# On the Implicit Encoding of Human Psychology in Large Language Model Representations

Jonathan Politzki[1]

[1]Jean Technologies, Inc.

February 25, 2026

**Abstract**

We present a logical argument, grounded in psycholinguistics, representation learning theory, and mechanistic interpretability, that large language models trained on text corpora have implicitly learned structured representations of human psychological traits. The argument proceeds in three steps. First, we establish that natural language is a reliable projection of the human mind: personality traits, values, emotional states, and cognitive styles systematically manifest in language patterns. Second, we show that modern language models learn internal representations that go beyond surface-level co-occurrence statistics, developing structured world models that include space, time, truth, and social cognition. Third, we connect these findings through the lens of representation learning: the features that emerge in a model's latent space are causally determined by the training objective $P(Y \mid X)$, and since language modeling requires predicting human-generated text, the resulting representations necessarily encode the psychological processes that generated that text. We discuss the implications and limitations of this finding — in particular, that task-specific training objectives determine which psychological features become salient, that semantic similarity in embedding space does not equate to deeper forms of understanding such as compatibility, and that current embedding architectures face fundamental constraints including dimensional collapse and the conflation of similarity with relevance. We argue that this implicit psychological encoding is a foundational capability with broad downstream applications, from personalization to human modeling, but that realizing these applications requires moving beyond naive similarity-based retrieval toward task-aware representation systems.

## 1 Introduction

The last decade of machine learning has produced a development that is simultaneously celebrated and profoundly misunderstood. Large language models — trained on vast corpora of human-generated text to predict the next token — have achieved remarkable performance on tasks ranging from translation to reasoning to creative writing. The popular understanding emphasizes the generative output: these models produce text that reads as if written by a knowledgeable human.

We argue that the generative capability, while impressive, obscures a more fundamental development: *what these models have learned about us in the process of learning to predict our words.*

The argument is simple in structure. Language is not arbitrary noise. Decades of psycholinguistic research establish that language is a systematic projection of the speaker's mind — their personality traits, values, emotional states, cognitive style, and social orientation leave reliable traces in the words they choose, the syntax they deploy, and the topics they engage

[Pennebaker, 2011, Tausczik and Pennebaker, 2010]. If a model is trained to predict human language with high fidelity, it must — as a necessary condition of that prediction task — develop internal representations of the psychological processes that generate language. A model that did not implicitly represent human extraversion, agreeableness, or emotional regulation could not accurately predict the linguistic patterns that these traits produce.

This is not a speculative hypothesis. The claim moved from theoretical to empirically confirmed with the advent of mechanistic interpretability. Bricken et al. [2023] decomposed language model activations into interpretable monosemantic features and found that the majority mapped to single human-interpretable concepts. Templeton et al. [2024] scaled this approach to production models, extracting tens of millions of features from Claude 3 Sonnet — and among them, features encoding unmistakably human psychological constructs: sycophancy, deception, bias, power-seeking, emotional manipulation. These are not abstract statistical patterns. They are interpretable features that correspond to specific human behavioral traits, existing as structured directions in the model's activation space. The discovery of these features is, to us, the most direct evidence that language models have internalized a structured model of human psychology — not as a design goal, but as a necessary byproduct of learning to predict human-generated text. Complementary probing studies have confirmed that personality traits, emotional states, and social dynamics are encoded in linearly extractable form [Gurnee and Tegmark, 2024, Kosinski, 2024].

The purpose of this paper is to formalize this argument, ground it in the relevant literatures, and examine its implications and limitations. We do not propose a specific system or architecture. Instead, we establish a foundational claim: **the latent representations of large language models contain structured encodings of human psychological traits, and this implicit knowledge is a general-purpose resource with broad downstream applications.**

We also identify critical limitations that must be addressed before this resource can be reliably exploited. Chief among these: the features that become salient in any learned representation are causally determined by the training objective. A model trained for next-token prediction learns a different organization of psychological features than one trained for compatibility matching or personality classification. Semantic similarity in embedding space — the dominant retrieval paradigm — does not capture deeper relational structures such as compatibility, complementarity, or causal influence. And current embedding architectures face well-known pathologies including dimensional collapse [Jing et al., 2022] and the curse of dimensionality that limit their capacity to faithfully represent the full complexity of a human being.

