Computational Hermeneutics: Evaluating Generative AI as a Cultural Technology

Cody Kommers
The Alan Turing Institute
ckommers@turing.ac.uk

Ruth Ahnert
Queen Mary
University of London

Maria Antoniak
University of Colorado

Emmanouil Benetos
Queen Mary
University of London

Steve Benford
University of Nottingham

Mercedes Bunz
King's College
London

Baptiste Caramiaux
Sorbonne Université

Shauna Concannon
Durham University

Martin Disley
University of Edinburgh

James Dobson
Dartmouth College

Yali Du
King's College London

Edgar Duéñez-Guzmán
Gibran AI

Kerry Francksen
University of Coventry

Evelyn Gius
Technische Universität Darmstadt

Jonathan Gray
King's College London

Ryan Heuser
University of Cambridge

Sarah Immel
University of Edinburgh

Richard Jean So
McGill University

Sang Leigh
Cornell University

Dalaki Livingston
University of Utah

Hoyt Long
University of Chicago

Meredith Martin
Princeton University

Georgia Meyer
London School of Economics

Daniela Mihai
University of Southampton

Ashley Noel-Hirst
Queen Mary
University of London

Kirsten Ostherr
Rice University

Deven Parker
University of Glasgow

Yipeng Qin
Cardiff University

Jessica Ratcliff
Cornell University

Emily Robinson
University of Exeter

Karina Rodriguez
University of Brighton

Adam Sobey
The Alan Turing Institute &
University of Southampton

Ted Underwood
University of Illinois
Urbana-Champaign

Aditya Vashistha
Cornell University

Matthew Wilkens
Cornell University

Youyou Wu
University College
London

Zheng Yuan
University of Sheffield

Drew Hemment
The Alan Turing Institute &
University of Edinburgh

Abstract

Generative AI (GenAI) systems are increasingly recognized as cultural technologies, yet current evaluation frameworks often treat culture as a variable to be measured rather than fundamental to the system's operation. Drawing on hermeneutic theory from the humanities, we argue that GenAI systems function as "context machines" that must inherently address three interpretive challenges: situatedness (meaning only emerges in context), plurality (multiple valid interpretations coexist), and ambiguity (interpretations naturally conflict). We present computational hermeneutics as an emerging framework offering an interpretive account of what GenAI systems do, and how they might do it better. We offer three principles for hermeneutic evaluation—that benchmarks should be iterative, not one-off; include people, not just machines; and measure cultural context, not just model output. This perspective offers a nascent paradigm for designing and evaluating contemporary AI systems: shifting from standardized questions about accuracy to contextual ones about meaning.

1 Introduction

Generative AI (GenAI) systems are cultural and social technologies [3, 31, 46, 82]. While this position is increasingly accepted as orthodoxy within the field, it can rely on a limited definition of culture. In practice culture is often treated as a secondary consideration, like a coat of paint or dash of seasoning that modifies the more "fundamental" aspects of the model: for example, as a bias to debug [8, 88], a constraint for generalizing from one context to another [11], a parameter in an ethical dilemma [19], or a source of variability in user preferences [36]. These approaches operationalize culture as a variable to be measured—often implying that it is an optional parameter to include in model evaluation, rather than a foundational aspect of the model's functioning.

However, most large models (especially those from industry) are not specialized systems designed to solve targeted, well-defined tasks. They are marketed as general systems built to generate a variety of cultural artifacts in a vast space of possible contexts. Cultural considerations are inextricable both from how these models are developed and from the open-ended, dialogic interfaces in which they are used. It is therefore crucial that we ask: How can we most effectively evaluate GenAI as a cultural technology?

In this Perspective, we offer an account of culture informed by the humanities. We argue that evaluation methods in AI often overlook an important conception of culture: not as a variable to be measured, but as a dynamic, contested space where social meaning is made [37, 38, 46]. This way of looking at culture challenges a core assumption in standard practices for AI benchmarking—that model performance is best understood through universal, standardized tasks with convergent solutions or goals [70]. While this approach works for well-defined tasks where "success" can be codified into a single, unique interpretation, culture is not this kind of task.

