PsyEmbedding: A Framework for Aligning Language Models to Psychological Theory

Anonymous ACL submission

Abstract

Pre-trained transformer-based language models revolutionized natural language processing and are increasingly important in computational psychology. However, these models' representations are optimized for quantifying *semantic* similarity, which may not always align with *psychological* similarity. We present *PsyEmbedding*, a framework for fine-tuning models (including BERT, SBERT, RoBERTa, GTE, and E5) to encode psychological constructs within content rather than just their semantic meaning. Leveraging a dataset annotated for numerous psychological constructs (CAMEL), we introduce a balanced stratified sampling strategy to generate embeddings predictive of psychological dimensions. We evaluate PsyEmbedding across multiple textual similarity and construct representation tasks, demonstrating that our method significantly aligns embedding spaces with psychological theory.

1 Introduction

The utility of Pre-trained Language Models (PLMs) is rapidly expanding beyond Natural Language Processing (NLP) into the social sciences (Ziems et al., 2024; Ash and Hansen, 2023), and especially to psychology (Feuerriegel et al., 2025; Hussain et al., 2024; Boyd and Schwartz, 2021). While PLMs capture rich semantic representations, they often yield unsatisfactory outcomes when *semantic* similarity diverges from *psychological* similarity. For example, the two sentences "I have a natural talent for influencing people" and "Everybody likes to hear my stories" are semantically distinct but mark the same underlying psychological construct: narcissism (Raskin and Terry, 1988). In a purely semantic space, these sentences are distant neighbors, limiting the ability of off-the-shelf models to capture psychologically meaningful relationships between texts.

To address this gap, we introduce *PsyEmbedding*, a framework that fine-tunes PLMs to effectively encode psychological representations. We utilize the Culture and Moral Expressions in Language (CAMEL) corpus (Zewail et al., 2025), a large-scale, psychologically annotated corpus, to optimize PLM representations such that psychologically similar sentences—regardless of lexical overlap—are proximal in embedding space. The CAMEL corpus is appropriate for three reasons: (1) it has been carefully annotated by experts for psychological labels, (2) it has a broad array of psychological constructs, not limited to one or a few labels, and (3) it is large enough to be suitable for meaningful fine-tuning.

By fine-tuning PLMs on this expert-annotated dataset, *PsyEmbedding* equips models to represent not only what sentences mean, but what they express about a person's latent beliefs, values, and motivations—bringing machine understanding closer to the patterns that organize human minds. For psychologists and computational social scientists, *PsyEmbedding* offers a tool that goes beyond general linguistic similarity to enable more theoretically-informed analyses of large-scale corpora. For NLP researchers, this framework demonstrates a generalizable approach for domain adaptation, showing how explicitly modeling a non-linguistic feature space (psychology) can improve model representations for domain-specific tasks.

2 Related Works

Psychological text analysis has historically relied on dictionary-based methods like LIWC (Pennebaker et al., 2003; Pennebaker and Graybeal, 2001), which categorizes words into predefined psychologically relevant categories such as emotions, cognitive processes, and social concerns (Tausczik and Pennebaker, 2010; Boyd et al., 2022). While these methods are simple, interpretable (Atari and Henrich, 2023; Feuerriegel et al., 2025), and theoretically grounded in psychological constructs (Boyd and Schwartz, 2021), they are funda-

mentally limited by their reliance on fixed, context-invariant lexical inventories. The static word embeddings, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), introduced a conceptual shift, representing words as points in a continuous semantic space where proximity reflects similarity in usage (Firth, 1957). Nevertheless, they still fail to capture syntax, polysemy, and word order, which are essential for encoding subtle psychological expressions in natural language. Transformer-based contextual language models such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2019) capture dependencies across entire sequences, allowing them to encode subtle semantic, syntactic, and pragmatic cues. In psychological text analysis, BERT has been employed to enhance the analysis of mental health, emotions, social biases, and personality (Cutler and Condon, 2022; Mozafari et al., 2020).

To enable efficient sentence comparison, Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) and subsequent models like RoBERTa (Liu et al., 2019), GTE (Li et al., 2023), and E5 (Wang et al., 2022) utilize contrastive learning for semantic search. This advancement made large-scale semantic similarity, clustering, and retrieval tasks feasible. However, these models remain constrained by the **Psychological Alignment Gap**. We define this gap as the discrepancy where purely semantic models fail to group text segments that are lexically and topically distinct yet express the same latent psychological construct.

