Article:
Kommers, C., Ahnert, R., Antoniak, M. et al. (2026) Computational hermeneutics: evaluating generative AI as a cultural technology. Frontiers in Artificial Intelligence, 9. 1753041. ISSN: 2624-8212

https://doi.org/10.3389/frai.2026.1753041

frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 26 February 2026
DOI 10.3389/frai.2026.1753041

OPEN ACCESS

EDITED BY
Meital Amzalag,
Holon Institute of Technology, Israel

REVIEWED BY
Gabriel Grill,
IT:U Interdisciplinary Transformation University Austria, Austria
Josh Andres,
University of New South Wales, Australia

*CORRESPONDENCE
Cody Kommers
ckommers@turing.ac.uk

RECEIVED 24 November 2025
REVISED 03 February 2026
ACCEPTED 06 February 2026
PUBLISHED 26 February 2026

CITATION
Kommers C, Ahnert R, Antoniak M, Benetos E, Benford S, Bunz M, Caramiaux B, Concannon S, Disley M, Dobson J, Du Y, Duéñez-Guzmán E, Francksen K, Gius E, Gray JWY, Heuser R, Immel S, So RJ, Leigh S, Livingston D, Long H, Martin M, Meyer G, Mihai D, Noel-Hirst A, Ostherr K, Parker D, Qin Y, Ratcliff J, Robinson E, Rodriguez K, Sobey A, Underwood T, Vashistha A, Wilkens M, Wu Y, Zheng Y and Hemment D (2026) Computational hermeneutics: evaluating generative AI as a cultural technology. *Front. Artif. Intell.* 9:1753041. doi: 10.3389/frai.2026.1753041

Computational hermeneutics: evaluating generative AI as a cultural technology

Cody Kommers[1]*, Ruth Ahnert[2], Maria Antoniak[3], Emmanouil Benetos[2], Steve Benford[4], Mercedes Bunz[5], Baptiste Caramiaux[6], Shauna Concannon[7], Martin Disley[8], James Dobson[9], Yali Du[5], Edgar Duéñez-Guzmán[10], Kerry Francksen[11], Evelyn Gius[12], Jonathan W. Y. Gray[5], Ryan Heuser[13], Sarah Immel[8], Richard Jean So[14], Sang Leigh[15], Dalaki Livingston[16], Hoyt Long[17], Meredith Martin[18], Georgia Meyer[19], Daniela Mihai[20], Ashley Noel-Hirst[2], Kirsten Ostherr[21], Deven Parker[22], Yipeng Qin[23], Jessica Ratcliff[15], Emily Robinson[24], Karina Rodriguez[25], Adam Sobey[1,20], Ted Underwood[26], Aditya Vashistha[15], Matthew Wilkens[15], Youyou Wu[27], Yuan Zheng[28] and Drew Hemment[1,29]

[1]The Alan Turing Institute, London, United Kingdom, [2]Queen Mary University of London, London, United Kingdom, [3]University of Colorado, Boulder, CO, United States, [4]University of Nottingham, Nottingham, United Kingdom, [5]King's College London, London, United Kingdom, [6]Sorbonne Université, Paris, France, [7]Durham University, Durham, United Kingdom, [8]University of Edinburgh, Edinburgh, United Kingdom, [9]Dartmouth College, Hanover, NH, United States, [10]Gibran AI, London, United Kingdom, [11]University of Coventry, Coventry, United Kingdom, [12]Technische Universität Darmstadt, Darmstadt, Germany, [13]University of Cambridge, Cambridge, MA, United States, [14]McGill University, Montreal, QC, Canada, [15]Cornell University, Ithaca, NY, United States, [16]University of Utah, Salt Lake City, UT, United States, [17]University of Chicago, Chicago, IL, United States, [18]Princeton University, Princeton, NJ, United States, [19]London School of Economics, London, United Kingdom, [20]University of Southampton, Southampton, United Kingdom, [21]Rice University, Houston, TX, United States, [22]University of Glasgow, Glasgow, United Kingdom, [23]Cardiff University, Cardiff, United Kingdom, [24]University of Exeter, Exeter, United Kingdom, [25]University of Brighton, Brighton, United Kingdom, [26]University of Illinois Urbana-Champaign, Champaign, IL, United States, [27]University College London, London, United Kingdom, [28]University of Sheffield, Sheffield, United Kingdom, [29]University of Edinburgh, Edinburgh, United Kingdom

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.

