Draft references for Baker, Ryan S.; Barany, Amanda (forthcoming) *Artificial Intelligence in Qualitative Research: New Possibilities.* Philadelphia, PA: Vassant Press.

Find the latest chapters online at
https://learninganalytics.upenn.edu/book/ai-qualitative-research.html

---

References
Version 1.0, 31 March 2026

AI Notkilleveryoneism Memes (2025) An engineer showed Gemini what another AI said about its code. https://x.com/AISafetyMemes/status/2000620127054598508

AI Weekly (2025) Palmer Lucky Explains His ChatGPT Hack to Get It to Do ANYTHING. https://www.youtube.com/shorts/qS4S_-p-zso

Andriushchenko, M., Croce, F., & Flammarion, N. Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks. (2023) In *The Thirteenth International Conference on Learning Representations*.

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., ... & Baldwin, B. (2024). Non-determinism of" deterministic" llm settings. *arXiv preprint arXiv:2408.04667*.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Baudrillard, J. (1981). *Simulacra and simulation*. University of Michigan press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Bennett, J. (2010). *Vibrant matter: A political ecology of things*. Duke University Press.

Blair, A. (2025) Gemini 3 is Evaluation-Paranoid and Contaminated. https://www.lesswrong.com/posts/8uKQyjrAgCcWpfmcs/gemini-3-is-evaluation-paranoid-and-contaminated?utm_source=chatgpt.com

Booth, R (2026) From ‘nerdy’ Gemini to ‘edgy’ Grok: how developers are shaping AI behaviours. *The Guardian,* 4 February 2026,

https://www.theguardian.com/technology/2026/feb/03/gemini-grok-chatgpt-claude-qwen-ai-chatbots-identity-crisis

Bostrom, N. (2025) *Superintelligence: Paths, Dangers, Strategies.* Oxford, UK: Oxford University Press.

Buckingham Shum, S. (2024) *Qreframer: a chatbot prompt that reveals your assumptions*. https://oercommons.org/courseware/lesson/114039/overview

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., … Erlingsson, Ú. (2021). *Extracting training data from large language models*. In **Proceedings of the 30th USENIX Security Symposium (USENIX Security '21)** (pp. 2633–2650).

Case, N. (2018). How to become a centaur. *Journal of Design and Science*, *3*(5).

Chen, L., Zaharia, M., & Zou, J. (2024). *How is ChatGPT's behavior changing over time? Harvard Data Science Review, 6*(2).

Chiang, T. (2023, February 9). *ChatGPT is a blurry JPEG of the web*. **The New Yorker**. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

Clark, H. H., & Brown, K. (2006). Context and common ground. *Concise encyclopedia of philosophy of language and linguistics*, 85-87.

Cleo Nardo (2023) The Waluigi Effect (mega-post). https://www.lesswrong.com/posts/D7PumeYTDPfBTp3i7/the-waluigi-effect-mega-post

Clifford, J. (1986). Introduction: Partial truths. In J. Clifford & G. E. Marcus (Eds.), *Writing culture: The poetics and politics of ethnography* (pp. 1–26). University of California Press.

Cummins, J., Elson, M., & Hussey, I. (2025). Cognitive dissonance in large language models is neither cognitive nor dissonant. *Proceedings of the National Academy of Sciences*, *122*(35), e2517912122.

DeLanda, M. (2006). *A new philosophy of society: Assemblage theory and social complexity*. Continuum.

Deleuze, G., & Guattari, F. (1980). *A thousand plateaus: Capitalism and schizophrenia* (B. Massumi, Trans.). University of Minnesota Press.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard business school technology & operations mgt. Unit working paper*, (24-013).

Derrida, J. (1967). *Of grammatology*. Baltimore, MD: Johns Hopkins University Press.

Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., ... & Koyejo, S. (2025, October). Syceval: Evaluating llm sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Vol. 8, No. 1, pp. 893-900).

Fine, M. (1994). *Working the hyphens: Reinventing self and other in qualitative research*. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 70–82). Sage.

Freund, Y. (2013) Artificial intelligence vs intelligence amplification. California Institute for Telecommunications and Information Technology. https://www.youtube.com/watch?v=sEGdszE86bY

Friese, S., Nguyen-Trung, K., Powell, S., & Morgan, D. (2025). Beyond Binary Positions: Making Space for Critical and Reflexive GenAI Integration in Qualitative Research. Manuscript under review.

Fu, J., Zhao, G., Deng, Y., Mi, Y., & Qian, X. (2024). *Learning to paraphrase for alignment with LLM preference*. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 2394–2407). Association for Computational Linguistics. https://aclanthology.org/2024.findings-emnlp.134/

Goffman, E. (1956). The Presentation of Self in Everyday Life. Doubleday

Goujon, V., & Ricci, D. (2024). “Shoggoth with Smiley Face”: Knowing-how and letting-know by analogy in artificial intelligence research. *Hybrid. Revue des arts et médiations humaines*, (12).

Heidegger, M. (1927) *Being and Time.* London, UK: SCM Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences*, *33*(2-3), 61-83.

Herr, K., & Anderson, G. L. (2005). *The Action Research Dissertation: A Guide for Students and Faculty*. SAGE.