## 1.1 Contributions

1. We formalize the logical chain from psycholinguistics (text reflects the mind) through representation learning theory ($P(Y \mid X)$ determines feature salience) to the conclusion that language models implicitly encode human psychology.

2. We survey the empirical evidence across three independent research programs — psycholinguistics, emergent capabilities, and mechanistic interpretability — showing convergent support for this claim.

3. We articulate the critical distinction between semantic similarity and task-relevant representation, arguing that downstream applications of personal embeddings require task-aware training objectives, not naive similarity search.

4. We identify open problems and limitations, establishing a research agenda for the emerging field of computational human representation.

# 2 Background: Representation Learning

Before presenting our argument, we establish the theoretical framework that connects training objectives to learned features.

## 2.1 Representations Are Shaped by Objectives

Bengio et al. [2013] defined the central goal of representation learning as discovering transformations of raw data that make subsequent learning tasks easier. A key insight from this framework is that the quality of a learned representation is always relative to a downstream task — there is no task-independent notion of a "good" representation.

More precisely, any learned representation is shaped by the objective function used to train it. Given data $X$ and an objective involving target $Y$, the model learns features of $X$ that are maximally informative about $Y$. This is not a limitation but a fundamental property: **the training objective causally determines which features of the input become salient in the learned representation.**

$$\text{Representation} = f_\theta(X) \quad \text{where} \quad \theta^* = \arg\min_\theta \; \mathcal{L}(f_\theta(X), Y) \tag{1}$$

When $Y$ is the next token in a text sequence, the model must learn features of $X$ (the preceding context) that predict human linguistic choices. When $Y$ is a compatibility label between two people, the model must learn features that predict relational outcomes. When $Y$ is a document relevance judgment, the model must learn features that predict topical alignment.

Each of these objectives produces a different organization of the latent space — a different "projection" of the same underlying data. This observation is critical for understanding both the power and the limitations of using language model representations for human modeling.

## 2.2 The Bitter Lesson and General Representations

Sutton [2019] observes that in the long run, general methods that leverage computation consistently outperform hand-engineered, domain-specific approaches. In representation learning, this manifests as a consistent finding: large models trained on diverse data learn more transferable features than small models trained on narrow domains [Goodfellow et al., 2016].

However, the No Free Lunch Theorem [Wolpert and Macready, 1997] provides an important caveat: averaged over *all possible* tasks, no single representation outperforms any other. In practice, this means that while general representations are broadly useful, task-specific fine-tuning consistently improves performance on any given downstream application. This tension — between generality and specialization — is central to the challenge of building useful personal embeddings, as we discuss in Section 5.

## 2.3 Dense Embedding Spaces

Modern embedding models map inputs to dense vectors in $\mathbb{R}^D$ where geometric proximity (typically cosine similarity) encodes semantic relatedness. The embedding paradigm, from Word2Vec [Mikolov et al., 2013] to contemporary sentence transformers [Reimers and Gurevych, 2019], has become the dominant approach for representing text in applications ranging from search to recommendation to matching.

The power of this paradigm lies in its compositionality: the famous $\text{king} - \text{man} + \text{woman} \approx \text{queen}$ example [Mikolov et al., 2013] demonstrates that linear structure in the embedding space corresponds to meaningful semantic relationships. This linear structure is not limited to simple analogies — it extends to personality traits, emotional valence, and other psychological dimensions, as we will show.

The limitation of this paradigm is equally important: **the geometry of the embedding space is entirely determined by the training objective.** A model trained for semantic similarity learns to place semantically similar texts nearby, regardless of whether that similarity is useful for the downstream task. Two job candidates may be semantically similar (similar backgrounds, similar language) yet entirely incompatible with the same role. This conflation of similarity with relevance is a fundamental constraint that we examine in Section 5.1.

# 3 The Argument: Text Is a Projection of the Mind

We now present the core argument in three steps: (1) language reliably reflects psychological traits, (2) language models learn structured representations that go beyond surface statistics, and (3) therefore, language model representations implicitly encode human psychology.

## 3.1 Step 1: Language as Psychological Fingerprint

The claim that language reflects the speaker's mind is not speculative — it is one of the most replicated findings in psycholinguistics.