To illustrate the challenge of evaluating cultural outputs, consider the act of writing a letter, painting a portrait, composing a song, cooking a meal, penning a journal entry, or even talking with a friend. While it is possible to assign a quantitative score to these outputs to describe how well the task was performed, that approach can miss the point of these activities in crucial ways. For example, reducing cultural activities to a set of proxy variables can trivialize them [100], while scalable "thin" metrics are often insufficient to capture key aspects of what makes them meaningful [49]. The structure of these tasks is such that the primary question is not about assessing how closely they cleave to a canonical ground truth. Rather it is about arbitrating among multiple, possibly conflicting, interpretations of their meaning within a specific frame of reference. This requires us to think about culture as an intrinsically different kind of "task" from those by which a model's performance has traditionally been judged.

Our position is that, as AI systems are increasingly deployed to (co-)produce cultural outputs, it is imperative that our methods of evaluation reflect the interpretive dimensions needed to characterize

This is a preprint. Submitted August 2025. Under review.

them more fully. To address this, we introduce hermeneutics—a core tradition in the humanities concerned with the theory and practice of interpretation—as a theoretical foundation for understanding and evaluating GenAI systems [64, 71, 76]. Having grappled with these questions for decades, if not centuries, the conceptual infrastructure of the humanities (via hermeneutics) can help articulate the grounds on which a given interpretation can be considered legitimate. Providing such an account of the interpretive nature of GenAI systems is a crucial step towards improving the way we design and evaluate them.

Thus, we present computational hermeneutics as an emerging framework offering an interpretive approach to the evaluation of GenAI systems. We argue that GenAI can, and should, be understood as "doing" interpretation in ways that reflect the entanglement of culture in their input, processes, and outputs. We offer three hermeneutic challenges that are inherent to such interpretive processes: situatedness, plurality, and ambiguity. Each of these already exists in one form or another in contemporary AI [1, 51, 82]; we further this existing work by suggesting how these challenges can be brought together within a hermeneutic frame. Finally, we offer three principles for developing hermeneutic methods of evaluating GenAI: that benchmarks should be iterative, not one-off; include people, not just machines; and measure cultural context, not just model output.

2 Computational Hermeneutics

Interpretation is the methodological bedrock of the humanities [37]. Generally speaking, what humanists do when studying cultural artifacts—whether a novel, historical event, or painting—is to construct an interpretation: an analysis of that artifact's meaning within its social or historical context. But this approach comes with an inherent challenge. How do we know whether a given interpretation is a good one? Hermeneutics is the method, justification, or separate interpretive process which gives credence or legitimacy to the original interpretation [12, 74]. This concept is foundational across many disciplines and practices, from legal and literary studies [53, 86] to debates in philosophy and aesthetics [77, 81]. It arises, in one form or another, whenever scholars confront epistemological problems of meaning.

A core concept within this tradition is the "hermeneutic circle" [22, 34, 40, 79]. This describes the interpretation of an artifact as an iterative process between understanding the meaning of a specific part of the artifact and the meaning of the artifact as a whole. For example, one could iteratively analyze the imagery depicted in a given line or stanza of a poem, then update one's conception about what the poem means in general—each time using the updated general theory to analyze the specific line, and vice versa. While the term is varied in its usage, what it typically means to analyze something hermeneutically is to engage in (and provide an account of) this iterative process of interpretation.

As applied to contemporary AI, we offer a notion of computational hermeneutics in two senses. The first sense is that AI models are fundamentally interpretive in a way that makes hermeneutic problems unavoidable; these challenges are intrinsic to GenAI's flexible production of sophisticated cultural artifacts such as texts and images. To categorize their outputs as binary "right" or "wrong" responses presents a similar profile of problems as asking whether *Anna Karenina* is a superior novel to *Jane Eyre*, whether the spiritual life prescribed in Laozi's *Tao Te Ching* is the right one, or whether Andy Warhol's soup can paintings were a critique, rather than a celebration, of American consumerism. Judgments on these matters are possible, but they depend crucially on the underlying assumptions of one's interpretive processes.

The second sense is that interpretive evaluation requires us to look at both specific and general aspects of the models, in the tradition of the hermeneutic circle. These models have both a general architecture (e.g., pre-training, vector representations, fine-tuning), as well as specific dialogic interactions with human users (e.g., context windows, prompts). We must look at both the system-level generalizations and context-specific outputs in interpreting the outputs of these models. Roughly speaking, partial analysis maps onto the "Chat" in ChatGPT, while holistic analysis maps onto the "GPT." Though these separate parts are interrelated, it is crucial to draw distinctions required for the evaluation of each on their own terms [23, 75].

