

A Vision for Cultural Alignment: Opportunities and Safety Imperatives for AI in Mental Health Support

Ratna Kandala¹ Akshata Kishore Moharir² Niva Manchanda¹ Samantha Adorno¹

¹ University of Kansas, USA

²Independent Researcher, USA

Abstract

As Large Language Models (LLMs) are increasingly integrated into mental health applications, the prevailing paradigm of *cultural neutrality* risks creating a technical debt that undermines model generalization and safety. Current systems/frameworks, while focusing on multilingual capabilities, often fail to account for the diverse social and cognitive norms of global populations. This vision paper explores the transition from culturally invariant models toward a research trajectory that operationalizes cultural dimensions such as power distance and individualism as fundamental architectural parameters. Rather than viewing cultural adaptation as a post-hoc fine-tuning task, we advocate for systems that internally encode cultural variability as a primary driver of the model's reasoning process. We outline a roadmap for architecturally encoding cultural variability as a first-class parameter in AI/ML pipelines, moving beyond surface-level translation toward systems where cultural dimensions fundamentally shape model reasoning.

1 Introduction

On a late evening in Mumbai, India, Priya, a tech professional, seeks chatbot support for mounting workplace mental stress. The system, trained on Western therapeutic frameworks, advises her to "set clear boundaries" and "prioritize her own needs". However, in Priya's cultural context where hierarchical respect and collective harmony are paramount (Markus and Kitayama, 1991; Kapoor et al., 2003; Chadda and Deb, 2013), such direct confrontation could jeopardize her professional standing and family stability.

This scenario illustrates a critical but largely overlooked challenge as LLMs are increasingly deployed in mental health applications (Ali et al., 2025): the implicit assumption of cultural neutrality (Joseph, 2025). Current mental health AI systems, while achieving impressive multilingual

capabilities and demonstrating significant performance in benchmarks (Arora et al., 2023b; Hua et al., 2025), fundamentally fail to account for the diverse social, cognitive, and cultural aspects (Harigian et al., 2020; Hershcovich et al., 2022; Pucci et al., 2026) that shape how mental health is understood, experienced, and addressed across global populations.

Recent studies iterated that leading LLMs exhibit systematic biases toward Western, Educated, Industrialized, Rich, and Democratic (WEIRD) cultural values, particularly emphasizing individualistic frameworks even when interacting with users from collective cultural contexts (Atari et al., 2023; Tao et al., 2024). These systems inherit psychological assumptions, for example, conversational therapy bots like Wysa and Woebot, developed almost entirely from Western clinical databases, incorrectly interpreted culturally characteristic Indian expressions, pathologizing statements such as "family pressure is my karma" as depressive symptoms requiring unsuitable therapeutic interventions (Lakshmi et al., 2025).

Foundational research in cultural psychology challenges these assumptions: emotions themselves are not universal but culturally constructed (Barrett, 2017; Jack et al., 2012), and therapeutic approaches embody culturally-specific emotion concepts (Hoemann et al., 2024). Mesquita's (Mesquita, 2022) distinction between MINE (Mental, INside person, Essentialist) and OURS (OUtside person, Relational, Situational) emotional models reveals how Western individualistic cultures prioritize internal feelings and personal autonomy, while collectivist cultures view emotions as serving relational harmony and social cohesion.

The stakes are substantial: from the concept of "self" to help-seeking behaviors, mental health is culturally constructed (Kirmayer, 2007; Kleinman, 1988). The DSM-5's inclusion of a Cultural Formulation Interview acknowledges that "culture

provides interpretive frameworks that shape the experience and expression of symptoms" (American Psychiatric Association, 2013b), yet current AI systems, despite processing text in dozens of languages, operate with reasoning patterns, therapeutic orientations, and conceptualizations of mental wellness rooted predominantly in Western individualistic paradigms (Cao et al., 2023; Kandala et al., 2025).

This cultural misalignment constitutes fundamental technical debt that undermines model generalization and safety (Software Engineering Institute, Carnegie Mellon University, 2015). When models trained on Western literature encounter users from high power-distance cultures, the resulting misalignment can lead to ineffective or harmful counsel. Current approaches that treat cultural adaptation as post-hoc translation fail because culture shapes the fundamental reasoning patterns embedded in support (Hershcovich et al., 2022).