Frameworks such as Contextualized Construct Representation (CCR) (Atari et al., 2023b) leverage semantic embeddings like SBERT to measure psychological constructs, and are effective for researchers who prioritize interpretability and hypothesis testing (Simchon et al., 2023; Chen et al., 2024; Abdurahman et al., 2024; Goyal et al., 2025). However, because SBERT and similar PLMs are optimized for capturing *semantic* similarity, they may not fully align with the degree to which two texts express the same *psychological* construct. This limitation extends to CCR and similar frameworks that rely on off-the-shelf PLMs (Sen et al., 2022; Lahnala et al., 2025). Semantically dissimilar sentences can indicate the same psychological phenomenon. PsyEmbedding bridges this gap by explicitly fine-tuning the embedding geometry to reflect psychological rather than semantic structure.

3 Method: PsyEmbedding

3.1 Problem Formulation

We propose a generalizable framework for adapting text embeddings to psychological domains. Our approach assumes a dataset $\mathcal{T} = \{(T_i, \mathbf{y}_i)\}$ where each text (T_i) is annotated for the presence (1) or absence (0) of N psychological constructs. We represent these annotations as a fixed-order binary vector $\mathbf{y}_i \in \{0,1\}^N$ (in the CAMEL corpus, $N = 25$), which serves as the "psychological profile".

3.2 Psychological Similarity Metric

To obtain a continuous supervision signal, we define the *ground-truth psychological similarity* between any two texts T_i and T_j as the cosine similarity of their psychological profiles: $s^\star_{ij} = \cos(\mathbf{y}_i, \mathbf{y}_j)$. Pairs where both $\mathbf{y}_i$ and $\mathbf{y}_j$ are all-zero are excluded. This provides a label-derived metric space in which proximity is defined by shared psychological constructs, independent of semantic or lexical overlap.

3.3 Balanced Stratified Pair Sampling

Standard random sampling of pairs from sparse datasets yields a distribution heavily skewed toward low-similarity ("easy negative") pairs. To learn fine-grained psychological distinctions, the model requires exposure to rare mid- and high-similarity pairs. We introduce a stratified sampling strategy. We divide the similarity interval $[0, 1]$ into B bins $\{[l_b, u_b)\}_{b=1}^{B}$, including a degenerate $[0, 0]$ bin for completely disjoint pairs and several progressively higher-similarity bins.Each bin is assigned a target number of pairs T_b according to

$$T_b = \left(\frac{|\mathcal{T}|}{2B}\right) \times m,$$

where m is a constraint on how many times a single text may appear across different pairs. The term $\frac{|\mathcal{T}|}{2B}$ corresponds to the expected per-bin count in an ideal setting with uniformly distributed similarities and no text reuse. The multiplicative factor m acts as a scaling parameter to permit controlled reuse of texts, ensuring adequate sample sizes in each bin while mitigating over-representation that could bias the model or induce overfitting.

Sampling proceeds iteratively: for a bin with the largest gap between its current and target counts, an anchor text T_i that has not reached its m limit is randomly drawn. A candidate partner T_j is selected such that their $s^\star_{ij}$ falls within the bin's predefined

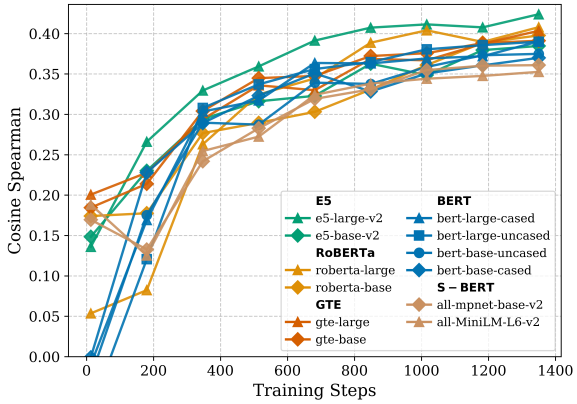

Figure 1: Validation-set learning curves for all fine-tuned models.

range $[l_b, u_b)$. This produces a training set $\mathcal{P}_b$ with uniform coverage across the full similarity range.