2.1 Hermeneutic Challenges for AI

With this framing in mind, we present three hermeneutic challenges for GenAI: situatedness, plurality, and ambiguity. Each of these challenges take aspects of a model that may seem arbitrary, peripheral, or in need of optimization—and recenters those apparently accidental features as significant choices worthy of theoretical reflection. We take addressing these challenges to be the main difference between accounting for culture as a variable versus culture as a site of social meaning-making.

2.1.1 Situatedness: Meaning only emerges in context.

A core principle across many (if not all) of the humanities is that context is key. What this expresses, typically, is that to interpret the meaning of a cultural artifact, one must look at the historical or social context in which it has been made, used, or perceived [34]. For example, a contemporary reader of *Huckleberry Finn* will inevitably have a different relation to the text from a reader in the 19th century America of the book's original publication. When the frame of reference shifts, so does the meaning. Cultural products are always generated within the bounds of a particular historical, cultural, or communicative context. This is the "situatedness" of meaning: an interpretation always takes a particular point of view, even if that perspective is only stated implicitly.

It can be easy to overlook this in contemporary AI interfaces, which often present the model as speaking from a god's eye point of view—that of the disembodied model which has seen, read, and synthesized more information than any one human ever could [41]. No such epistemically totalitarian "view from nowhere" exists in any legitimate sense [39]. Within a hermeneutic frame, the point is not to build and evaluate models that aim to achieve this universal, monolithic perspective. Rather it is for the specific perspective being offered to be clearly identified and understood as just that: a specific perspective.

2.1.2 Plurality: One person's bias is another person's values.

Interpretation is inherently plural, because different communities rely on distinct frameworks for making sense of the world. What appears as meaningful artistic expression to one group may seem inappropriate or offensive to another; what counts as authoritative fact in one tradition may be dismissed as unsubstantiated assertion in another. As is widely held in the humanities, multiple valid interpretations can coexist without requiring resolution into a single "correct" reading. Any AI model intended for use in different cultural contexts must grapple with the observation that what looks like arbitrary cultural bias from one perspective is often the same thing that gives a sense of meaning and value in another.

AI systems face this challenge directly because they serve users with distinct values while being trained on materials whose authors often disagree. Generative models are therefore both one and many: reflecting specific curatorial decisions, but also containing contradictory voices [21, 80, 94]. Recent work on pluralistic, thick, or full-stack alignment recognizes that human values naturally conflict and advocates for systems that can accommodate this diversity [51, 56, 82]. However, while pluralistic alignment focuses on adjusting model behavior to reflect different values, the deeper challenge lies in how we evaluate such systems. Standard evaluation frameworks assume convergent solutions—that there is a standard candle against which model performance can be definitively compared. Cultural tasks, by contrast, do not converge to single solutions: success cannot be determined by proximity to a ground truth but must account for the legitimacy of multiple interpretations within their respective contexts. This requires fundamentally rethinking evaluation from measuring accuracy to assessing appropriateness across different cultural frameworks.

2.1.3 Ambiguity: Interpretations naturally conflict.

In hermeneutics, meaning is not something that exists as a fixed property of a text or cultural artifact, inertly awaiting discovery. Rather, meaning emerges through what Gadamer calls the "fusion of horizons"—the dynamic interaction between the interpreter's background and the artifact being interpreted [34]. This process is intrinsically ambiguous. The space of possible mappings between potentially relevant features of the interpreter's background and the artifact is combinatorially large, and therefore a definitive interpretation is not computationally tractable. To offer a particular kind of interpretation (e.g., feminist, post-colonial, techno-optimist) is to ease this intractability by specifying an a priori constraint on which features to consider. More generally, Gadamer emphasizes the role

of "play" in interpretation—that creative, open-ended consideration of tensions between different meanings offers a way of exploring this space of interpretive possibilities. It is therefore crucial that ambiguity be maintained in articulating this interpretive space, rather than being flattened into a specific mode of interpretation.