KEYWORDS
culture, GenAI, interpretation, meaning, societal impact

1 Introduction

Generative AI (GenAI) systems are cultural and social technologies (Bender et al., 2021; Farrell et al., 2025; Klein et al., 2025; Sorensen et al., 2024). Every aspect of these systems—from the social data they are trained on, to the benchmarks by which they are evaluated, to the societal effects of their outputs—depends on the complex web of norms, assumptions, meanings, practices, and social dynamics that make up culture and context. Their development is both influenced by culture (e.g., drawing on data reflecting perspectives from 21st century internet culture, while having been developed by engineers operating in a particular milieu) and in turn influences culture (e.g., shaping the kind of content people create and consume, while integrating into an ever-expanding set of production pipelines across society).

While this position is increasingly accepted as orthodoxy within the field of AI, 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 (Bolukbasi et al., 2016; Tao et al., 2024), a constraint for generalizing from one context to another (Yong Cao et al., 2023), a parameter in an ethical dilemma (Freitas et al., 2021), or a source of variability in user preferences (Ge et al., 2024). 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, the most prominent frontier models—many of which not only power popular general-purpose AI systems (e.g., ChatGPT, Gemini) but also underlie task-specific applications (e.g., Microsoft Copilot, Notion)—are not specialized systems designed to solve targeted, well-defined tasks. They are designed, and often 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 (Geertz, 1973; Hall, 1997; Klein et al., 2025). 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 (Raji et al., 2021). 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 (Zhou et al., 2025), while scalable "thin" metrics are often insufficient to capture key aspects of what makes them meaningful (Kommers et al., 2025b). 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 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 (Mohr et al., 2015; Rebera et al., 2025; Romele et al., 2020). 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 toward 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 use this active phrasing—to "do" interpretation—to reflect the fact that interpretive considerations are inextricably bound up in the processes and structures underlying these systems. The data, architecture, and algorithms of these systems are not static mirrors reflecting back invariant, disinterested projections of the world. They must be understood as comprising interpretive stakes, decisions, and processes. On the other hand, it is nonetheless crucial to recognize that GenAI do not "do" interpretation in the same way as humans and to avoid unduly imbuing them with anthropomorphic intentions (DeVrio et al., 2025; Akbulut et al., 2024).

With this in mind, 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 (Akbulut et al., 2025; Lazar and Nelson, 2023; Sorensen et al., 2024). 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 (Geertz, 1973). 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 (Caputo, 2018; Ricoeur, 1981). This concept is foundational across many disciplines and practices, from legal and literary studies (Levinson and Mailloux, 1988; Szondi, 1995) to debates in philosophy and aesthetics (Rosen, 2003; Simpson, 2020). It arises, in one form or another, whenever scholars confront epistemological problems of meaning.

A core concept within this tradition is the "hermeneutic circle" (Dilthey, 1989; Gadamer, 1960; Heidegger, 1927; Schleiermacher, 1998). 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 and context-specific generalizations 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 (Dobson, 2019; Ringler, 2024).

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 re-centers 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 (Gadamer, 1960). 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 (Hemment et al., 2025). No such epistemically totalitarian "view from nowhere" exists in any legitimate sense (Haraway, 1988). 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. For example, recent work has empirically demonstrated how GenAI systems can collapse perspectives into an idealized form, showing how further mechanisms are needed to maintain the individuation of distinct perspectives (Heuser, 2025).

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 (Desai et al., 2024; Sharma et al., 2024; Veselovsky et al., 2025b). Recent work on pluralistic, thick, or full-stack alignment recognizes that human values naturally conflict and advocates for systems that can accommodate this diversity (Lazar and Nelson, 2023; Lowe et al., 2025; Sorensen et al., 2024). 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 (Leibo et al., 2024). For example, what is viewed as AI "slop" in one context may be valued as a legitimate source of meaning or seen as having aesthetic resonance in another (Kommers et al., 2025a).