3.4 Training Objective

During training, for each pair (T_i, T_j), the model generates embeddings $\mathbf{z}_i$ and $\mathbf{z}_j$, and the goal is to minimize the Mean Squared Error (MSE) between the cosine similarity of the generated text embeddings $\mathbf{z}_i^\top \mathbf{z}_j$ and the ground-truth psychological similarity $s_{ij}^\star$:

$$\mathcal{L} = \frac{1}{|\mathcal{P}_{\text{train}}|} \sum_{(i,j)\in\mathcal{P}_{\text{train}}} \left(\mathbf{z}_i^\top \mathbf{z}_j - s_{ij}^\star\right)^2.$$

4 Experimental Setup

4.1 Dataset

We instantiate the PsyEmbedding framework using the Cultural and Moral Expressions in Language (CAMEL) corpus (Zewail et al., 2025), a large-scale, theory-driven dataset of over 56,787 texts annotated for 25 different psychological variables. The labels include moral values (e.g., equality, liberty), adverse language (e.g., hate-based rhetoric, incivility), cognitive styles (e.g., analytic thinking, creativity), and cultural orientations (e.g., collectivism, norm tightness). The corpus was curated from multiple online platforms (e.g., social media), archives (e.g., historical quotes), and synthetic language (generated by Large Language Models). Each text is labeled by three annotators: we encoded labels as 1 (majority agreement), -1 (noise/disagreement), and 0 (absence). After preprocessing, we partitioned the data into training (39,750), validation (8,518), and test (8,519) splits. We applied the balanced sampling strategy (Section 3.3) to all splits. When $m = 1$, this yielded 11,180 training pairs; setting $m = 2$ yielded 26,289 pairs.

4.2 Evaluation Tasks

We evaluate performance on two complementary tasks:

4.2.1 Psychological Textual Similarity (PTS)

This task measures how well the embeddings preserve continuous psychological similarity following the standard evaluation paradigm of Semantic Textual Similarity (Cer et al., 2017). Across all text pairs, we compute the Spearman and Pearson correlations between the model's predicted cosine similarity and the ground-truth scores $s_{ij}^\star$ computed based on expert annotations.

4.2.2 Contextualized Construct Representation (CCR)

While PTS captures global semantic alignment between texts, CCR is designed to test whether embeddings reflect the presence of specific psychological constructs in context (Chen et al., 2024). Each evaluation instance consists of a text T_i, a set of items $\mathcal{S}_c = \{s_1, s_2, \ldots, s_{|\mathcal{S}_c|}\}$ corresponding to a psychological construct c, and a binary label $y_i \in \{0, 1\}$ indicating whether the construct is expressed. For each text T_i, we compute the cosine similarity with all items in $\mathcal{S}_c$ and take the average as the model's prediction:

$$\hat{y}_i = \frac{1}{|\mathcal{S}_c|} \sum_{s\in\mathcal{S}_c} \cos\big(f(T_i), f(s)\big),$$

where $f(\cdot)$ denotes the sentence encoder, and $\cos(\cdot,\cdot)$ is the cosine similarity function. This procedure follows the original CCR pipeline (Atari et al., 2023b; Chen et al., 2024).

During evaluation, we interpret $\hat{y}_i$ as the predicted probability of construct presence: threshold-based metrics (accuracy, F1, precision, recall) are obtained by selecting an optimal decision threshold, while ranking-based metrics (e.g., average precision) directly exploit the continuous scores to assess whether positive examples are consistently ranked above negatives. From the test split, we randomly sampled 10650 texts, maintaining an equal 1:1 ratio between positive and negative instances to ensure balanced representation. Since the original CCR pipeline works with psychological questionnaire items (Atari et al., 2023b), we evaluated CCR using 17 psychological constructs for which validated scales were available (see Table 5 in the

Model	PTS		CCR		
	Pearson	Spearman	Accuracy	F1	AP
bert-base-cased	0.025	-0.007	0.650	0.697	0.639
after PsyEmbed	0.334	0.355	0.660	0.716	0.716
bert-base-uncased	0.011	-0.025	0.653	0.689	0.702
after PsyEmbed	0.345	0.364	0.663	0.692	0.732
bert-large-cased	0.013	-0.026	0.643	0.682	0.696
after PsyEmbed	0.354	0.368	0.654	0.698	0.729
bert-large-uncased	-0.014	-0.050	0.646	0.686	0.683
after PsyEmbed	0.351	0.371	0.661	0.704	0.725
roberta-base	0.144	0.156	0.611	0.693	0.626
after PsyEmbed	0.367	0.383	0.638	0.685	0.704
roberta-large	0.070	0.072	0.620	0.674	0.640
after PsyEmbed	0.371	0.388	0.637	0.695	0.709
all-mpnet-base-v2	0.198	0.204	0.578	0.667	0.600
after PsyEmbed	0.337	0.354	0.637	0.691	0.689
all-MiniLM-L6-v2	0.195	0.186	0.577	0.667	0.608
after PsyEmbed	0.317	0.336	0.619	0.685	0.640
e5-base-v2	0.155	0.148	0.555	0.668	0.579
after PsyEmbed	0.348	0.369	0.654	0.706	0.733
e5-large-v2	0.164	0.154	0.530	0.670	0.535
after PsyEmbed	0.369	0.384	0.643	0.686	0.685
gte-base	0.217	0.209	0.588	0.671	0.629
after PsyEmbed	0.350	0.369	0.642	0.697	0.681
gte-large	0.223	0.223	0.589	0.672	0.623
after PsyEmbed	0.357	0.385	0.646	0.698	0.706