**Foundational evidence.** Pennebaker [2011] demonstrated that even the most forgettable aspects of language — function words like pronouns, prepositions, and articles — serve as reliable fingerprints of personality, emotional states, social relationships, and cognitive style. These patterns are not consciously controlled; they emerge from the underlying psychological processes that generate speech and writing.

Tausczik and Pennebaker [2010] systematized this insight through the Linguistic Inquiry and Word Count (LIWC) framework, showing that word category frequencies correlate significantly with Big Five personality traits, emotional regulation, attentional focus, and social orientation. Hundreds of subsequent studies have validated and extended these findings across languages, platforms, and populations.

**Open-vocabulary confirmation.** Schwartz et al. [2013] moved beyond closed dictionaries to an open-vocabulary analysis of 700 million words from 75,000 Facebook users, revealing that personality traits produce systematic, pervasive variations across all aspects of language use — not just word choice, but topic selection, phrasing patterns, and syntactic complexity. The signal is not confined to a few diagnostic words; personality permeates the full distribution of linguistic behavior.

**Machines exceed human judgment.** Youyou et al. [2015] demonstrated that computational models trained on Facebook Likes achieved personality prediction accuracy ($r = 0.56$) exceeding that of friends and family members. With 300 Likes, the computer outperformed a person's spouse. While this study used behavioral data rather than text, it establishes the broader principle: computational analysis of human behavioral traces recovers personality structure more accurately than human intuitive judgment.

**Meta-analytic confirmation.** Moreno et al. [2021] conducted a meta-analysis of 23 independent studies on computational personality prediction from text, finding significant correlations ($r = 0.26$–$0.30$) between Big Five traits and computationally-derived language features. Performance improved when both semantic *and* syntactic features were used, suggesting that personality signal exists at multiple levels of linguistic representation.

**Summary.** The evidence is unambiguous: natural language is a reliable, multi-level projection of the speaker's psychological state. Any model that learns to predict or represent language with high fidelity must, as a necessary consequence, develop internal representations of the psychological processes that generate that language.

### 3.2 Step 2: Language Models Learn Structured World Models

The second step establishes that modern language models do not merely learn surface-level co-occurrence statistics. They develop internal representations with genuine structural properties — linear organization, causal relationships, and conceptual abstraction — that go qualitatively beyond pattern matching.

**Spatial and temporal representations.** Gurnee and Tegmark [2024] demonstrated that Llama-2 models learn linear representations of physical space and time. Individual neurons and directions in activation space encode geographic coordinates and historical dates in a structured, continuous manner. These are not memorized associations but genuine geometric representations of abstract physical concepts — learned purely from text.

**Emergent world models.** Li et al. [2023] showed that a GPT model trained only to predict legal Othello moves — with zero explicit knowledge of the game — spontaneously develops an internal representation of the board state. Causal intervention experiments confirmed that this representation is not epiphenomenal: it causally drives the model's outputs. The model learned a world model as a byproduct of sequence prediction.

**Latent knowledge.** Burns et al. [2023] demonstrated that language models develop internal representations that encode truth versus falsehood in a structured, logically consistent way. Using unsupervised probing, they extracted latent knowledge from model activations that outperformed zero-shot prompting — showing that models have richer internal knowledge than what they express through outputs.

**Theory of Mind.** Most directly relevant to our argument, Kosinski [2024] showed that GPT-4 solves 75% of false-belief tasks, matching the performance of 6-year-old children. Strachan et al. [2024] extended this finding with the most comprehensive Theory of Mind evaluation to date, showing GPT-4 performing at or above human levels on identifying indirect requests, false beliefs, and misdirection. The ability to model other minds — to represent what another person believes, wants, and intends — is a core psychological capability that has emerged without explicit training.

**Behavioral simulation.** Park et al. [2023] demonstrated that LLM-powered agents, given only brief personality descriptions, produce believable emergent social behaviors: forming relationships, coordinating events, and making plans consistent with their described traits. The models have internalized enough about human psychology and social dynamics to simulate coherent human behavior from text-based personality descriptions alone.