This position paper argues for a paradigm shift: from culturally invariant models toward systems that architecturally encode cultural variability. We focus on Hofstede's Individualism-Collectivism dimension (Hofstede, 2001), a cultural framework that profoundly influences mental health conceptualizations. Our contributions are threefold: First, we provide critical analysis of cultural neutrality as technical debt, documenting concrete failure modes. Second, we outline a framework for cultural architecture, sketching technical approaches to encode cultural dimensions as fundamental parameters rather than post-hoc adaptations (Arora et al., 2023b). Third, we present a research roadmap that identifies priority directions for datasets, evaluation frameworks, and deployment safeguards needed to transition from surface-level translation to deep cultural alignment.

2 Background and Motivation

Cross-cultural research in psychology systematically challenges the universality of mental health constructs as cultural context shapes how individuals experience, express, and seek help for distress, reflecting deep differences in self-concept, social relationships, and meaning-making. These differences profoundly influence help-seeking: research with Asian American adolescents shows that collectivistic values reduce professional help-seeking even controlling for severity, mediated by family shame and loss of face (Shea and Yeh, 2008; Zheng

et al., 2021). Cultural pressures against disclosure stem from traditional norms where "excessive self-disclosure and strong emotional expression" are seen as "disruptive acts against collective harmony and family honor" (Kim et al., 2001). Asian Americans view requests for social support with concern about "burdening others, disrupting the harmony of the group, making the problem worse, and bringing shame to oneself or one's family" (Kim et al., 2008). Family involvement differs accordingly: collectivistic cultures expect family participation in treatment while individualistic cultures prioritize confidentiality (Chadda and Deb, 2013), meaning AI built on Western privacy assumptions may reduce engagement in collectivistic contexts.

LLMs frame work-family conflict as individual optimization (boundary setting, time management), assuming renegotiating obligations is acceptable, clashing with evidence that collectivist settings rely on norm conformity and social support (Liu et al., 2022). Post-hoc adaptation-translation, prompt engineering, and fine-tuning are insufficient because they preserve underlying reasoning patterns (Hershcovich et al., 2022). In addition to these, prompt engineering ("respond from a collectivist perspective") works inconsistently, is overridden in multi-turn dialogue, and leaves internal representations unchanged (Arora et al., 2023b). Also, small cultural fine-tuning datasets adjust style without altering deep WEIRD priors; models revert to individualistic defaults under distribution shift (Masoud et al., 2025). Therefore, we argue that a cultural architecture would inherently integrate culture: representing relational selves, encoding social roles and obligations, and learning culturally grounded notions of harm, help-seeking, and coping.

3 Toward Culturally-Aware AI Architecture: A Research Agenda

In this section, we outline our vision for *cultural architecture*: designing systems to encode cultural variability as an architectural parameter rather than treating it as post-hoc localization. We propose core design principles, an operational strategy for cultural dimensions, and sketch four complementary technical approaches. Our goal here is not to present a complete solution- culturally oriented AI remains an open challenge, but to chart a path forward for industry researchers working towards leveraging AI for the rising mental health challenges globally.

3.1 Core Principles of Cultural Architecture

Culture as architectural parameter. Similar to language-specific tokens in multilingual models (Lample and Conneau, 2019), cultural parameters condition model behavior, not as post-hoc localization but as first-class architectural components influencing reasoning, generation, and interaction patterns.

Explicit modeling vs. implicit default. Current systems claim neutrality while defaulting to Western norms. When an LLM advises "set boundaries with your parents" in response to family conflict, it implicitly assumes individualistic autonomy is appropriate. Explicit cultural modeling makes such assumptions transparent and adjustable.

Multi-cultural competence. Drawing inspiration from neurocognitive code-switching in bilingual speakers, systems should handle multiple cultural frameworks simultaneously, rather than requiring separate fine-tuned models per culture, enabling natural transitions between cultural contexts.

Cultural parameters as priors, not constraints. Cultural dimensions describe population tendencies, not individual determinants. A person from a collectivistic culture may value personal autonomy in specific contexts. Systems use cultural information as Bayesian priors that update through interaction (Gelman et al., 2013), not rigid categorizations.

Support for hybrid identities. Second-generation immigrants, multicultural families, and globally mobile individuals navigate multiple cultural frameworks (Benet-Martínez and Haritatos, 2005). Systems need composable representations that blend cultural parameters fluidly, not binary categorizations.