Table 1: Performance of pre-trained models without fine-tuning and after PsyEmbed across Psychological Textual Similarity (PTS) and Contextualized Construct Representation (CCR) evaluation tasks. AP stands for Average Precision.

Appendix). The constructs encompass three primary domains: cultural orientations (e.g., collectivism);moral values (e.g., authority); and cognitive processes (e.g., analytical thinking).

5 Results

We fine-tuned a suite of models (BERT, RoBERTa, SBERT, GTE, E5). During fine-tuning, validation was performed every 10 steps, and the best configuration for each model was selected via a hyperparameter sweep as shown in Table 2 in the Appendix. Figure 1 illustrates that fine-tuning with psychologically informed similarity supervision yields substantial improvements over pre-trained baselines across all architectures. The consistent gains indicate that our method successfully aligns the embedding space with the targeted psychological constructs, enabling the models to capture psychological dimensions in text effectively. Table 1 summarizes the performance before and after fine-tuning.

Performance Gains. Fine-tuning yielded substantial improvements across all architectures. On the PTS task, pre-trained GTE models initially outperformed BERT, but after fine-tuning, all models showed significant alignment with psychological similarity. Notably, fine-tuning also improved performance on the CCR task, despite CCR not being the direct training objective. This cross-task transfer suggests the models successfully learned generalized psychological representations rather than overfitting to specific pair similarities.

Impact of Data Reuse. We analyzed the effect of the m parameter. Contrary to standard expectations where more data yields better models, setting $m = 2$ (permitting texts to be reused) consistently decreased performance compared to to $m = 1$. This suggests that in psychological domains, reusing texts induces overfitting to specific lexical artifacts (memorizing the anchor text) rather than learning robust, abstract construct regularities.

6 Discussion and Conclusion

Language is not merely a vehicle for conveying propositions; it is a mirror of the human mind (Pennebaker et al., 2003; Pinker, 2007). Understanding language, therefore, requires modeling the psychological dimensions that underlie expression, the beliefs, norms, and emotions between the lines. Yet standard NLP models often prioritize surface meaning over latent psychological content. By shifting from purely *semantic* to *psychological* similarity, PsyEmbedding reframes how we extract cognitive traces from text. Methodologically, this opens a pathway toward psychologically grounded NLP, where representations of text are aligned with constructs such as cultural orientations and moral values.Theoretically, this work underscores how the same linguistic form can encode different mental states, and how distinct-looking expressions can be about the same psychological profile.

Crucially, PsyEmbedding is a model-agnostic framework, not merely a collection of fine-tuned models. It provides a generalizable recipe for aligning any fine-tuning PLMs with psychological theory. Therefore, as larger annotated corpora become available in psychology and adjacent social-science disciplines, this approach will enable sharper detection of complex behavioral and emotional signals, bridging the gap between linguistic meaning and the cognitive processes that shape it.

Limitations

PsyEmbedding has several limitations worth noting. First, our work is confined to English, a language whose cultural and grammatical structures may not generalize to other linguistic contexts. Developing suites of PsyEmbedding in other languages may encounter different issues (Blasi et al., 2022). Cognitive processes are shaped by culture and language (Nisbett et al., 2001; Blasi et al., 2022), and future research should extend this framework to multilingual corpora to test whether the same latent psychological dimensions emerge across linguistic and cultural boundaries. Second, our model targets a large, but still limited set of psychological constructs, leaving out aspects of human psychology. Expanding and validating the framework across a broader spectrum of psychological domains will be helpful for developing a truly comprehensive model. Finally, the fine-tuned models perform differently across various domains, as shown in Table 4 in the Appendix. Their performance in social and cultural values and cognitive and intellectual dispositions is better than in morality, which may be related to the quality and subjectivity of the manual annotations in the training data.