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 (Gadamer, 1960). 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 (Navigli, 2009). 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 dialog, or creating visual art—the goal is not necessarily to eliminate semantic uncertainty but to work productively within it (Gaver et al., 2003). 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 (Empson, 1930). However, current evaluation frameworks often treat this ambiguity as noise to be minimized rather than a generative resource (Yadav et al., 2021). 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 (Leibo et al., 2024; Veselovsky et al., 2025a). For example, recent approaches have sought to tease out the inherently multiplicitous perspectives contained within GenAI systems, developing processes for negotiating among conflicting viewpoints held within the model architecture (Li et al., 2024).

3 Generative AI systems as "context machines"

In this section, we argue that GenAI systems "do" interpretation as a fundamental capacity (Dobson, 2022)—and therefore evaluation of their performance is subject to the three hermeneutic challenges described above. Even so, it is important to note that the interpretive processes underlying these systems are distinct from those of human interpreters (Placani, 2024); while conversational systems may superficially adopt the voice of a human perspective, they should not be mistaken as inveterately human (Peter et al., 2025). For instance, such 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 toward 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 (Ethayarajh, 2019; Kozlowski et al., 2019; Stoltz and Taylor, 2021). Such embeddings are a means of encoding highly sophisticated co-occurrence statistics (Turney and Pantel, 2010). In language models, they are learned by poring over vast corpora of text (Mikolov et al., 2013; Pennington et al., 2014). In vision models, vectors of pixel values are often encoded as feature maps capturing edges, textures, and semantic patterns (Bengio et al., 2012; Mihai and Hare, 2021). Decoding these embeddings is also an interpretive act. This process is often probabilistic, accommodating a plurality of possible interpretations (James et al., 2013; Yang et al., 2023). 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 (Vaswani et al., 2017). 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 do not just "read in" context; they help create it

Interpretation does not only occur in isolation within GenAI models; these systems also co-construct interpretations in collaboration with humans (Frauenberger, 2019). 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 (Desai et al., 2024); formulate objective functions that reflect a specific set of goals, values, and assumptions (Lazar and Nelson, 2023); fine-tune system behavior through mechanisms like reinforcement learning from human feedback (Ouyang et al., 2022); and "engineer" prompts in order to elicit certain kinds of responses (Chen et al., 2025). At multiple layers of the system, human annotators—who can themselves offer conflicting interpretations (Frenda et al., 2024)—can provide feedback on ambiguous cases, rank responses, or supply preference scores, effectively staging a dialog where the AI's provisional interpretations can be contested and refined.

From machine to human, AI systems affect important mental capacities like metacognition (Tankelevitch et al., 2024); elicit different assumptions about relational norms [e.g., AI as assistant vs. therapist (Earp et al., 2025)]; act as thought-partners, for example by summarizing documents people would otherwise have to read—or skim—in full (Collins et al., 2024); shape human responses by explaining their own decisions (Doshi-Velez and Kim, 2017); and enable novel kinds of experience, such as certain creative practices (Caramiaux and Fdili Alaoui, 2022; Hemment et al., 2024; Murray-Browne and Tigas, 2021). Examples of how this approach has been employed include applications of assemblage thinking to study how AI is deployed (Tseng, 2023), as well as design frameworks that incorporate interpretive practices into multiple steps of the development process (Andres et al., 2025). 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 (Chang et al., 2024; Eriksson et al., 2025; Raji et al., 2021)—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 (Kapoor et al., 2024; McIntosh et al., 2025; Reuel-Lamparth et al., 2024; Schlangen, 2021). Furthermore, the implicit goal of benchmarking is often not to develop stronger metrics for specialized cases (though see Chiu et al. (2024) and Underwood et al. (2025)) 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 Norah Alzahrani et al. (2024), Ethayarajh and Jurafsky (2020), Koch and Peterson (2024), Raji et al. (2021), and Srivastava et al. (2023).

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 (Chang et al., 2024; Eriksson et al., 2025)—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 (Mizrahi et al., 2024). By contrast, cultural outputs are always part of an evolving conversation, whether a literal dialog or as a part of a broader evolutionary process (Brinkmann et al., 2023). 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 (Burnell et al., 2023). Overall, hermeneutic evaluations should seek to iteratively assess both the model's holistic capabilities, as well as its behavior in specific circumstances.