3.2 Operationalizing Cultural Dimensions

The central challenge is transforming qualitative cultural constructs into quantifiable parameters. We propose Hofstede’s Individualism-Collectivism (I-C) dimension as our starting point, a 0-100 scale derived from validated surveys across 76 countries (Hofstede, 2001). The USA scores 91 (highly individualistic), China scores 20 (highly collectivistic), correlating with observable patterns in therapeutic practices and help-seeking behaviors (Triandis, 1995). We normalize this to a continuous parameter $c \in [0,1]$ conditioning model behavior, where $c=1$ represents maximum individualism and $c=0$ represents maximum collectivism. For identical input

("I’m stressed about family obligations conflicting with my career goals"), individualistic conditioning ($c=0.9$) yields: "It’s important to honor your own aspirations. What boundaries might help?" while collectivistic conditioning ($c=0.2$) yields: "How might you discuss this with your family to find a path that fulfills your responsibilities...?" Current LLMs default to individualistic patterns regardless of context.

Beyond I-C and avoiding stereotypes. Comprehensive systems would incorporate additional dimensions-Power Distance (therapeutic relationship expectations), Uncertainty Avoidance (diagnostic ambiguity comfort), as a multi-dimensional cultural vector. Critical safeguards prevent stereotyping: cultural parameters are continuous (not categorical), serve as Bayesian priors that update through interaction (Gelman et al., 2013), and are user-adjustable. Users can explicitly override assumptions, just as they set language preferences.

Hybrid identities. Bicultural individuals, second-generation immigrants, multicultural families navigate multiple frameworks simultaneously (Benet-Martínez and Haritatos, 2005). Systems must support: (a) uncertainty estimates updating with evidence, (b) contextual variation (collectivistic framing for family discussions, individualistic for career contexts), and (c) cultural code-switching between frameworks. This requires continuous, compositional cultural representations rather than discrete categories.

3.3 Technical Approaches

Below, we present four technical approaches to implement the idea discussed so far. It is crucial to note that these are not mutually exclusive, and robust systems may combine multiple strategies, as deemed necessary. These approaches could be viewed as different entry points into the problem, each with distinct trade-offs w.r.t data requirements, computational cost, interpretability, and ease of deployment. This requires moving from discrete cultural categories toward continuous, compositional cultural representations, a technical challenge we address in the next section.

Option 1: Culturally-Conditioned Language Models: The most direct approach extends the conditional generation paradigm: modern LLMs already condition on various factors - control codes for style, domain and other attributes (Keskar et al., 2019; Ouyang et al., 2022). We propose adding cultural parameters to this conditioning set via cul-

tural tokens - analogous to the language identifiers in multilingual models (Lample and Conneau, 2019) - such as [I] or [C] to specify individualism-collectivism orientation. For instance, when a user discusses family-career conflict, a system conditioned on individualistic parameters might respond: "This is your life and your decision. What feels more authentic to you?" In contrast, collectivistic conditioning yields: "Let's think about how to discuss this with your family in a way that honors their hopes while expressing your interests." Same input, culturally-appropriate variation in framing.

Option 2: Mixture-of-Cultural-Experts: Mixture of Experts (MoE) architectures can offer specialized modules for different cultural reasoning patterns, combined via learned gating (Shazeer et al., 2017). For example, Expert 1 specializes in individualistic therapeutic reasoning (trained on Western therapy data); Expert 2 handles collectivist patterns (e.g., Asian/Latin American contexts); Expert 3 manages hybrid bicultural reasoning. The gating network routes inputs to appropriate experts based on cultural parameters and context, enabling natural cultural code-switching for bicultural users. This modularity improves interpretability (in this case, examining which experts' activation reveals cultural framing), enables extensibility (adding cultural expertise without full retraining), and reduces interference during training. In this regard, Google's Switch Transformer (Fedus et al., 2022), and multilingual models with language-specific experts (Lepikhin et al., 2021) provide technical precedent.

Option 3: Meta-Learning for Cultural Adaptation: Both previous approaches require substantial culturally-labeled training data. But high-quality mental health data is scarce for most non-Western contexts (Saxena et al., 2007). Meta-learning offers a solution: train models to learn *how to adapt* to new cultural contexts with minimal data (Finn et al., 2017). Model-Agnostic Meta-Learning (MAML) finds model parameters that, after a few gradient steps on a new cultural context, achieve strong performance. This addresses data scarcity: given just 10-20 examples of culturally-appropriate therapeutic interactions in an under-resourced context (e.g., indigenous communities), a meta-learned model could adapt effectively while preserving general capabilities. Meta-learning has been successfully applied to domain adaptation (Li et al., 2018), personalization (Finn et al., 2019), and cross-lingual transfer (Nooralahzadeh et al., 2020). Cultural adaptation is a natural extension of learning to learn

across cultural distributions rather than linguistic ones.