**Summary.** Language models trained on text develop structured internal representations of space, time, truth, game states, social cognition, and personality. These are not surface-level statistical artifacts; they are organized representations with geometric structure, causal efficacy, and behavioral consequence. The question is not whether these models represent the world — it is *how much* of the world, and how faithfully.

### 3.3 Step 3: Therefore, LLM Representations Encode Human Psychology

The conclusion follows directly from the premises:

1. Language is a systematic projection of the speaker's psychological state (Step 1).
2. Language models learn structured, causally efficacious internal representations of the patterns in their training data (Step 2).
3. Language models are trained on human-generated text (premise).
4. Therefore, language model representations necessarily encode the psychological processes that generate human language.

This is not merely a theoretical deduction. It has been directly confirmed by mechanistic interpretability research. Bricken et al. [2023] used sparse autoencoders to decompose model activations into interpretable features, finding that 70% of extracted features mapped cleanly to single human-interpretable concepts. Templeton et al. [2024] scaled this approach to Claude 3 Sonnet, extracting tens of millions of features ranging from concrete concepts ("Golden Gate Bridge") to highly abstract ones ("deception," "sycophancy," "bias," "safety concerns"). The existence of features for psychological and social concepts — not just physical or linguistic ones — directly demonstrates that production-scale language models learn disentangled, interpretable representations of human-relevant psychological constructs.

The argument can be stated more concisely through the lens of Equation 1: the training objective of language modeling is $P(\text{next\_token} \mid \text{context})$. Since the next token was written by a human with particular psychological traits, and those traits systematically influence token choice (Step 1), a model that minimizes prediction error must represent those traits internally (Step 2). The representations are implicit — they were not designed to encode personality — but they are structured and extractable.

## 4 Implications for Downstream Applications

If language model representations implicitly encode human psychological traits, this has immediate implications for any application that requires understanding, modeling, or matching people. We briefly survey the landscape.

### 4.1 The Current State: Narrow, Siloed Representations

Today, every platform that personalizes its experience builds its own model of the user. Dating applications learn embeddings optimized for romantic compatibility from in-app behavior. Hiring platforms learn embeddings from resumes and interview transcripts. Content recommendation systems learn embeddings from engagement patterns. Social networks learn embeddings from interaction graphs.

Each of these representations captures a narrow slice of the person, optimized for a single task, in a format that is incompatible with every other platform. The same individual is remodeled from scratch each time they join a new service, losing all prior knowledge about who they are. This is the **cold-start problem** at the individual level, and it is entirely a consequence of siloed, task-specific representation.

### 4.2 The Opportunity: Pre-Trained Psychological Representations

The finding that general-purpose language models already encode psychological traits suggests a different paradigm: rather than learning person representations from scratch for each task, *begin with the implicit psychological knowledge already present in the pre-trained model* and adapt it for specific downstream objectives.

This parallels the broader trajectory of NLP, where task-specific models trained from scratch have been largely replaced by pre-trained models fine-tuned for specific tasks [Devlin et al., 2019]. The pre-trained model provides a rich feature space; fine-tuning selects and reorganizes the features relevant to the task at hand.

For person representation, this implies a two-stage pipeline:

1. **Extract**: Use the pre-trained model's implicit psychological features as a starting representation of a person, derived from any available text (profile, writing, communication).

2. **Adapt**: Fine-tune or project this representation for the specific downstream task (compatibility matching, role fit, content recommendation, etc.).

The critical insight is that different downstream tasks require different features of the same person to become salient. A dating platform needs romantic attachment style, lifestyle preferences, and values alignment to be foregrounded. A hiring platform needs professional skills, work style, and cultural fit. A mentorship platform needs learning style, domain expertise, and communication compatibility. The underlying person is the same; the *projection* differs.

This is directly analogous to the domain-adapted embedding models that have emerged for specialized retrieval: legal embedding models [Xiao et al., 2023] foreground legal reasoning features that general models underweight; financial embedding models foreground risk and regulatory features. In the same way, person-adapted embedding models could foreground the specific psychological features relevant to each application domain.

## 4.3 Downstream Application Taxonomy

We identify several categories of applications that would benefit from psychologically-informed person representations:

**Compatibility matching.** Dating, hiring, team formation, mentorship pairing. Requires representations that encode not just who a person *is*, but how they relate to others — complementarity, shared values, dealbreaker asymmetries.