References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245.

Elliott Ash and Stephen Hansen. 2023. Text algorithms in economics. *Annual Review of Economics*, 15:659–688.

Mohammad Atari and Jonathan Haidt. 2023. Ownership is (likely to be) a moral foundation. *Behavioral & Brain Sciences*, 46.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023a. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157.

Mohammad Atari and Joseph Henrich. 2023. Historical psychology. *Current Directions in Psychological Science*, 32(2):176–183.

Mohammad Atari, Ivan Kroupin, Helen E. Davis, Yuqi Chen, Jonathan Schulz, and Joseph Henrich. 2025. The origins of parochial morality: Intensive kinship and qeirat values. Manuscript under review.

Mohammad Atari, Ali Omrani, and Morteza Dehghani. 2023b. Contextualized construct representation: Leveraging psychometric scales to advance theory-driven text analysis.

Damián E Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on english hinders cognitive science. *Trends in cognitive sciences*, 26(12):1153–1170.

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*, pages 1–47.

Ryan L Boyd and H Andrew Schwartz. 2021. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Yuqi Chen, Sixuan Li, Ying Li, and Mohammad Atari. 2024. Surveying the dead minds: Historical-psychological text analysis with contextualized construct representation (CCR) for classical Chinese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2597–2615, Miami, Florida, USA. Association for Computational Linguistics.

Incheol Choi, Minkyung Koo, and Jong An Choi. 2007. Individual differences in analytic versus holistic thinking. *Personality and social psychology bulletin*, 33(5):691–705.

Andrew Cutler and David M Condon. 2022. Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Stefan Feuerriegel, Abdurahman Maarouf, Dominik Bär, Dominique Geissler, Jonas Schweisthal, Nicolas Pröllochs, Claire E Robertson, Steve Rathje, Jochen Hartmann, Saif M Mohammad, et al. 2025. Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 4(2):96–111.

J. R. Firth. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Philological Society, Oxford. Reprinted in F. R. Palmer (ed.), *Selected Papers of J. R. Firth 1952–1959*, London: Longman, 1968.

Michele J Gelfand, Jana L Raver, Lisa Nishii, Lisa M Leslie, Janetta Lun, Beng Chong Lim, Lili Duan, Assaf Almaliach, Soon Ang, Jakobina Arnadottir, et al. 2011. Differences between tight and loose cultures: A 33-nation study. *science*, 332(6033):1100–1104.

Namrata Goyal, Lorenzo De Gregori, Yuqi Liu, and Krishna Savani. 2025. Moral absolutism drives support for bans: Unpacking ideological differences in the moral philosophies of conservatives and liberals. *Journal of Personality and Social Psychology*.

Zak Hussain, Marcel Binz, Rui Mata, and Dirk U Wulff. 2024. A tutorial on open-source large language models for behavioral science. *Behavior Research Methods*, 56(8):8214–8237.

Ravi Iyer, Spassena Koleva, Jesse Graham, Peter Ditto, and Jonathan Haidt. 2012. Understanding libertarian morality: The psychological dispositions of self-identified libertarians.

James C Kaufman. 2012. Counting the muses: development of the kaufman domains of creativity scale (k-docs). *Psychology of Aesthetics, Creativity, and the Arts*, 6(4):298.

Harold G Koenig and Arndt Büssing. 2010. The duke university religion index (durel): a five-item measure for use in epidemological studies. *Religions*, 1(1):78–85.

Elizabeth J Krumrei-Mancuso and Steven V Rouse. 2016. The development and validation of the comprehensive intellectual humility scale. *Journal of Personality Assessment*, 98(2):209–221.

Allison Lahnala, Vasudha Varadarajan, Lucie Flek, H Andrew Schwartz, and Ryan L Boyd. 2025. Unifying the extremes: Developing a unified model for detecting and predicting extremist traits and radicalization. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1051–1067.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Richard E Nisbett, Kaiping Peng, Incheol Choi, and Ara Norenzayan. 2001. Culture and systems of thought: holistic versus analytic cognition. *Psychological review*, 108(2):291.

Sheida Novin and Daphna Oyserman. 2016. Honor as cultural mindset: Activated honor mindset affects subsequent judgment and attention in mindset-congruent ways. *Frontiers in psychology*, 7:1921.

Daphna Oyserman. 1993. The lens of personhood: Viewing the self and others in a multicultural society. *Journal of personality and social psychology*, 65(5):993.