Option 4: Multi-Task Learning with Cultural Objectives: Rather than hoping cultural awareness emerges from standard pre-training, we propose incorporating it explicitly through auxiliary objectives (Ruder, 2017). (a) *Cultural value prediction:* Given a therapy dialogue snippet, predict associated cultural dimension values (e.g., individualism-collectivism score) to encourage learning culturally-sensitive representations (Arora et al., 2023a). (b) *Contrastive cultural learning:* Train models to assign higher probability to culturally-appropriate responses than inappropriate ones for a given context (Chen et al., 2020), directly optimizing for cultural alignment rather than hoping it emerges from language modeling alone. (c) *Cultural reasoning tasks:* Include explicit reasoning about cultural appropriateness, explaining why responses are/aren't suitable, generating culturally-adapted paraphrases, or predicting cross-cultural interpretation differences. This multi-task approach builds cultural awareness from pre-training rather than treating it as post-hoc adaptation.

4 Ethical Considerations and Path Forward

Architecting cultural variability into LLMs for mental health raises critical tensions. We must avoid exporting WEIRD assumptions as defaults while also preventing culture from becoming a static template that overrides individual needs. Current LLMs show systematic cultural value skews (Masoud et al., 2025) and reproduce cultural stereotypes in interpersonal contexts (Yuan et al., 2025) - behaviors that are safety-relevant in mental health applications because they influence what systems optimize for (self-advocacy versus relational harmony) and treat as risk signals (family involvement versus coercion). In this context, three ethical risks demand attention: **(1) Reducing individuals to stereotypes** if culture is implemented as fixed labels rather than soft, reversible priors that generate options rather than directives (American Psychiatric Association, 2013a); **(2) Governance and privacy questions about who decides cultural appropriateness**, training data curation, policy constraints, and evaluator selection are governance choices requiring stakeholder involvement (Delgado et al., 2023); and **(3) Power dynamics** in cultural data collection that can mirror extrac-

tive patterns, requiring consentful, context-specific governance and contestability mechanisms (Lewis et al., 2020).

Technical and deployment challenges. Significant open questions remain: High-quality mental health data is scarce for most non-Western cultures - how do we build culturally-aware systems without reinforcing data colonialism (Couldry and Mejias, 2019)? How do systems determine cultural context without stereotyping based on location, yet without delaying appropriate responses? Standard NLP metrics cannot assess cultural appropriateness; we need evaluation frameworks involving cross-cultural mental health experts and community validation (Benrimoh et al., 2023). Supporting diaspora, multicultural identities, and within-culture variation, where users code-switch across contexts, requires moving beyond national or language proxies toward preference elicitation (Kirmayer and Minas, 2000). Finally, regulatory frameworks (FDA, CE marking) assume fixed systems; culturally-adaptive AI requires new approaches to safety validation.

Recommendations for industry. For organizations deploying mental health AI globally, we recommend: (1) *Start with transparency* make current cultural assumptions explicit and allow users to set cultural preferences alongside language selection; (2) *Phased deployment* - begin with cultural conditioning (Option 1) using validated dimensions, monitoring for misalignments before advancing to complex architectures; (3) *Invest in culturally-diverse evaluation*-standard benchmarks reflect Western contexts, requiring new datasets developed with mental health professionals from target cultures; (4) *Regulatory engagement* work with regulators to develop validation frameworks for adaptive systems; and (5) *Community partnership* - meaningful collaboration with mental health communities across cultures in data sharing, validation, and governance.

Research priorities for the NLP community. Urgent questions include: What constitutes valid metrics for cultural appropriateness beyond country-level survey averages (Masoud et al., 2025)? What abstention and clarification policies are appropriate when cultural signals conflict? What participatory mechanisms give communities substantive power over objectives rather than only post-hoc feedback (Delgado et al., 2023)? How can cultural conditioning avoid encoding stereotypes, given evidence that LLMs amplify cultural biases (Yuan et al., 2025)?