**Personalization and recommendation.** Content, product, and experience recommendation based on deep preference modeling rather than surface-level engagement signals.

**AI agent personalization.** Memory systems, AI companions, and personal assistants that maintain a persistent model of the user's personality, communication style, and preferences.

**Human-computer and human-human interaction.** Tools that facilitate communication by modeling the psychological profiles of participants — detecting escalation in disputes, identifying semantic misunderstandings, adjusting communication style.

**Self-understanding.** Systems that help individuals understand their own psychological patterns — the modern, empirically grounded successor to personality tests and self-report instruments.

Each of these applications requires a different task-specific objective, and therefore a different organization of the representation space. This is not a limitation but a feature: the same underlying psychological knowledge can be projected into many task-relevant geometries.

# 5 Limitations and Open Problems

The finding that language models encode human psychology is a foundational observation, not a complete solution. Several critical limitations must be addressed before this implicit knowledge can be reliably exploited.

## 5.1 Semantic Similarity Is Not Compatibility

The dominant paradigm for embedding-based applications is similarity search: given a query, retrieve the items with the highest cosine similarity in embedding space. For document retrieval, this is often appropriate — a query about "climate change policy" should retrieve documents about climate change policy.

For person matching, similarity is fundamentally insufficient. Two people can be highly similar (same background, same values, same communication style) yet incompatible as romantic partners, colleagues, or collaborators. Compatibility is a *relational* property that depends on the interaction between two people's traits, not just the similarity of their individual representations.

Consider: an extremely introverted person seeking a partner does not want the most similar person (another extreme introvert); they may want someone moderately extroverted who draws them out. A startup seeking a CTO does not want someone maximally similar to the existing team; they want someone with complementary skills.

This means that naive similarity search over person embeddings will systematically fail for compatibility-oriented applications. Task-specific training objectives that directly optimize for compatibility — using contrastive losses on compatibility-labeled pairs rather than similarity-labeled pairs — are necessary. Empirical evidence confirms this: contrastive models trained on dealbreaker-annotated pairs achieve 79% accuracy on "similarity trap" evaluations that expose the gap between surface similarity and genuine compatibility [Politzki, 2026d]. The geometric constraints of encoding bilateral compatibility into unilateral vectors further complicate this picture [Politzki, 2026e].

## 5.2 Representations Are Task-Dependent

As established in Section 2, the training objective determines which features become salient. A language model trained for next-token prediction organizes its representations to predict linguistic patterns. A model fine-tuned for sentiment analysis reorganizes to foreground emotional valence. A model fine-tuned for compatibility matching reorganizes to foreground relational features.

This means there is no single "correct" embedding of a person. The same individual will have different representations depending on the task. This is not a bug — it reflects the genuine fact that different aspects of a person are relevant in different contexts. But it does mean that claims of "universal person embeddings" must be qualified: universality refers to the breadth of the underlying feature space, not to a single fixed-point representation.

The analogy to domain-adapted embedding models is instructive. Just as there is no single embedding model that is optimal for both legal document retrieval and medical literature search, there may be no single person embedding that is optimal for both dating and hiring. The base model provides a rich general-purpose feature space; task-specific adaptation selects the relevant subspace.

## 5.3 Dimensional Collapse and the Curse of Dimensionality

Current embedding architectures face two related pathologies.

**Dimensional collapse.** Jing et al. [2022] showed that contrastive learning objectives can produce representations that utilize only a fraction of the available dimensions, effectively reducing the representational capacity. In the context of person embeddings, this means that important psychological dimensions may be collapsed into a low-rank subspace, losing discriminative information.

**The curse of dimensionality.** As embedding dimensionality increases, data points become increasingly sparse in the space, making meaningful geometric relationships harder to learn and exploit. This creates a tension: representing the full complexity of a human being may require thousands of dimensions, but learning reliable structure in such high-dimensional spaces requires correspondingly more data.

These pathologies suggest that naive scaling of embedding dimensionality is not sufficient. Structured approaches — such as hierarchical embeddings, multi-resolution representations, or compositional architectures that decompose a person into interpretable sub-representations — may be necessary to faithfully represent human complexity within the constraints of current embedding geometry.