Ambiguity has long been of interest in AI, often with the goal of resolving it [66]. Semantic disambiguation tasks, for instance, aim to determine which meaning of a polysemous word is intended in a given context—clarifying whether "light" is used to signify illumination or weight. Such tasks are crucial for many applications, but they represent only one way of engaging with ambiguity. When AI systems generate cultural outputs—whether composing poetry, engaging in dialogue, or creating visual art—the goal is not necessarily to eliminate semantic uncertainty but to work productively within it [35]. A poem that resolves all its ambiguities loses much of its interpretive richness; a conversation that admits only one reading of each utterance becomes sterile [27]. However, current evaluation frameworks often treat this ambiguity as noise to be minimized rather than a generative resource [97]. While semantic disambiguation tasks can be useful, elimination of ambiguity is not the only—or even the primary—goal when it comes to cultural outputs. Instead, evaluation should assess how well systems navigate ambiguity productively, maintaining the interpretive flexibility that enables meaningful cultural engagement across diverse contexts [52, 93].

3 Generative AI systems as "Context Machines"

In this section, we argue that GenAI systems "do" interpretation as a fundamental capacity [24]—and therefore evaluation of their performance is subject to the three hermeneutic challenges described above. These interpretive processes take place both internally within a model, as well as dialogically in their interactions with people. Providing a more comprehensive account of the interpretive nature of these systems is a crucial step towards improving the way we design and evaluate them.

We posit that GenAI systems can be broadly understood as "context machines." At core, GenAI systems are designed to answer the question: given the current context, what is the next relevant token, pixel, or other value? This ability to consolidate a broader set of contextual cues into a unified representation is supported by a variety of architectural features—but most notably by vector space embeddings [29, 50, 85]. Such embeddings are a means of encoding highly sophisticated co-occurrence statistics [90]. In language models, they are learned by poring over vast corpora of text [62, 69]. In vision models, vectors of pixel values are often encoded as feature maps capturing edges, textures, and semantic patterns [4, 61]. Decoding these embeddings is also an interpretive act. This process is often probabilistic, accommodating a plurality of possible interpretations [43, 98]. Informally, these vectors are designed to capture the "meaning" of words or images; more concretely, they are a highly nuanced way of describing the context in which a word is likely to occur.

Generative models work as well as they do because (as is a common refrain in the humanities) context matters—so much so that if you get it right, a lot of other important things follow. Vector space embeddings are therefore subject to a similar question as humanistic inquiry: How do we know whether a given interpretation, as encoded by an embedding, is a good one? Accordingly, GenAI systems are faced with the three hermeneutic challenges described above: the outputs of these systems are situated (the "meaning" of one token is defined relationally within the context of other tokens); plural (there are multiple legitimate interpretations of what counts as the next most likely token); and ambiguous (the probabilistic decoding process maintains rather than resolves semantic uncertainty).

Our position is that Generative AI systems both "do" interpretation, and that they can do it better. For example, the self-attention mechanism of the transformer architecture can be read as a way of relating partial and holistic interpretations [92]. It allows the model to iteratively update its understanding of individual tokens based on their relationship to the broader sequence, and vice versa—in other words, the hermeneutic circle in action.

3.1 AI systems don't just "read in" context; they help create it.

GenAI models do not just perform interpretation in isolation; they also co-construct interpretations in collaboration with humans [32]. A hermeneutic perspective on AI is not just about building systems that can interpret like humans, as a substitute or proxy for human expertise. Rather it is about recognizing how interpretation itself emerges through interaction between humans and machines. In

this view, interpretive capacity arises not only within the model but through the design of interactions and interfaces that frame it.

The effects of this collaboration are bidirectional. From human to machine, people decide what data the systems are trained on [21]; formulate objective functions that reflect a specific set of goals, values, and assumptions [51]; fine-tune system behavior through mechanisms like reinforcement learning from human feedback [68]; and "engineer" prompts in order to elicit certain kinds of responses [16]. At multiple layers of the system, human annotators—who can themselves offer conflicting interpretations [33]—can provide feedback on ambiguous cases, rank responses, or supply preference scores, effectively staging a dialogue where the AI's provisional interpretations can be contested and refined.

From machine to human, AI systems affect important mental capacities like metacognition [87]; elicit different assumptions about relational norms (e.g., AI as assistant vs therapist [26]); act as thought-partners, for example by summarizing documents people would otherwise have to read—or skim—in full [18]; shape human responses by explaining their own decisions [25]; and enable novel kinds of experience, such as certain creative practices [13, 42, 65]. Together, humans and GenAI systems form an interpretive feedback loop. Far from a separate isolated entity that the system merely "reads in," AI systems can exert a direct influence on the cultural context in which they operate.