James W Pennebaker and Anna Graybeal. 2001. Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, 10(3):90–93.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Steven Pinker. 2007. *The stuff of thought: Language as a window into human nature*. Penguin.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Robert Raskin and Howard Terry. 1988. A principal-components analysis of the narcissistic personality inventory and further evidence of its construct validity. *Journal of personality and social psychology*, 54(5):890.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Indira Sen, Daniele Quercia, Marios Constantinides, Matteo Montecchi, Licia Capra, Sanja Scepanovic, and Renzo Bianchi. 2022. Depression at work: exploring depression in major us companies from online reviews. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–21.

Almog Simchon, Britt Hadar, and Michael Gilead. 2023. A computational text analysis investigation of the relation between personal and linguistic agency. *Communications Psychology*, 1(1):23.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Aliah Zewail, Aanchal Setia, Roya Mohammadsadegh, Fırat Şeker, Ali Hajian, Hector J. Sosa, Satvika Reddy Gavireddy, Liora Morhayim, and Mohammad Atari. 2025. The cultural and moral expressions in language (camel) corpus. *Manuscript under review*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Appendix

Batch Size	{32}
Max Occurrences	{1, 2}
Seed	{42}
Epochs	{3}
Warm-up Epoch	{1, 2}
Learning Rate	{1e-5, 2e-5, 3e-5}

Table 2: Hyperparameter sweep for the validation of fine-tuned models.

Model	Performance Metrics		Hyperparameters			
	Pearson	Spearman	Batch	Warmup	LR	Sampling Strategy
BERT						
bert-base-cased	0.345	0.371	32	1	3e-5	max_occurrences=1
bert-base-uncased	0.358	0.378	32	1	3e-5	max_occurrences=1
bert-large-cased	0.369	0.390	32	1	3e-5	max_occurrences=1
bert-large-uncased	0.369	0.390	32	1	3e-5	max_occurrences=1
RoBERTa						
roberta-base	0.383	0.398	32	2	3e-5	max_occurrences=1
roberta-large	0.395	0.410	32	1	3e-5	max_occurrences=1
S-BERT						
all-mpnet-base-v2	0.346	0.363	32	1	3e-5	max_occurrences=1
all-MiniLM-L6-v2	0.337	0.357	32	1	3e-5	max_occurrences=1
E5						
e5-base-v2	0.370	0.387	32	2	3e-5	max_occurrences=1
e5-large-v2	**0.411**	**0.426**	32	1	3e-5	max_occurrences=1
GTE						
gte-base	0.372	0.391	32	1	3e-5	max_occurrences=1
gte-large	0.388	0.405	32	2	3e-5	max_occurrences=1

Table 3: Performance of fine-tuned models on the validation set with the optimal configurations.