## 5.4 Strategic Self-Presentation and Context Dependence

People do not produce text that is a pure, unfiltered projection of their psychology. They engage in strategic self-presentation [Goffman, 1959], emphasizing different facets of themselves in different contexts. A person's LinkedIn profile, dating profile, and text messages to close friends may project genuinely different aspects of the same underlying personality.

This does not invalidate the foundational argument — the psychological signal is still present, merely filtered — but it does mean that any person representation derived from a single text source captures a *context-dependent projection* of the person, not the full underlying psychology. Richer representations may require integrating signals across multiple contexts, raising both technical challenges (how to fuse heterogeneous text sources) and ethical challenges (how to respect context-dependent privacy expectations).

## 5.5 The Question of Stable Core Identity

The entire enterprise of person embedding assumes that people possess stable psychological traits worth capturing. The evidence is mixed. Big Five personality traits show moderate stability over decades [Roberts and DelVecchio, 2000], and core values are similarly persistent. But behavior is also highly context-dependent [Mischel, 1968], and people genuinely change over time.

A faithful person representation must therefore be understood as a distribution over possible behaviors and presentations, not a fixed point in embedding space. This has implications for architecture (static vectors may be insufficient; temporal or distributional representations may be needed) and for application design (representations should be updatable, and downstream systems should account for uncertainty).

## 5.6 Multi-Embedding and Bi-Encoder Architectures

A single dense vector may be fundamentally insufficient to represent the full complexity of a human being. Just as modern information retrieval has moved toward multi-vector representations (ColBERT [Khattab and Zaharia, 2020]), late-interaction models, and cross-encoder rerankers that capture fine-grained token-level interactions, person representation may require:

- **Multi-vector representations**: Separate embeddings for different facets of identity (personality, values, skills, preferences), composed at query time for task-specific matching.

- **Bi-encoder architectures**: Separate encoders for different text types (profiles vs. behavior logs vs. writing samples), with learned fusion.
- **Hierarchical representations**: Coarse-grained identity embeddings for rapid filtering, fine-grained sub-embeddings for nuanced matching — potentially leveraging Matryoshka-style multi-resolution training [Kusupati et al., 2022].

The design space for person representation architectures is largely unexplored, and mapping it is a key direction for future work.

# 6 Discussion

## 6.1 What This Paper Does and Does Not Claim

We claim that language model representations implicitly encode human psychological traits, as a necessary consequence of being trained to predict human-generated text. We support this claim with convergent evidence from psycholinguistics, emergent capabilities research, and mechanistic interpretability.

We do *not* claim that current embedding models provide ready-made person representations. The gap between implicit psychological encoding and useful person embeddings is significant, requiring task-specific training objectives, architectural innovations, and careful treatment of the limitations enumerated in Section 5.

We do *not* claim that embeddings can fully capture a human being. People are not vectors. But useful computational approximations of psychological traits are achievable and increasingly necessary for the applications that define the intelligence age.

## 6.2 The Broader Context: Why This Matters Now

Every major AI application that interacts with people — recommendation, matching, personalization, communication, education — requires some model of the user. Today, these models are built from scratch, in silos, using narrow behavioral signals. The finding that pre-trained language models already contain rich psychological representations suggests a fundamentally different approach: start from the implicit knowledge, adapt for the task.

This shift parallels the broader transition in NLP from task-specific architectures to pre-trained foundation models. The "foundation model" for person representation may not require a new training paradigm — it may already exist in the representations learned by current language models, waiting to be extracted, organized, and applied.

## 6.3 Ethical Considerations

The ability to computationally represent human psychology raises profound ethical questions. Who owns a person's embedding? Who should have access? Can embeddings be used for manipulation, surveillance, or discrimination? These questions are not new — they arise with any personal data — but the comprehensiveness of psychologically-informed embeddings makes them more acute.

We believe that responsible development requires: (1) individual ownership and control of personal embeddings, (2) irreversibility guarantees that prevent reconstruction of source text from embeddings, (3) purpose limitation that restricts embeddings to consented applications, and (4) neutral, trusted infrastructure that is not controlled by any single application or platform. We leave the detailed design of such infrastructure to future work, noting only that the technical capability described in this paper makes the governance question urgent.

