# AI Tools Can Enhance, Not Threaten, Generalizability

Zhicheng Lin
Department of Psychology, Yonsei University

Using large language models (LLMs) to replace human participants suffers from fundamental fallacies: overgeneralization from WEIRD (Western, Educated, Industrialized, Rich, Democratic) samples, conflation of linguistic form with psychological content, and neglect of embodied and social dimensions of cognition [1]. Crockett and Messeri [2] extend this critique, arguing that AI as surrogates perpetuates generalizability problems by entrenching WEIRD samples and decontextualized tasks. Their analysis is correct but incomplete: it conflates LLMs as surrogates that replace participants with LLMs as tools that make diverse human research more tractable.

The persistent lack of generalization and diversity in cognitive sciences reflects structural barriers: historical populations are inaccessible, cross-cultural studies costly, marginalized communities resistant to conventional recruitment. As methodological tools paired with human validation, LLMs lower these barriers without creating surrogates. **Figure 1** operationalizes this distinction. When human data collection is feasible—including cross-cultural studies and rare populations—LLMs function as pilots that refine designs before human validation. When genuinely infeasible—historical populations, extreme contexts—LLMs serve as computational models validated through theoretical coherence and expert assessment. Three pathways illustrate this distinction.

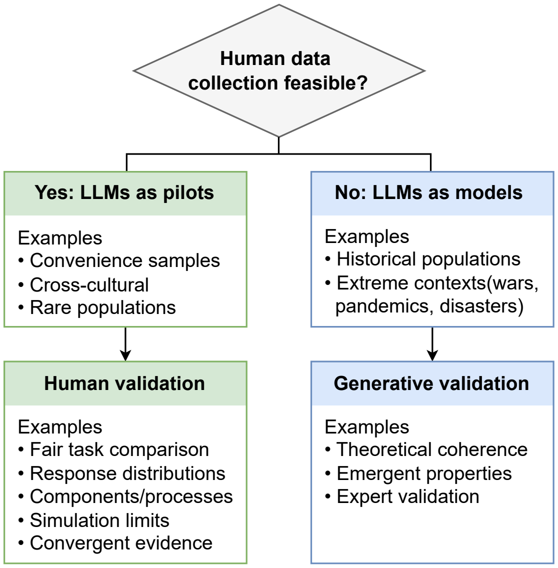

***Figure 1. Decision framework for using LLMs to enhance generalizability in cognitive science.*** *The framework distinguishes two categories based on whether human data collection is feasible. Left (feasible): when studying accessible populations—including convenience samples, cross-cultural populations, and rare groups—LLMs function as methodological tools to refine research designs before mandatory human validation. This pilot-testing approach addresses generalizability threats by identifying instrument failures, cultural blind spots, and measurement non-equivalence before committing resources to international or hard-to-reach samples.*

*Validation requirements include fair task comparison (ensuring construct-appropriate measurement), response distribution analysis (examining central tendencies and variability), component/process testing (validating constituent mechanisms), simulation limit probing (identifying boundary conditions), and convergent evidence (triangulating across multiple empirical benchmarks). Right (infeasible): when studying temporally inaccessible populations (historical eras) or ethically/logistically constrained contexts (wars, pandemics, disasters), LLMs function as computational models that expand generalizability along dimensions unavailable to human recruitment. These applications require generative validation through theoretical coherence (consistency with established psychological frameworks), emergent properties (higher-order patterns arising from component interactions), and expert assessment (domain specialist evaluation). Both categories address generalizability constraints—the left by making diverse human research more tractable, the right by accessing phenomena otherwise excluded from empirical study—but differ in their evidentiary standards based on the availability of direct human comparison data [4].*

**LLMs make the impossible tractable**

Temporal inaccessibility is absolute—we cannot recruit participants from past eras. This constrains psychology to a narrow temporal slice of human variation. LLMs trained on historical corpora—classical Chinese texts, Victorian literature, medieval manuscripts—enable investigation of psychological constructs across centuries [3].

Validation differs fundamentally from contemporary human research. We cannot compare LLM outputs to what Confucius actually thought but can validate against historical records, textual consistency, and theoretical coherence [4]. The research question itself shifts: not "What did historical figures think?" but "Do contemporary psychological constructs show temporal stability or cultural specificity?" If a moral judgment framework calibrated on 21st-century Americans fails when applied to Song Dynasty texts, this suggests the framework lacks generalizability—informing contemporary cross-cultural research design.

Historical LLMs add temporal dimensions of variation that complement rather than substitute for spatial diversity in contemporary samples. Crockett and Messeri's DEAD (Decontextualized, Engineered, Anonymized, Disembodied) critique deflates here. We study texts produced in specific historical and cultural contexts, not abstract cognition from lived experience [5].

**Efficient hypothesis refinement before expensive validation**

Cross-cultural data collection requires international collaborators, instrument translation, cultural adaptation, and local ethics approval—processes taking years and substantial funding [6]. Poorly designed instruments waste these resources. LLMs enable rapid prototyping: testing whether experimental manipulations translate across linguistic contexts, identifying ambiguous items likely to produce measurement non-equivalence, and flagging cultural blind spots before committing to international data collection.