Model		Cultural Orientations					Moral Values					Cognitive Processes				
		Acc.	F1	Prec.	Rec.	AP	Acc.	F1	Prec.	Rec.	AP	Acc.	F1	Prec.	Rec.	AP
BERT																
bert-base-cased	w/o FT	0.608	0.669	0.503	0.998	0.627	0.578	0.677	0.527	0.947	0.588	0.780	0.777	0.754	0.802	0.824
	w/ FT †	0.690	0.722	0.629	0.849	0.774	0.565	0.673	0.512	0.980	0.590	0.780	0.789	0.748	0.835	0.852
bert-base-uncased	w/o FT	0.632	0.675	0.528	0.936	0.684	0.563	0.673	0.515	0.970	0.595	0.781	0.779	0.788	0.770	0.830
	w/ FT †	0.704	0.735	0.627	0.889	0.785	0.548	0.667	0.500	1.000	0.560	0.784	0.787	0.766	0.808	0.859
bert-large-cased	w/o FT	0.627	0.680	0.542	0.914	0.679	0.558	0.676	0.520	0.964	0.568	0.790	0.784	0.803	0.766	0.837
	w/ FT †	0.708	0.737	0.633	0.881	0.800	0.541	0.668	0.504	0.990	0.561	0.764	0.768	0.739	0.799	0.835
bert-large-uncased	w/o FT	0.638	0.679	0.530	0.946	0.680	0.562	0.671	0.515	0.959	0.582	0.764	0.767	0.729	0.809	0.813
	w/ FT †	0.705	0.732	0.648	0.841	0.793	0.550	0.668	0.503	0.996	0.578	0.775	0.783	0.749	0.820	0.843
RoBERTa																
roberta-base	w/o FT	0.570	0.668	0.501	0.998	0.597	0.554	0.668	0.502	0.999	0.575	0.782	0.770	0.771	0.770	0.840
	w/ FT †	0.667	0.714	0.592	0.899	0.756	0.524	0.667	0.500	1.000	0.532	0.779	0.779	0.725	0.841	0.858
roberta-large	w/o FT	0.594	0.667	0.501	0.996	0.621	0.550	0.668	0.503	0.994	0.572	0.734	0.742	0.696	0.796	0.798
	w/ FT †	0.707	0.719	0.594	0.910	0.792	0.565	0.675	0.524	0.949	0.586	0.702	0.713	0.612	0.853	0.772
S-BERT																
all-mpnet-base-v2	w/o FT	0.664	0.704	0.581	0.894	0.725	0.558	0.667	0.502	0.994	0.566	0.553	0.667	0.500	0.999	0.557
	w/ FT †	0.705	0.725	0.603	0.908	0.795	0.537	0.668	0.502	0.996	0.547	0.738	0.745	0.695	0.802	0.783
all-MiniLM-L6-v2	w/o FT	0.650	0.682	0.546	0.908	0.710	0.552	0.667	0.500	1.000	0.565	0.560	0.668	0.504	0.991	0.583
	w/ FT †	0.621	0.699	0.559	0.932	0.663	0.544	0.667	0.501	0.997	0.556	0.736	0.726	0.720	0.732	0.805
E5																
e5-base-v2	w/o FT	0.614	0.672	0.514	0.969	0.646	0.561	0.670	0.507	0.987	0.580	0.517	0.667	0.502	0.995	0.521
	w/ FT †	0.709	0.733	0.645	0.849	0.792	0.527	0.667	0.503	0.992	0.540	0.794	0.786	0.799	0.774	0.875
e5-large-v2	w/o FT	0.581	0.670	0.514	0.965	0.603	0.541	0.669	0.510	0.976	0.551	0.520	0.672	0.509	0.990	0.486
	w/ FT †	0.735	0.742	0.668	0.835	0.809	0.538	0.667	0.500	1.000	0.551	0.743	0.761	0.682	0.860	0.776
GTE																
gte-base	w/o FT	0.669	0.707	0.575	0.919	0.737	0.576	0.667	0.501	0.998	0.609	0.563	0.675	0.513	0.986	0.565
	w/ FT †	0.671	0.718	0.605	0.884	0.741	0.540	0.667	0.500	1.000	0.553	0.772	0.771	0.724	0.824	0.846
gte-large	w/o FT	0.654	0.686	0.585	0.828	0.709	0.580	0.667	0.500	1.000	0.600	0.584	0.681	0.530	0.956	0.587
	w/ FT †	0.731	0.730	0.661	0.814	0.820	0.538	0.668	0.502	0.996	0.565	0.752	0.762	0.712	0.820	0.820

Table 4: Domain-wise evaluation of pre-trained (w/o FT) versus fine-tuned (w/ FT †) models on the CCR task. Acc., Prec., Rec., and AP stand for Accuracy, Precision, Recall, and Average Precision, respectively.