What consent and data control models are feasible in stigmatized mental health contexts (Lewis et al., 2020)?

5 Conclusion

A credible counter-argument holds that core psychological needs transcend cultural boundaries, making cultural layers unnecessary complexity. However, while needs may be universal, their *expression* is strictly cultural, ignoring these nuances imposes majority values. Cultural architecture is not merely an academic concern but a clinical safety and ethical imperative. Digital mental health support cannot be culturally neutral; it must be culturally aware, adaptive, and accountable to the communities it serves.

References

- Maisam Ali, Shahid Ali, Qaiser Abbas, Zeeshan Abbas, and Seung Won Lee. 2025. [Artificial intelligence for mental health: A narrative review of applications, challenges, and future directions in digital health](#). *Digital Health*, 11:20552076251395548.
- American Psychiatric Association. 2013a. [Cultural formulation interview \(cfi\)](#). PDF. DSM-5 Cultural Formulation Interview.
- American Psychiatric Association. 2013b. [Diagnostic and Statistical Manual of Mental Disorders](#), 5 edition. American Psychiatric Association Publishing.
- Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023a. [Probing pre-trained language models for cross-cultural differences in values](#). In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.
- Daman Arora, Himanshu Singh, and Mausam. 2023b. [Have LLMs advanced enough? a challenging problem solving benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.
- Mohammad Atari, Meng J. Xue, Peter S. Park, Damián E. Blasi, and Joseph Henrich. 2023. [Which humans?](#) OSF Preprints.
- Lisa Feldman Barrett. 2017. *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt, Boston, MA.
- Verónica Benet-Martínez and Jana Haritatos. 2005. The bicultural identity integration scale: Development and validation. *Journal of personality*, 73(4).

- David Benrimoh, Sonia Israel, Rebeca Fratila, Kelly Perlman, Marc Miresco, and Srividya Iyer. 2023. [Your robot therapist is not your therapist: Understanding the role of ai-powered mental health chatbots](#). *Frontiers in Digital Health*, 5:1278186.
- Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. [Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study](#). In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.
- Rakesh K Chadda and Koushik S Deb. 2013. [Indian family systems, collectivistic society and psychotherapy](#). *Indian Journal of Psychiatry*, 55(Suppl 2):S299–S305. Documents how Indian collectivism emphasizes family over individual needs.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.
- Nick Couldry and Ulises A. Mejias. 2019. Data colonialism: Rethinking big data’s relation to the contemporary subject. *Television & New Media*, 20(4):336–349.
- Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. [The participatory turn in ai design: Theoretical foundations and the current state of practice](#). In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO ’23)*.
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–40.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.
- Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. 2019. Online meta-learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 1920–1930. PMLR.
- Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian data analysis*, 3rd edition. CRC press.
- Keith Harrigan, Carlos Aguirre, and Mark Dredze. 2020. [Do models of mental health based on social media data generalize?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788, Online. Association for Computational Linguistics.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and strategies in cross-cultural NLP](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.
- Katie Hoemann, Maria Gendron, Alyssa N. Crittenden, Sillo M. Mangola, Evaline S. Endeko, Ève Dussault, Lisa Feldman Barrett, and Batja Mesquita. 2024. [What we can learn about emotion by talking with the hadza](#). *Perspectives on Psychological Science*, 19(1):173–200.
- Geert Hofstede. 2001. *Culture’s consequences: Comparing values, behaviors, institutions and organizations across nations*, 2nd edition. Sage publications.
- Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2025. [A scoping review of large language models for generative tasks in mental health care](#). *npj Digital Medicine*, 8(1):230.
- Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. 2012. [Facial expressions of emotion are not culturally universal](#). *Proceedings of the National Academy of Sciences*, 109(19):7241–7244.
- Jeena Joseph. 2025. [Algorithmic bias in public health ai: a silent threat to equity in low-resource settings](#). *Frontiers in Public Health*, 13:1643180.
- Ananth Kandala, Ratna Kandala, Akshata Kishore Moharir, Niva Manchanda, and Sunaina Singh Rathod. 2025. [Cross-lingual mental health ontologies for Indian languages: Bridging patient expression and clinical understanding through explainable AI and human-in-the-loop validation](#). In *NLP-AI4Health*, pages 16–24, Mumbai, India. Association for Computational Linguistics.
- Suraj Kapoor, Paul C. Hughes, John R. Baldwin, and Joel Blue. 2003. [The relationship of individualism–collectivism and self-construals to communication styles in india and the united states](#). *International Journal of Intercultural Relations*, 27(6):683–700.
- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. [Ctrl: A conditional transformer language model for controllable generation](#). *Preprint*, arXiv:1909.05858. ArXiv:1909.05858v2.