4 Operationalizing Hermeneutics in AI

Typically, AI benchmarking assumes universal, standardized tasks with convergent solutions [15, 28, 70]—an approach fundamentally at odds with a hermeneutic perspective on culture. While benchmarks are key drivers of progress in AI, they often do not offer especially strong standards for what they purport to measure [45, 59, 73, 78]. Furthermore, the implicit goal of benchmarking is often not to develop stronger metrics for specialized cases (though see [17, 91]) but something more like one-task-suite-to-rule-them-all, a comprehensive assessment that would give an unequivocal, decisive answer to the question of which model is better at what [2, 30, 47, 70, 83].

Our hermeneutic framing challenges this paradigm by reimagining the kinds of questions that can be asked with AI benchmarks: shifting from standardized questions about accuracy to contextual ones about meaning. From this perspective, no such comprehensive task suite can be developed, because the "task" of creating cultural outputs means too many different things in too many different contexts. Attempts to standardize cultural production into a comprehensive assessment often seek to scrub away this context; we advocate that such context must be embraced. We offer three ways of making AI benchmarks that better reflect a hermeneutic lens on culture—by making them iterative, not just one-off; including people, not just machines; and measuring cultural context, not just model output.

4.1 Benchmarks should be iterative, not just one-off.

The hermeneutic circle suggests that interpretation depends on an iterative process between part and whole. By contrast, benchmarks typically apply a score—often scalar values such as accuracy, precision, recall, F1, or BLEU scores [15, 28]—to quantify the model's performance in a given domain. Hermeneutics benchmarking suggests two modifications that can be made to this approach.

First, evaluation is both limited and unreliable when it scores performance based on a single prompt [63]. By contrast, cultural outputs are always part of an evolving conversation, whether a literal dialogue or as a part of a broader evolutionary process [9]. Evaluation should accordingly be iterative, unfolding over multiple prompts or exchanges that reflect the evolving interpretive context.

Second, evaluation must take into account both the model as whole and the specific dialogic frame in which a given output is elicited. For example, the focus of benchmarking on aggregate metrics indicating average performance rather than instance-by-instance evaluations limits generalizability [10]. Overall, hermeneutic evaluations should seek to iteratively assess both the model's holistic capabilities, as well as its behavior in specific circumstances.

4.2 Benchmarks should include people, not just machines.

The interpretive processes underlying GenAI are inextricably bound up in collaboration with the people using them [60]. Benchmarks should therefore not just consider AI performance in isolation

but ought to also measure the effects of different interactive configurations. For example, current approaches to the assessment of creativity in narrative generation range from automated metrics to expert human judgment [7, 14, 58]; but these often treat creativity as a model property rather than a relational phenomenon.

A hermeneutic approach would evaluate how human-AI collaboration produces interpretations, examining not just outputs but the interpretive dialogue that generates them. This builds on a wide range of efforts in AI evaluation which increasingly recognize that benchmarks cannot be divorced from their communicative context [17, 20, 95, 96]. Overall, hermeneutic evaluation requires benchmarks that assess interactivity rather than isolated performance, examining not just outputs but the interpretive dialogue that generates them.

4.3 Benchmarks should measure cultural context, not just model output.

Individual interpretations of meaning depend on cultural context [48]—yet standard evaluation practices treat context as secondary to model performance metrics. Thin signals of like/dislike, positive/negative, or use/disuse cannot provide this contextual grounding [49]. Rather, we need hermeneutic approaches for putting contextual use cases on equal footing with general model capacities.

Partially, this is simply a suggestion to evaluate AI in the context in which it will be used [1, 55, 57, 60, 89]. For example, frameworks like HELM recognize the need for contextually dependent approaches beyond accuracy [54]. This can help address issues with current benchmarks, such as failure to capture real-world utility [67], or by adapting general processes to better fit situational needs [84].

But more pointedly, digging deeper into contextualized scenarios allows us to probe different aspects of the model. Rather than asking whether a response is correct, hermeneutic evaluation can assess how and why a response achieves appropriateness within its specific cultural framework [5, 52]. Evaluation must treat cultural context not as a constraint on model performance, but as the medium through which such performance emerges.