This approach makes subsequent human validation more efficient without replacing it [7,8]. An unexpected LLM response is diagnostic, not validating. If an LLM trained on Chinese corpora responds incoherently to a moral dilemma translated from English, this signals the scenario may not translate conceptually—warranting revision before recruiting Chinese participants.

Researchers still conduct the cross-cultural study; LLM piloting reduces the probability of expensive failures.

The economic barrier matters. When piloting costs drop from tens of thousands of dollars for international samples to hundreds for LLM testing, more groups can design truly diverse studies. This becomes surrogate creation only if researchers publish findings based solely on LLM performance without plans for human validation, or if LLM use makes them less likely to recruit diverse participants.

## Building capacity for real-world diversity engagement

Ethical and practical constraints limit exposure to diverse clinical populations during training. A recent study developed PATIENT-Ψ to simulate varied CBT patient presentations—different cultural backgrounds, socioeconomic contexts, presenting problems—for therapist training. Validation focuses on skill acquisition rather than absolute population accuracy [9]. Trainees must learn to recognize diverse presentations and select culturally appropriate interventions. The LLM need not perfectly represent any demographic group; it must provide sufficient variability for trainees to develop adaptive clinical reasoning.

Training with simulated patients prepares clinicians to work competently with real clients from marginalized communities but does not substitute for supervised practice with actual clients or ongoing community engagement [10]. The validation metric—whether trainees subsequently demonstrate better outcomes with diverse real patients—determines whether this application lowers barriers or creates the illusion of generalization.

## Conclusion

Crockett and Messeri's critique is essential, but it conflates distinct uses of LLMs. The persistent WEIRD dominance reflects genuine structural barriers—temporal inaccessibility, resource demands, institutional obstacles. These barriers neither excuse methodological laziness nor disappear through critique alone.

LLMs enhance generalizability when they expand phenomena inaccessible to human recruitment (historical populations, extreme contexts) or make human validation more efficient (instrument piloting, training for diversity engagement). Validation relies on theoretical coherence, historical records, and contemporary human benchmarks when appropriate.

LLMs threaten generalizability when they substitute for feasible human recruitment—demographic prompting to make claims about marginalized groups without community engagement, publishing findings based solely on LLM performance, or enabling researchers to avoid culturally grounded research practices. Clear standards distinguishing tools from surrogates will determine whether LLMs expand or narrow cognitive science's scope. The technology is agnostic—the research practices surrounding it are not.

## Correspondence

Zhicheng Lin, Department of Psychology, Yonsei University, Seoul, 03722, Republic of Korea (zhichenglin@gmail.com; X: @ZLinPsy)

**Acknowledgments**
The writing was supported by the National Key R&D Program of China STI2030 Major Projects (2021ZD0204200). The funder had no role in the decision to publish or in the preparation of the manuscript. I used Claude Sonnet 4.5 and Gemini 2.5 Pro for proofreading the manuscript, following the prompts described at https://www.nature.com/articles/s41551-024-01185-8.

## Declaration of interests

None declared by the author.

## References

1. Lin, Z. (2025) Six fallacies in substituting large language models for human participants. *Adv. Meth. Pract. Psychol. Sci.* 8, 25152459251357566. 10.1177/25152459251357566
2. Crockett, M.J. and Messeri, L. (2025) AI Surrogates and illusions of generalizability in cognitive science. *Trends Cogn. Sci.* 10.1016/j.tics.2025.09.012
3. Chen, Y. *et al.* (2024). Surveying the dead minds: Historical-psychological text analysis with contextualized construct representation (CCR) for classical Chinese. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024). Association for Computational Linguistics
4. Lin, Z. (2025) Large language models as psychological simulators: A methodological guide. *arXiv:2506.16702*. 10.48550/arXiv.2506.16702
5. Varnum, M.E.W. *et al.* (2024) Large Language Models based on historical text could offer informative tools for behavioral science. *Proc. Natl. Acad. Sci. U. S. A.* 121, e2407639121. 10.1073/pnas.2407639121
6. Gjersing, L. *et al.* (2010) Cross-cultural adaptation of research instruments: Language, setting, time and statistical considerations. *BMC Med. Res. Methodol.* 10, 13. 10.1186/1471-2288-10-13
7. Adhikari, D.M. *et al.* (2025). Exploring LLMs for automated generation and adaptation of questionnaires. Proceedings of the 7th ACM Conference on Conversational User Interfaces. Association for Computing Machinery
8. Lu, S.-C. *et al.* (2025) Can machine translation match human expertise? Quantifying the performance of large language models in the translation of patient-reported outcome measures (PROMs). *Journal of Patient-Reported Outcomes* 9, 94. 10.1186/s41687-025-00926-w
9. Wang, R. *et al.* (2024). PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics
10. Atzor, M.-C. *et al.* (2024) Effectiveness of internet-based training on psychotherapists' transcultural competence: A randomized controlled trial. *Journal of Cross-Cultural Psychology* 55, 260-277. 10.1177/00220221231221095