Existing benchmarks have begun to incorporate multi-turn iterative processes into their evaluation practices. Notable examples include assessments of chatbot capabilities in more than a dozen distinct tasks which evaluate performance over the course of an interaction with a human interlocutor (Bai et al., 2024), as well as assessments of perceived anthropomorphism in language models which show that the interpretation of model behavior as social (e.g., "relationship building" via empathic, emotionally-validating responses) only take place after multiple turns of interaction (Ibrahim et al., 2025a).

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 (Messeri and Crockett, 2024). 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 (Boisson et al., 2025; Chakrabarty et al., 2024; Marco et al., 2025); 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 dialog 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 (Chiu et al., 2024; Denton et al., 2020; Weidinger et al., 2024; Weidinger et al., 2023). Overall, hermeneutic evaluation requires benchmarks that assess interactivity rather than isolated performance, examining not just outputs but the interpretive dialog that generates them.

This practice is increasingly adopted in AI benchmarking. For example, assessments of harms from GenAI systems are sensitive to a larger range of potential issues—such as social manipulation or cognitive overreliance—only by evaluating the model capabilities in conjunction with their use by a human (Ibrahim et al., 2025b). Likewise, a recent benchmark looking at cultural expectation incorporates over 10,000 human annotations, reflecting norms and judgments based on the lived experience of people from a given cultural domain (Nayak et al., 2025).

4.3 Benchmarks should measure cultural context, not just model output

Individual interpretations of meaning depend on cultural context (Kommers and DeDeo, 2025)—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 (Kommers et al., 2025b). 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 (Akbulut et al., 2025; Liao and Xiao, 2023; Malaviya et al., 2025; Messeri and Crockett, 2024; Tomaszewska and Biecek, 2024). For example, frameworks like HELM recognize the need for contextually dependent approaches beyond accuracy (Liang et al., 2023). This can help address issues with current benchmarks, such as failure to capture real-world utility (Ott et al., 2022), or by adapting general processes to better fit situational needs (Staufer et al., 2025).

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 (Bhutani et al., 2024; Leibo et al., 2024). Evaluation must treat cultural context not as a constraint on model performance, but as the medium through which such performance emerges.

Some benchmarks are beginning to incorporate these kinds of contextual markers. For instance, a recent benchmark contrasts sociocultural norms for Chinese vs. American viewers of AI-generated videos (Varimalla et al., 2025). Similarly, a recent dataset organizes feedback from human raters based on demographic information, allowing for distinction in cultural judgments (Rastogi et al., 2026). In summary, it is worth noting that the strongest of exemplars of hermeneutic evaluation tend to adopt all three recommendations: they are iterative, incorporate human participants, and sensitive to sociocultural variation.

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.

In this article, we have focused on the evaluation of GenAI systems via benchmarks. We offer this as a potentially effective means by which scholars with a humanistic background can help shape the direction of technical development in AI. Benchmarks are an important part of how the field of AI progresses and understands its own progress. However, in practice benchmarks often fall short of meaningfully assessing what they purport to measure (Kapoor et al., 2024; McIntosh et al., 2025; Reuel-Lamparth et al., 2024; Schlangen, 2021), and it is widely acknowledged that better benchmarks are needed to support ethical and effective development of AI (Blagec et al., 2023; Ren et al., 2024; Zhao et al., 2025). One possible systemic cause of this is proxy failure (John et al., 2024): that the field's monocultural overreliance on standardized performance metrics is inadequate to capture the kinds of things we really want AI to do (Koch and Peterson, 2024; Kommers et al., 2025b; Zhou et al., 2025). This gives researchers with expertise in operationalizing tricky social or cultural concepts a useful lever for influencing this technology's metrics for success. But while we have focused on evaluation, this is not the only way to employ a hermeneutic perspective in AI. For example, on-going debates look at the cultural and social underpinnings of a model's training data (Mihalcea et al., 2025; Ravichander et al., 2025).