5 Discussion

Computational hermeneutics represents a potential shift in how we conceptualize GenAI systems. Rather than treating culture as a variable to be controlled or optimized away, we propose recognizing it as a foundational aspect of how these systems operate. This reframing transforms GenAI from answer-generating machines into interpretive partners—systems designed to engage with the situatedness, plurality, and ambiguity that characterize individual and collective human meaning-making.

It is widely acknowledged that better benchmarks are needed to support ethical and effective development of AI [6, 72, 73, 78, 99]. One possible systemic cause of this is proxy failure [44]: that the field's monocultural overreliance on standardized performance metrics is inadequate to capture the kinds of things we really want AI to do [47, 49, 100]. We offer the emerging framework of computational hermeneutics as a potential means of rethinking how we evaluate AI from the ground up—as a set of technologies that does not just participate in culture by accident, but as systems which fundamentally shape, and are shaped by, cultural meaning.

References

[1] Canfer Akbulut, Kevin Robinson, Maribeth Rauh, Isabela Albuquerque, Olivia Wiles, Laura Weidinger, Verena Rieser, Yana Hasson, Nahema Marchal, Iason Gabriel, et al. Century: A framework and dataset for evaluating historical contextualisation of sensitive images. In *The Thirteenth International Conference on Learning Representations*, 2025.

[2] Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[4] Yoshua Bengio, Aaron C Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, 1(2665):2012, 2012.

[5] Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. Seegull multilingual: a dataset of geo-culturally situated stereotypes. *arXiv preprint arXiv:2403.05696*, 2024.

[6] Kathrin Blagec, Jakob Kraiger, Wolfgang Frühwirt, and Matthias Samwald. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *Journal of Biomedical Informatics*, 137:104274, 2023.

[7] Joanne Boisson, Zara Siddique, Hsuvas Borkakoty, Dimosthenis Antypas, Luis Espinosa Anke, and Jose Camacho-Collados. Automatic extraction of metaphoric analogies from literary texts: Task formulation, dataset construction, and evaluation. *arXiv preprint arXiv:2412.15375*, 2024.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29, 2016.

[9] Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L Griffiths, Joseph Henrich, et al. Machine culture. *Nature Human Behaviour*, 7(11):1855–1868, 2023.

[10] Ryan Burnell, Wout Schellaert, John Burden, Tomer D Ullman, Fernando Martinez-Plumed, Joshua B Tenenbaum, Danaja Rutar, Lucy G Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, 2023.

[11] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 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, May 2023. Association for Computational Linguistics.

[12] John D Caputo. *Hermeneutics: Facts and interpretation in the age of information*. Penguin UK, 2018.

[13] Baptiste Caramiaux and Sarah Fdili Alaoui. "Explorers of unknown planets": Practices and politics of artificial intelligence in visual arts. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), November 2022.

[14] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? Large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[15] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), March 2024.

[16] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*, 2023.

[17] Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. A computational framework for behavioral assessment of LLM therapists. *arXiv preprint arXiv:2401.00820*, 2024.

[18] Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024.

[19] Julian De Freitas, Andrea Censi, Bryant Walker Smith, Luigi Di Lillo, Sam E Anthony, and Emilio Frazzoli. From driverless dilemmas to more practical commonsense tests for automated vehicles. *Proceedings of the National Academy of Sciences*, 118(11):e2010202118, 2021.

[20] Remi Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*, 2020.

[21] Meera A Desai, Irene V Pasquetto, Abigail Z Jacobs, and Dallas Card. An archival perspective on pretraining data. *Patterns*, 5(4), 2024.

[22] Wilhelm Dilthey. *Introduction to the human sciences*, volume 1. Princeton University Press, 1989.

[23] James E Dobson. *Critical digital humanities: The search for a methodology*. University of Illinois Press, 2019.

[24] James E Dobson. Vector hermeneutics: On the interpretation of vector space models of text. *Digital Scholarship in the Humanities*, 37(1):81–93, 2022.

[25] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[26] Brian D Earp, Sebastian Porsdam Mann, Mateo Aboy, Edmond Awad, Monika Betzler, Marietjie Botes, Rachel Calcott, Mina Caraccio, Nick Chater, Mark Coeckelbergh, et al. Relational norms for human-AI cooperation. *arXiv preprint arXiv:2502.12102*, 2025.