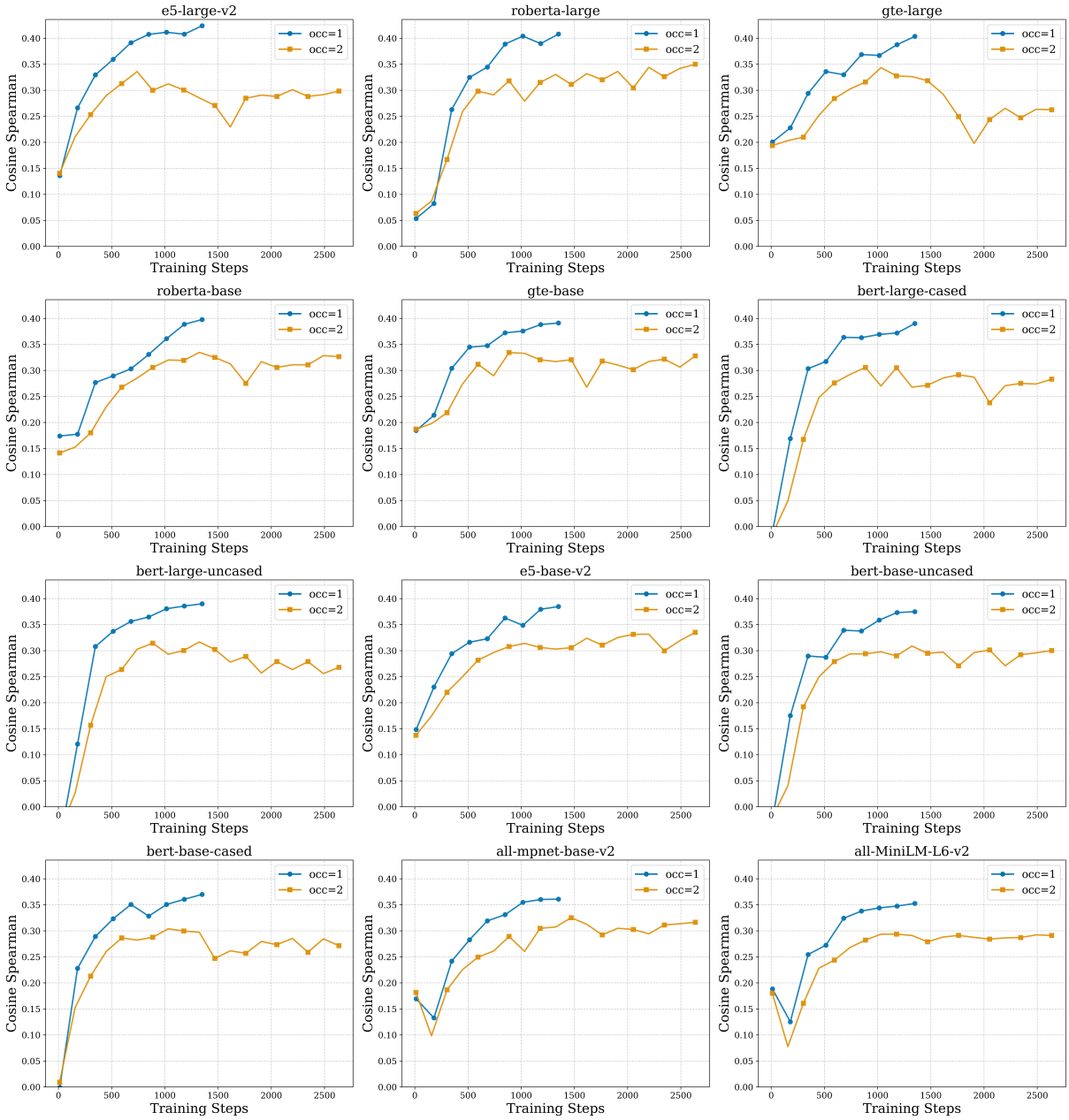

Figure 2: Validation-set learning curves across different max_occurrences settings.

Domain	Construct	Item Example	Source
Cultural Orientations	Collectivism	In general, I accept the decisions made by my group.	Oyserman (1993)
	Individualism	I determine my own destiny.	Oyserman (1993)
	Kinship	I prefer to have close connections with my blood family.	Atari et al. (2025)
	Religion	In my life, I experience the presence of the Divine (i.e., God).	Koenig and Büssing (2010)
	Honor	I prefer to live with honor, even if it means I will earn less money.	Novin and Oyserman (2016)
	Tightness	People in this country almost always comply with social norms.	Gelfand et al. (2011)
Moral Values	Authority	I think it is important for societies to cherish their traditional values.	Atari et al. (2023a)
	Loyalty	It upsets me when people have no loyalty to their country.	Atari et al. (2023a)
	Care	Caring for people who have suffered is an important virtue.	Atari et al. (2023a)
	Proportionality	It makes me happy when people are recognized on their merits.	Atari et al. (2023a)
	Equality	Our society would have fewer problems if people had the same income.	Atari et al. (2023a)
	Purity	I believe chastity is an important virtue.	Atari et al. (2023a)
	Ownership	In an ideal society, everyone would respect everyone else's rights of ownership.	Atari and Haidt (2023)
	Liberty	The government interferes far too much in our everyday lives.	Iyer et al. (2012)
Cognitive Processes	Analytical Thinking	Everything in the world is intertwined in a causal relationship.	Choi et al. (2007)
	Creativity and Innovation	Coming up with a new way to think about an old debate.	Kaufman (2012)
	Intellectual Humility	I can respect others, even if I disagree with them in important ways.	Krumrei-Mancuso and Rouse (2016)

Table 5: Example Items from Psychological Scales Across Different Domains.