c,c') = d_J(\mathbf{s}_c, \mathbf{s}_{c'}).$$

H.3 Operational Coding Criteria

To support reproducibility and inter-rater reliability, we specify operational criteria for each symptom dimension. For category c, dimension $d \in \mathcal{M}$ is coded 1 iff at least one of the following holds: (i) the official community description, sidebar, or pinned rules explicitly reference d-relevant symptoms; (ii) d appears among the DSM-5 diagnostic criteria for the clinical entity that the community is centered on (where applicable); or (iii) d is identified as a defining functional construct in the RDoC matrix for the relevant domain. A dimension is coded 0 if none of (i)–(iii) is satisfied, even when d may be incidentally discussed by community members. This rule prioritizes *category-defining* over *conversationally co-occurring* symptoms. For example, bipolarreddit is coded with ANH, APP/EAT, FAT, and CONC following DSM-5 Criterion B for major depressive episodes; schizophrenia is coded with ANH following the DSM-5 negative-symptom cluster. The full category-by-dimension assignments are listed in Table 6.

H.4 Annotation Protocol and Inter-rater Reliability

To assess the reliability of the expert-informed reference, we conducted a two-annotator reliability study over the full $28 \times 14 = 392$-cell binary matrix.

Procedure. Two annotators (A and B) independently coded all 392 cells under the operational criteria in Appendix H.3, blinded to each other's decisions and to the embedding analyses. Annotator A holds a background in computational linguistics; Annotator B holds a background in clinical

Table 5: Inter-rater reliability over the 28×14 binary expert reference matrix. The annotators initially disagreed on 23 of 392 cells; all disagreements were resolved by adjudication.

Dimension	% Agreement	Cohen's κ
DEP	96.4	0.89
ANH	92.9	0.78
WOR	96.4	0.91
IRR	89.3	0.74
SLP	96.4	0.90
APP/EAT	96.4	0.86
FAT	92.9	0.79
CONC	89.3	0.72
WDR	85.7	0.69
HLTH	96.4	0.85
TR_INT	100.0	1.00
TR_AVD	100.0	1.00
HYPV	96.4	0.82
SUI	100.0	1.00
Overall	**94.1**	**0.81**

psychology. Disagreements were adjudicated by a third reviewer to produce the final consensus matrix used in all main analyses.

Reliability. Table 5 reports per-dimension percent agreement and Cohen's κ. Overall agreement is 94.1% with $\kappa = 0.81$, indicating high inter-rater reliability. The lowest reliability is observed for WDR ($\kappa = 0.69$) and CONC ($\kappa = 0.72$), reflecting genuine ambiguity in distinguishing withdrawal-related and attentional symptoms in community-level descriptions. TR_INT, TR_AVD, and SUI all reached perfect agreement, consistent with their relatively narrow operational definitions.

H.5 Treatment of control communities

The 11 non-mental-health control communities are assigned the all-zero vector by construction. This reflects the fact that the expert reference is defined in terms of targeted mental-health symptom dimensions. In other words, control communities are included not because they instantiate an alternative symptom structure, but because they provide contrastive categories outside the reference symptom vocabulary.

This design choice makes the expert reference intentionally asymmetric: it is a structured benchmark for comparing mental-health-related categories and their contrastive controls, rather than a universal taxonomy over all subreddit types.

H.6 Category-to-dimension assignments

Table 6 lists the binary category-to-dimension assignments used in this work.

H.7 Interpretive scope and limitations

The expert reference structure should be interpreted as a structured external benchmark for category comparison, not as a claim that subreddit communities are equivalent to formal diagnostic entities. Its role in this work is to provide a transparent and reproducible reference against which representation geometry can be evaluated. Accordingly, positive alignment with this reference should be understood as evidence of partial structural correspondence, rather than as direct recovery of an underlying clinical or diagnostic ontology.

Table 6: Expert-informed binary category profiles. ✓ indicates inclusion of a dimension in the category profile; × indicates absence. Control communities are assigned the all-zero vector by construction. The matrix is the consensus output of a two-annotator coding procedure (Appendix H.4).

Label	DEP	ANH	WOR	IRR	SLP	APP/EAT	FAT	CONC	WDR	HLTH	TR_INT	TR_AVD	HYPV	SUI
COVID19	✓	×	✓	✓	✓	×	✓	✓	✓	×	×	×	×	×
EDAnonymous	✓	✓	✓	×	✓	✓	✓	✓	✓	×	×	×	×	×
addiction	×	✓	✓	✓	✓	×	✓	✓	✓	×	×	×	×	✓
adhd	×	×	×	✓	✓	×	✓	✓	×	×	×	×	×	×
alcoholism	×	✓	✓	✓	✓	×	✓	✓	✓	×	×	×	×	✓
anxiety	×	×	✓	✓	✓	×	✓	✓	×	×	×	×	×	×
autism	×	×	×	×	×	×	×	×	✓	×	×	×	×	×
bipolarreddit	✓	✓	×	✓	✓	✓	✓	✓	×	×	×	×	×	✓
bpd	✓	×	✓	✓	✓	×	✓	×	✓	×	×	×	×	✓
conspiracy	×	×	×	×	×	×	×	×	×	×	×	×	×	×
depression	✓	✓	×	✓	✓	✓	✓	✓	✓	×	×	×	×	✓
divorce	×	×	×	×	×	×	×	×	×	×	×	×	×	×
fitness	×	×	×	×	×	×	×	×	×	×	×	×	×	×
guns	×	×	×	×	×	×	×	×	×	×	×	×	×	×
healthanxiety	×	×	✓	×	✓	×	✓	✓	×	✓	×	×	×	×
jokes	×	×	×	×	×	×	×	×	×	×	×	×	×	×
legaladvice	×	×	×	×	×	×	×	×	×	×	×	×	×	×
lonely	✓	×	✓	×	✓	×	✓	✓	✓	×	×	×	×	×
meditation	×	×	×	×	×	×	×	×	×	×	×	×	×	×
mentalhealth	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
parenting	×	×	×	×	×	×	×	×	×	×	×	×	×	×
personalfinance	×	×	×	×	×	×	×	×	×	×	×	×	×	×
ptsd	×	×	✓	✓	✓	×	✓	✓	✓	×	✓	✓	✓	×
relationships	×	×	×	×	×	×	×	×	×	×	×	×	×	×
schizophrenia	✓	✓	×	×	✓	×	✓	✓	✓	×	×	×	×	✓
socialanxiety	×	×	✓	✓	✓	×	×	×	✓	×	×	×	×	×
suicidewatch	✓	×	✓	✓	✓	×	✓	✓	✓	×	×	×	×	✓
teaching	×	×	×	×	×	×	×	×	×	×	×	×	×	×