[27] William Empson. *Seven Types of Ambiguity*. 1930.

[28] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation. *arXiv preprint arXiv:2502.06559*, 2025.

[29] Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics.

[30] Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online, November 2020. Association for Computational Linguistics.

[31] Henry Farrell, Alison Gopnik, Cosma Shalizi, and James Evans. Large AI models are cultural and social technologies. *Science*, 387(6739):1153–1156, 2025.

[32] Christopher Frauenberger. Entanglement HCI the next wave? *ACM Trans. Comput.-Hum. Interact.*, 27(1), November 2019.

[33] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*, pages 1–28, 2024.

[34] Hans-Georg Gadamer. *Truth and method*. 1960.

[35] William W. Gaver, Jacob Beaver, and Steve Benford. Ambiguity as a resource for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, page 233–240, New York, NY, USA, 2003. Association for Computing Machinery.

[36] Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. How culture shapes what people want from AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[37] Clifford Geertz. *The interpretation of cultures*. Basic Books, 1973.

[38] S Hall. Representation: Cultural representations and signifying practices. *Culture*, 1997.

[39] Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599, 1988.

[40] Martin Heidegger. *Being and time*. 1927.

[41] Drew Hemment, Cody Kommers, and colleagues. Doing AI differently: Rethinking the foundations of AI via the humanities. Technical report, London: The Alan Turing Institute, 2025.

[42] Drew Hemment, Dave Murray-Rust, Vaishak Belle, Ruth Aylett, Matjaz Vidmar, and Frank Broz. Experiential AI: Between arts and explainable AI. *Leonardo*, 57(3):298–306, 2024.

[43] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R*, volume 103. Springer, 2013.

[44] Yohan J John, Leigh Caldwell, Dakota E McCoy, and Oliver Braganza. Dead rats, dopamine, performance metrics, and peacock tails: Proxy failure is an inherent risk in goal-oriented systems. *Behavioral and Brain Sciences*, 47:e67, 2024.

[45] Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *arXiv preprint arXiv:2407.01502*, 2024.

[46] Lauren Klein, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton, and David Mimno. Provocations from the humanities for generative AI research. *arXiv preprint arXiv:2502.19190*, 2025.

[47] Bernard J Koch and David Peterson. From protoscience to epistemic monoculture: How benchmarking set the stage for the deep learning revolution. *arXiv preprint arXiv:2404.06647*, 2024.

[48] Cody Kommers and Simon DeDeo. Sense-making, cultural scripts, and the inferential basis of meaningful experience. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

[49] Cody Kommers, Drew Hemment, Maria Antoniak, Joel Z Leibo, Hoyt Long, Emily Robinson, and Adam Sobey. Meaning is not a metric: Using LLMs to make cultural context legible at scale. *arXiv preprint arXiv:2505.23785*, 2025.

[50] Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019.

[51] Seth Lazar and Alondra Nelson. AI safety on whose terms?, 2023.

[52] Joel Z Leibo, Alexander Sasha Vezhnevets, Manfred Diaz, John P Agapiou, William A Cunningham, Peter Sunehag, Julia Haas, Raphael Koster, Edgar A Duéñez-Guzmán, William S Isaac, et al. A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv:2412.19010*, 2024.

[53] Sanford Levinson and Steven Mailloux. *Interpreting law and literature: A hermeneutic reader*. Northwestern University Press, 1988.

[54] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[55] Q Vera Liao and Ziang Xiao. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*, 2023.

[56] Ryan Lowe, Joe Edelman, Tan Zhi-Xuan, Oliver Klingefjord, Ellie Hain, Vincent Wang, Atrisha Sarkar, Michiel A Bakker, Fazl Barez, Matija Franklin, et al. Full-stack alignment: Co-aligning AI and institutions with thicker models of value. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025.

[57] Chaitanya Malaviya, Joseph Chee Chang, Dan Roth, Mohit Iyyer, Mark Yatskar, and Kyle Lo. Contextualized evaluations: Judging language model responses to underspecified queries. *arXiv preprint arXiv:2411.07237*, 2024.

[58] Guillermo Marco, Julio Gonzalo, and Víctor Fresno. The reader is the metric: How textual features and reader profiles explain conflicting evaluations of AI creative writing. *arXiv preprint arXiv:2506.03310*, 2025.

[59] Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Dan Xu, Paul Watters, and Malka N Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence*, 2025.