- Bryan S. K. Kim, Donald R. Atkinson, and Dana Umemoto. 2001. [Asian cultural values and the counseling process: Current knowledge and directions for future research](#). *The Counseling Psychologist*, 29(3):342–352.
- Heejung S. Kim, David K. Sherman, and Shelley E. Taylor. 2008. [Culture and social support](#). *American Psychologist*, 63(6):518–526.
- Laurence J. Kirmayer. 2007. [Psychotherapy and the cultural concept of the person](#). *Transcultural Psychiatry*, 44(2):232–257.
- Laurence J Kirmayer and Harry Minas. 2000. The future of cultural psychiatry: an international perspective. *The Canadian Journal of Psychiatry*, 45(5):438–446.
- Arthur Kleinman. 1988. *Rethinking Psychiatry: From Cultural Category to Personal Experience*. Free Press, New York.
- Divya S. Lakshmi, Visakh Mohan, P. S. Reesa, Elizabeth Alexander, Neethu Varghese, and Alphonsa Kurian. 2025. [Whose bias gets coded? Psychology’s role in decolonizing AI](#). *Frontiers in Psychology*, 16:1657227.
- Guillaume Lample and Alexis Conneau. 2019. [Cross-lingual language model pretraining](#). In *Advances in Neural Information Processing Systems*.
- Dmitry Lepikhin, HyounJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations (ICLR)*.
- Jason Edward Lewis, the Indigenous Protocol, and AI Working Group. 2020. Indigenous protocol and artificial intelligence position paper. Position paper.
- Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Xuejun Liu, Shuming Zhao, and Jun Li. 2022. [Power distance belief and workplace communication: The mediating role of fear of authority](#). *International Journal of Environmental Research and Public Health*, 19(5):2932. Published online 2022; documents obstacles to workplace communication in high power distance contexts.
- Hazel Rose Markus and Shinobu Kitayama. 1991. [Culture and the self: Implications for cognition, emotion, and motivation](#). *Psychological Review*, 98(2):224–253.
- Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2025. [Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.
- Batja Mesquita. 2022. *Between Us: How Cultures Create Emotions*. W. W. Norton & Company, New York, NY.
- Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. [Zero-shot cross-lingual transfer with meta learning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Giulia Pucci, Emily Hemendinger, Ruizhe Li, Gavin Abercrombie, Tanvi Dinkar, and Arabella Sinclair. 2026. [Food noise & false safety: A systematic evaluation of how llms fail to adapt to eating disorder queries with clinician feedback](#). *Preprint*, arXiv:2606.02444.
- Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
- Shekhar Saxena, Graham Thornicroft, Martin Knapp, and Harvey Whiteford. 2007. [Resources for mental health: scarcity, inequity, and inefficiency](#). *The Lancet*, 370(9590):878–889.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*.
- Munyi Shea and Christine J. Yeh. 2008. [Asian american students’ cultural values, stigma, and relational self-construal: Correlates of attitudes toward professional help seeking](#). *Journal of Mental Health Counseling*, 30(2):157–172.
- Software Engineering Institute, Carnegie Mellon University. 2015. A field study of technical debt. Blog post.
- Yan Tao, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. [Cultural bias and cultural alignment of large language models](#). *PNAS Nexus*, 3(9):pgae346.
- Harry C. Triandis. 1995. *Individualism & collectivism*. Westview press.

- Hang Yuan, Zhongyue Che, Yue Zhang, Shao Li, Xi-
anger Yuan, Liqin Huang, Xiaomeng Hu, Kaiping
Peng, and Siyang Luo. 2025. [The cultural stereotype
and cultural bias of chatgpt](#). *Journal of Pacific Rim
Psychology*, 19:18344909251355673.
- Shaofeng Zheng, Takahiko Masuda, Masahiro Mat-
sunaga, Yasuki Noguchi, Yasuki Noguchi, Yohsuke
Ohtsubo, Hidenori Yamasue, and Keiko Ishii. 2021.
[Cultural differences in social support seeking: The
mediating role of empathic concern](#). *PLOS ONE*,
16(12):e0262001.