## 7 Conclusion

We have presented a logical argument, supported by convergent empirical evidence, that large language models trained on text implicitly learn structured representations of human psychological traits. The argument rests on three well-established pillars: language is a systematic projection of the mind, language models learn structured world models that go beyond surface statistics, and the training objective of language modeling necessarily requires encoding the psychological processes that generate human text.

This finding is foundational rather than prescriptive. We do not propose a specific system for exploiting this implicit knowledge. Instead, we establish the base claim and identify the critical challenges that must be addressed: the task-dependence of representations, the insufficiency of semantic similarity for relational tasks, the pathologies of current embedding architectures, and the philosophical complexity of representing a human being as a vector.

The downstream applications are broad — compatibility matching, personalization, AI agents, communication tools, self-understanding — and each requires task-specific adaptation of the general psychological features that language models have already learned. Building the bridge from implicit encoding to useful application is the research program that follows from this paper.

This paper is the first in a series. The immediate next question is: if different training objectives produce different representations of the same person, what is the geometric relationship between those representations? Do representations trained on different domains converge on shared structure, and can that structure be measured, aligned, and exploited? A companion survey [Politzki, 2026a] addresses these questions by examining the landscape of methods for geometric alignment across representation spaces. A third paper [Politzki, 2026b] proposes one specific technical mechanism — Shared Matryoshka Embeddings — for cross-domain prefix alignment. A fourth [Politzki, 2026c] synthesizes the preceding work into a concrete research program for universal human embeddings.

## References

Bengio, Y., Courville, A., and Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*, Anthropic, 2023.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering Latent Knowledge in Language Models Without Supervision. In *Proceedings of ICLR*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 2019.

Goffman, E. *The Presentation of Self in Everyday Life.* Anchor Books, 1959.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning.* MIT Press, 2016.

Gurnee, W. and Tegmark, M. Language Models Represent Space and Time. In *Proceedings of ICLR*, 2024.

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. Understanding Dimensional Collapse in Contrastive Self-supervised Learning. In *ICLR*, 2022.

Khattab, O. and Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of SIGIR*, 2020.

Kosinski, M. Evaluating Large Language Models in Theory of Mind Tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sber, A., Shanber, V., Simhadri, H. V., and Jain, P. Matryoshka Representation Learning. In *Advances in Neural Information Processing Systems*, 2022.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *Proceedings of ICLR*, 2023.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of ICLR Workshop*, 2013.

Mischel, W. *Personality and Assessment.* Wiley, 1968.

Moreno, J. D., Martinez-Huertas, J. A., Olmos, R., Jorge-Botana, G., and Botella, J. Can personality traits be measured analyzing written language? A meta-analytic study on computational methods. *Personality and Individual Differences*, 177:110818, 2021.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of UIST*, 2023.

Pennebaker, J. W. *The Secret Life of Pronouns: What Our Words Say About Us.* Bloomsbury Press, 2011.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP*, 2019.

Roberts, B. W. and DelVecchio, W. F. The Rank-Order Consistency of Personality Traits from Childhood to Old Age: A Quantitative Review of Longitudinal Studies. *Psychological Bulletin*, 126(1):3–25, 2000.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9):e73791, 2013.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024.

Sutton, R. S. The Bitter Lesson. Blog post, `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`, 2019.

Tausczik, Y. R. and Pennebaker, J. W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

Templeton, A., Conerly, T., Marcus, J., Lindamey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*, Anthropic, 2024.

Wolpert, D. H. and Macready, W. G. No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. C-Pack: Packaged Resources To Advance General Chinese Embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Youyou, W., Kosinski, M., and Stillwell, D. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

Politzki, J. Geometric Alignment Across Representation Spaces: A Survey of Measurement, Methods, and Convergence. Working paper, 2026.

Politzki, J. Shared Matryoshka Embeddings: Cross-Domain Prefix Alignment of Human Representations. Working paper, 2026.

Politzki, J. Towards Universal Human Embeddings. Working paper, 2026.

Politzki, J. Compatibility Is Not Similarity: Contrastive Learning for Asymmetric Human Matching. Working paper, 2026.

Politzki, J. The Dyadic Compression Problem: Does Contrastive Compatibility Training Learn Identity or Matching Geometry? Working paper, 2026.