More generally, the hermeneutic tradition points to how powerful technological systems cannot be considered only in isolation, without appreciation of their environmental and societal consequences; this is a juncture at which crucial debates are being held and to which we hope a hermeneutic perspective can contribute. 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.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

CK: Writing – original draft, Writing – review & editing. RA: Writing – original draft, Writing – review & editing. MA: Writing – review & editing, Writing – original draft. EB: Writing – review & editing, Writing – original draft. SB: Writing – review & editing, Writing – original draft. MB: Writing – review & editing, Writing – original draft. BC: Writing – review & editing, Writing – original draft. SC: Writing – review & editing, Writing – original draft. MD: Writing – review & editing, Writing – original draft. JD: Writing – review & editing, Writing – original draft. YD: Writing – review & editing, Writing – original draft. ED-G: Writing – review & editing, Writing – original draft. KF: Writing – review & editing, Writing – original draft. EG: Writing – review & editing, Writing – original draft. JG: Writing – original draft, Writing – review & editing. RH: Writing – review & editing, Writing – original draft. SI: Writing – review & editing, Writing – original draft. RS: Writing – original draft, Writing – review & editing. SL: Writing – review & editing, Writing – original draft. DL: Writing – review & editing,

Writing – original draft. HL: Writing – original draft, Writing – review & editing. MM: Writing – review & editing, Writing – original draft. GM: Writing – review & editing, Writing – original draft. DM: Writing – review & editing, Writing – original draft. AN-H: Writing – review & editing, Writing – original draft. KO: Writing – review & editing, Writing – original draft. DP: Writing – original draft, Writing – review & editing. YQ: Writing – review & editing, Writing – original draft. JR: Writing – review & editing, Writing – original draft. ER: Writing – original draft, Writing – review & editing. KR: Writing – review & editing, Writing – original draft. AS: Writing – original draft, Writing – review & editing. TU: Writing – original draft, Writing – review & editing. AV: Writing – review & editing, Writing – original draft. MW: Writing – original draft, Writing – review & editing. YW: Writing – original draft, Writing – review & editing. YZ: Writing – original draft, Writing – review & editing. DH: Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the Alan Turing Institute under Lloyd's Register Foundation grant ATI/100004. This work also supported by the Arts and Humanities Research Council UK.

Acknowledgments

The authors wish to thank the reviewers who helped us present this material as clearly and effective as possible with their feedback.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author YW declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Akbulut, C., Weidinger, L., Manzini, A., Gabriel, I., and Rieser, V. (2024). All too human? Mapping and mitigating the risk from anthropomorphic AI. In Proceedings of the AAAI/ACM conference on AI, ethics, and society (Vol. 7. pp. 13–26).

Akbulut, M., Kevin Robinson, K., Maribeth Rauh, M., Isabela Albuquerque, I., Olivia Wiles, O., Laura Weidinger, L., et al. (2025). "Century: A framework and dataset for evaluating historical contextualisation of sensitive images" in The thirteenth international conference on learning representations.

Andres, J., Danta, C., Bianchi, A., Farzanfar, S., Milena Fernandez-Nieto, G., Becker, A., et al. (2025). "A scenario-based design pack for exploring multimodal human–GenAI relations" in Proceedings of the 27th international conference on multimodal interaction, 145–154.

Bai, G., Liu, J., Bu, X., He, Y., Liu, J., Zhou, Z., et al. (2024). Mt-bench-101: a fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots: can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (New York, NY: Association for Computing Machinery), 610–623.

Bengio, Y., Courville, A. C., and Vincent, P. (2012). Unsupervised feature learning and deep learning: a review and new perspectives. *CoRR* 1.

Bhutani, M., Robinson, K., Prabhakaran, V., Dave, S., and Dev, S. (2024). "SeeGULL multilingual: a dataset of geo-culturally situated stereotypes" in Proceedings of the 62nd annual meeting of the Association for Computational Linguistics, vol. 2, 842–854.

Blagec, K., Kraiger, J., Frühwirt, W., and Samwald, M. (2023). Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *J. Biomed. Inform.* 137:104274. doi: 10.1016/j.jbi.2022.104274

Boisson, J., Siddique, Z., Borkakoty, H., Antypas, D., Anke, L. E., and Camacho-Collados, J. (2025). "Automatic extraction of metaphoric analogies from literary texts: task formulation, dataset construction, and evaluation" in Proceedings of the 31st international conference on computational linguistics, 6692–6704.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* 29.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., et al. (2023). Machine culture. *Nat. Hum. Behav.* 7, 1855–1868. doi: 10.1038/s41562-023-01742-2

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., et al. (2023). Rethink reporting of evaluation results in AI. *Science* 380, 136–138. doi: 10.1126/science.adf6369

Caputo, J. D. (2018). *Hermeneutics: Facts and interpretation in the age of information*. London, UK: Penguin UK.