[60] Lisa Messeri and Molly J Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024.

[61] Daniela Mihai and Jonathon Hare. Learning to draw: Emergent communication through sketching. *Advances in Neural Information Processing Systems*, 34:7153–7166, 2021.

[62] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013.

[63] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

[64] John W Mohr, Robin Wagner-Pacifici, and Ronald L Breiger. Toward a computational hermeneutics. *Big Data & Society*, 2(2):2053951715613809, 2015.

[65] Tim Murray-Browne and Panagiotis Tigas. Emergent interfaces: Vague, complex, bespoke and embodied interaction between humans and computers. *Applied Sciences*, 11(18):8531, 2021.

[66] Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69, 2009.

[67] Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793, 2022.

[68] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[69] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[70] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

[71] Andrew P Rebera, Lode Lauwaert, and Ann-Katrien Oimann. Hidden risks: Artificial intelligence and hermeneutic harm. *Minds and Machines*, 35(3):33, 2025.

[72] Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan Kim, et al. Safetywashing: Do AI safety benchmarks actually measure safety progress? *Advances in Neural Information Processing Systems*, 37:68559–68594, 2024.

[73] Anka Reuel-Lamparth, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. Betterbench: Assessing AI benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813, 2024.

[74] Paul Ricoeur. *Hermeneutics and the human sciences: Essays on language, action and interpretation*. Cambridge University Press, 1981.

[75] Hannah Ringler. Computation and hermeneutics. *Computational Humanities*, page 1967, 2024.

[76] Alberto Romele, Marta Severo, and Paolo Furia. Digital hermeneutics: From interpreting with machines to interpretational machines. *AI & SOCIETY*, 35(1):73–86, 2020.

[77] Stanley Rosen. *Hermeneutics as politics*. Yale University Press, 2003.

[78] David Schlangen. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, 2021.

[79] Friedrich Schleiermacher. *Schleiermacher: Hermeneutics and criticism: and other writings*. Cambridge University Press, 1998.

[80] Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–29, 2024.

[81] Lorenzo C Simpson. *Hermeneutics as critique: Science, politics, race, and culture*. Columbia University Press, 2020.

[82] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302, 2024.

[83] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[84] Leon Staufer, Mick Yang, Anka Reuel, and Stephen Casper. Audit cards: Contextualizing ai evaluations. *arXiv preprint arXiv:2504.13839*, 2025.

[85] Dustin S Stoltz and Marshall A Taylor. Cultural cartography with word embeddings. *Poetics*, 88:101567, 2021.

[86] Peter Szondi. *Introduction to literary hermeneutics*. Cambridge University Press, 1995.

[87] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2024.

[88] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346, 2024.

[89] Paulina Tomaszewska and Przemysław Biecek. Position: Do not explain vision models without context. *Proceedings of Machine Learning Research*, 235, 2024.

[90] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

[91] Ted Underwood, Laura K Nelson, and Matthew Wilkens. Can language models represent the past without anachronism? *arXiv preprint arXiv:2505.00030*, 2025.

[92] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[93] Veniamin Veselovsky, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L Griffiths, and Arvind Narayanan. Localized cultural knowledge is conserved and controllable in large language models. *arXiv preprint arXiv:2504.10191*, 2025.

[94] Veniamin Veselovsky, Benedikt Stroebl, Gianluca Bencomo, Dilip Arumugam, Lisa Schut, Arvind Narayanan, and Thomas L Griffiths. Hindsight merging: Diverse data generation with language models. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

[95] Laura Weidinger, John Mellor, Bernat Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, Mark Díaz, A Bergman, Mikel Rodriguez, et al. Star: Sociotechnical approach to red teaming language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21516–21532, 2024.

[96] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative AI systems. *arXiv preprint arXiv:2310.11986*, 2023.

[97] Apurwa Yadav, Aarshil Patel, and Manan Shah. A comprehensive review on resolving ambiguities in natural language processing. *AI Open*, 2:85–92, 2021.

[98] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4):1–39, 2023.

[99] Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. Assessing and understanding creativity in large language models. *Machine Intelligence Research*, 22(3):417–436, 2025.

[100] Naitian Zhou, David Bamman, and Isaac L. Bleaman. Culture is not trivia: Sociocultural theory for cultural NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria, July 2025. Association for Computational Linguistics.