Heston, T. F., & Gillette, J. (2025). Large Language Models Demonstrate Distinct Personality Profiles. *Cureus*, *17*(5).

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1-55.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA USA: MIT press.

Janus (2022) Simulators. https://www.google.com/url?q=https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx&sa=D&source=docs&ust=1771817617517855&usg=AOvVaw1Cf-maJmW4bHzqCy_J-Skh

Jowsey, T., Braun, V., Clarke, V., Lupton, D., & Fine, M. (2025). We reject the use of generative artificial intelligence for reflexive qualitative research. *Qualitative Inquiry*, 10778004251401851.

Krakowski, S., Luger, J., & Raisch, S. (2023). Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal*, *44*(6), 1425-1452.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive science*, *4*(2), 195-208.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge university press.

Li, K., Liu, T., Bashkansky, N., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2024). Measuring and controlling instruction (in) stability in language model dialogs. *arXiv preprint arXiv:2402.10962*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024a). Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, *12*, 157-173.

Lopez-Fierro, S., & Nguyen, H. (2024). Making Human-AI contributions transparent in qualitative coding. In *Proceedings of the 17th International Conference on Computer-Supported Collaborative Learning-CSCL 2024, pp. 3-10*. International Society of the Learning Sciences.

Meherab, M. M., Billah, M. M., Rahman, K. S., Sharmin, L., Islam, T., Mahmud, Z. Z., ... & Bhuiyan, T. (2026). Advancing NLP Equity: A Secondary Benchmark Evaluation of Multilingual Language Models for Underrepresented Languages. In *Second Workshop on Language Models for Underserved Communities (LM4UC)*.

Mowshowitz, Z. (2025). *OpenAI model differentiation 101*. Don't Worry About the Vase (Substack). https://thezvi.substack.com/p/openai-model-differentiation-101

Mwatkins, Rumbelow, J. (2023) SolidGoldMagikarp II: technical details and more recent findings. Available online at https://www.lesswrong.com/posts/Ya9LzwEbfaAMY8ABo/solidgoldmagikarp-ii-technical-details-and-more-recent

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435–450

Newell, A., & Simon, H. A. (1961). Computer Simulation of Human Thinking: A theory of problem solving expressed as a computer program permits simulation of thinking processes. *Science*, *134*(3495), 2011-2017.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., ... & Olah, C. (2022). In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative agents: Interactive simulacra of human behavior*. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)* (pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). SAGE.

Perrigo, B. (2023). *The new AI-powered Bing is threatening users. That's no laughing matter*. *TIME*. https://time.com/6256529/bing-openai-chatgpt-danger-alignment/

Pringle, E. (2023). *Microsoft's ChatGPT-powered Bing is becoming a pushy pick-up artist that wants you to leave your partner: 'You're married, but you're not happy'*. *Fortune*. https://fortune.com/2023/02/17/microsoft-chatgpt-bing-romantic-love/

Rettberg, J. W., & Wigers, H. (2025). AI-generated stories favour stability over change: homogeneity and cultural stereotyping in narratives generated by gpt-4o-mini. *arXiv preprint arXiv:2507.22445*.

Robertson, A. (2024, February 21). *Google apologizes for "missing the mark" after Gemini generated racially diverse Nazis*. **The Verge**. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

Roddenberry, G. (1964) *Star Trek.* Los Angeles, CA: NBC.

Sartre, J-P. (1946) Existentialism is Humanism. New Haven, CT USA: Yale University Press.

Shaffer, D. W., & Ruis, A. R. (2021, January). How we code. In *International Conference on Quantitative Ethnography* (pp. 62-77). Cham: Springer International Publishing.

St. Amant, R. (2014) The Use of Tools. North Carolina State University Department of Computer Science Technical Report TR-2014-4.

Stephenson, N. (1994). *Snow crash*. Penguin UK.

Tang, K. S. (2025). AI-textuality: Expanding intertextuality to theorize human-AI interaction with generative artificial intelligence. *Applied Linguistics*, article amaf016.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). *Cultural bias and cultural alignment of large language models*. **PNAS Nexus, 3**(9), pgae346.

Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, *36*, 74952-74965.

Wang, A., Morgenstern, J., & Dickerson, J. P. (2024). *Large language models that replace human participants can harmfully misportray and flatten identity groups*. arXiv:2402.01908.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022a) Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022b). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Wickelgren, W. A. (1995). *How to solve mathematical problems*. Courier Corporation.

Xu, Z., Liu, Y., Deng, G., Li, Y., & Picek, S. (2024, August). A comprehensive study of jailbreak attack versus defense for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 7432-7449).

Yang, Z., Zhang, Y., Liu, T., Yang, J., Lin, J., Zhou, C., & Sui, Z. (2024). *Can large language models always solve easy problems if they can solve harder ones?* In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 1531–1555). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.92

Yuli_Ban (2023) Pink Shoggoths: What does alignment look like in practice? https://www.lesswrong.com/posts/9y5RpyyFJX4GaqPLC/pink-shoggoths-what-does-alignment-look-like-in-practice?utm_source=chatgpt.com

Zhang, A., Tanzer, G., Marcheret, E., & Mortensen, D. R. (2024). Shortcomings of LLMs for Low-Resource Translation: Retrieval and Understanding are Both the Problem. *arXiv preprint arXiv:2406.15625*.