Caramiaux, B., and Fdili Alaoui, S. (2022). "Explorers of unknown planets": practices and politics of artificial intelligence in visual arts. *Proc. ACM hum.-Comput. Interact.* 6.

Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., and Wu, C.-S. (2024). "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: Association for Computing Machinery).

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., et al. (2024). A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* 15. doi: 10.1145/3641289

Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns* 6. doi: 10.1016/j.patter.2025.101260

Chiu, Y. Y., Sharma, A., Lin, I. W., and Althoff, T. (2024). A computational framework for behavioral assessment of LLM therapists. *arXiv preprint arXiv*.

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., et al. (2024). Building machines that learn and think with people. *Nat. Hum. Behav.* 8, 1851–1863. doi: 10.1038/s41562-024-01991-9

Denton, R., Hanna, A., Amironesei, R., Smart, A., Nicole, H., and Scheuerman, M. K. (2020). Bringing the people back in: contesting benchmark machine learning datasets. *arXiv preprint arXiv*.

Desai, M. A., Pasquetto, I. V., Jacobs, A. Z., and Card, D. (2024). An archival perspective on pretraining data. *Patterns* 5. doi: 10.1016/j.patter.2024.100966

DeVrio, A., Cheng, M., Egede, L., Olteanu, A., and Blodgett, S. L. (2025). "A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies"

in Proceedings of the 2025 CHI conference on human factors in computing systems, 1–18.

Dilthey, W. (1989). *Introduction to the human sciences, volume 1*. Princeton, NJ: Princeton University Press.

Dobson, J. E. (2019). *Critical digital humanities: The search for a methodology*. Chicago, Illinois: University of Illinois Press.

Dobson, J. E. (2022). Vector hermeneutics: on the interpretation of vector space models of text. *Digit. Scholarsh. Humanit.* 37, 81–93.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *Stat* 1050:2.

Earp, B. D., PorsdamMann, S., Aboy, M., Awad, E., Betzler, M., Botes, M., et al. (2025). Relational norms for human-AI cooperation. *arXiv preprint arXiv*.

Empson, W. (1930). Seven types of ambiguity.

Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., et al. (2025). Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation. *Proc. AAAI/ACM Conf. AI Ethics Soc.* 8. doi: 10.1609/aies.v8i1.36595

Ethayarajh, K. (2019). "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)* (Hong Kong, China: Association for Computational Linguistics), 55–65.

Ethayarajh, K., and Jurafsky, D. (2020). "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)* (New York City, NY: Association for Computational Linguistics), 4846–4853.

Farrell, H., Gopnik, A., Shalizi, C., and Evans, J. (2025). Large AI models are cultural and social technologies. *Science* 387, 1153–1156. doi: 10.1126/science.adt9819

Frauenberger, C. (2019). Entanglement HCI the next wave? *ACM Trans. Comput. Hum. Interact.* 27. doi: 10.1145/3364998

Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., Panizzon, R., Cignarella, A. T., et al. (2024). Perspectivist approaches to natural language processing: a survey. *Lang. Resour. Eval.*, 1–28.

Freitas, J. D., Censi, A., WalkerSmith, B., DiLillo, L., Anthony, S. E., and Frazzoli, E. (2021). From driverless dilemmas to more practical commonsense tests for automated vehicles. *Proc. Natl. Acad. Sci. USA* 118:e2010202118. doi: 10.1073/pnas.2010202118

Gadamer, H.-G. (1960). Truth and method.

Gaver, W. W., Beaver, J., and Benford, S. (2003). "Ambiguity as a resource for design" in *Proceedings of the SIGCHI conference on human factors in computing systems* (New York, NY: Association for Computing Machinery), 233–240.

Ge, X., Xu, C., Misaki, D., Markus, H. R., and Tsai, J. L. (2024). "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: Association for Computing Machinery).

Geertz, C. (1973). *The interpretation of cultures*. New York City, NY: Basic Books.

Hall, S. (1997). Representation: cultural representations and signifying practices. *Culture*.

Haraway, D. (1988). Situated knowledges: the science question in feminism and the privilege of partial perspective. *Feminist Stud.* 14, 575–599.

Heidegger, M. (1927). Being and time.

Hemment, D., Kommers, C., et al. (2025). *Doing AI differently: Rethinking the foundations of AI via the humanities. Technical report*. London: The Alan Turing Institute.

Hemment, D., Murray-Rust, D., Belle, V., Aylett, R., Vidmar, M., and Broz, F. (2024). Experiential AI: between arts and explainable AI. *Leonardo* 57, 298–306.

Heuser, R. (2025). Cultural collapse: toward a generative formalism for AI cultural production. *Anthology Computers Humanities* 3, 575–588. doi: 10.63744/usvuyzsiapvy

Ibrahim, L., Akbulut, C., Elasmar, R., Rastogi, C., Kahng, M., Morris, M. R., et al. (2025a). Multi-turn evaluation of anthropomorphic behaviours in large language models. *arXiv preprint arXiv:2502.07077*.

Ibrahim, L., Huang, S., Ahmad, L., Bhatt, U., and Anderljung, M. (2025b). Towards interactive evaluations for interaction harms in human-AI systems. In Proceedings of the AAAI/ACM conference on AI, ethics, and society (Vol. 8. pp. 1302–1310).

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R, volume 103*. New York City, NY: Springer.

John, Y. J., Caldwell, L., McCoy, D. E., and Braganza, O. (2024). Deadrats, dopamine, performance metrics, and peacock tails: proxy failure is an inherent risk in goal-oriented systems. *Behav. Brain Sci.* 47:e67.

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. (2024). AI agents that matter. *arXiv preprint arXiv*.

Klein, L., Martin, M., Brock, A., Antoniak, M., Walsh, M., Johnson, J. M., et al. (2025). Provocations from the humanities for generative AI research. *arXiv preprint arXiv*.

Koch, B. J., and Peterson, D. (2024). From protoscience to epistemic monoculture: how benchmarking set the stage for the deep learning revolution. *arXiv preprint arXiv*.

Kommers, C., and DeDeo, S. (2025). "Sense-making, cultural scripts, and the inferential basis of meaningful experience" in Proceedings of the annual meeting of the cognitive science society, vol. 47.

Kommers, C., Duede, E., Gordon, J., Holtzman, A., McNulty, T., Stewart, S., et al. (2025a). Why slop matters. *arXiv preprint arXiv*:2601.06060.

Kommers, C., Hemment, D., Antoniak, M., and Leibo, J. Z. (2025b). Meaning is not a metric: using LLMs to make cultural context legible at scale. *arXiv preprint arXiv*.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The geometry of culture: analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* 84, 905–949.

Lazar, S., and Nelson, A. (2023). AI safety on whose terms? *Science* 381:138. doi: 10.1126/science.adi8982

Leibo, J. Z., Vezhnevets, A. S., Diaz, M., Agapiou, J. P., Cunningham, W. A., Sunehag, P., et al. (2024). A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv*.

Levinson, S., and Mailloux, S. (1988). *Interpreting law and literature: A hermeneutic reader*. Chicago, Illinois: North- western University Press.

Li, M., Chen, J., Chen, L., and Zhou, T. (2024). "Can LLMs speak for diverse people? Tuning LLMs via debate to generate controllable controversial statements" in Findings of the Association for Computational Linguistics ACL 2024, 16160–16176.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., et al. (2023). Holistic evaluation of language models. *Transactions Machine Learning Research*.

Liao, Q. V., and Xiao, Z. (2023). Rethinking model evaluation as narrowing the sociotechnical gap. *arXiv preprint arXiv*.

Lowe, R., Edelman, J., Zhi-Xuan, T., Klingefjord, O., Hain, E., Wang, V., et al. (2025). "Full-stack alignment: co-aligning AI and institutions with thicker models of value" in 2nd workshop on models of human feedback for AI alignment.

Malaviya, C., Chang, J. C., Roth, D., Iyyer, M., Yatskar, M., and Lo, K. (2025). Contextualized evaluations: judging language model responses to underspecified queries. *Trans. Assoc. Comput. Linguist.* 13, 878–900.

Marco, G., Gonzalo, J., and Fresno, V. (2025). The reader is the metric: how textual features and reader profiles explain conflicting evaluations of AI creative writing. *arXiv preprint arXiv*.

McIntosh, T., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., et al. (2025). Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Trans. Artif. Intell.*

Messeri, L., and Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature* 627, 49–58. doi: 10.1038/s41586-024-07146-0

Mihai, D., and Hare, J. (2021). Learning to draw: emergent communication through sketching. *Adv. Neural Inf. Process. Syst.* 34, 7153–7166.

Mihalcea, R., Ignat, O., Bai, L., Borah, A., Chiruzzo, L., Jin, Z., et al. (2025). Why AI is WEIRD and shouldn't be this way: towards AI for everyone, with everyone, by everyone. In Proceedings of the AAAI conference on artificial intelligence (Vol. 39. pp. 28657–28670).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* 26.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., and Stanovsky, G. (2024). State of what art? A call for multi-prompt LLM evaluation. *Trans. Assoc. Comput. Linguist.* 12, 933–949. doi: 10.1162/tacl

Mohr, J. W., Wagner-Pacifici, R., and Breiger, R. L. (2015). Toward a computational hermeneutics. *Big Data Soc.* 2:2053951715613809. doi: 10.1177/2053951715613809

Murray-Browne, T., and Tigas, P. (2021). Emergent interfaces: vague, complex, bespoke and embodied interaction between humans and computers. *Appl. Sci.* 11:8531. doi: 10.3390/app11188531

Navigli, R. (2009). Word sense disambiguation: a survey. *ACM Computing Surveys* 41, 1–69.

Nayak, S., Bhatia, M., Zhang, X., Rieser, V., Hendricks, L. A., Van Steenkiste, S., et al. (2025). "Culturalframes: assessing cultural expectation alignment in text-to-image models and evaluation metrics" in Findings of the Association for Computational Linguistics: EMNLP, 20918–20953.

Norah Alzahrani, N., Hisham Alyahya, H., Yazeed Alnumay, Y., Sultan AlRashed, S., Shaykhah Alsubaie, S., Yousef Al- mushayqih, Y., et al. (2024). 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. Association for Computational Linguistics.

Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J., and Samwald, M. (2022). Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nat. Commun.* 13:6793. doi: 10.1038/s41467-022-34591-0

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* 35, 27730–27744.

Pennington, J., Socher, R., and Manning, C. (2014). "GloVe: global vectors for word represen- tation" in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (Doha, Qatar: Association for Computational Linguistics), 1532–1543.

Peter, S., Riemer, K., and West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proc. Natl. Acad. Sci. USA* 122:e2415898122. doi: 10.1073/pnas.2415898122

Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. *AI Ethics* 4, 691–698. doi: 10.1007/s43681-024-00419-4

Raji, I. D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. (2021). AI and the everything in the whole wide world benchmark. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Rastogi, C., Teh, T. H., Mishra, P., Patel, R., Wang, D., Diaz, M., et al. (2026). Whose view of safety? A deep DIVE dataset for pluralistic alignment of text-to-image models. In The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Ravichander, A., Fisher, J., Sorensen, T., Lu, X., Antoniak, M., Lin, B. Y., et al. (2025). "Information-guided identification of training data imprint in (proprietary) large language models" in Proceedings of the 2025 conference of the nations of the Americas chapter of the Association for Computational Linguistics: Human language technologies, vol. 1, 1962–1978.

Rebera, A. P., Lauwaert, L., and Oimann, A.-K. (2025). Hidden risks: artificial intelligence and hermeneutic harm. *Minds Mach.* 35:33. doi: 10.1007/s11023-025-09733-0

Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., et al. (2024). Safetywashing: do AI safety benchmarks actually measure safety progress? *Adv. Neural Inf. Process. Syst.* 37, 68559–68594.

Reuel-Lamparth, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. J. (2024). Betterbench: assessing AI benchmarks, uncovering issues, and establishing best practices. *Adv. Neural Inf. Process. Syst.* 37, 21763–21813.

Ricoeur, P. (1981). *Hermeneutics and the human sciences: Essays on language, action and interpretation*. Cambridge, UK: Cambridge University Press.

Ringler, H. (2024). Computation and hermeneutics. *Computational Humanities*:1967.

Romele, A., Severo, M., and Furia, P. (2020). Digital hermeneutics: from interpreting with machines to interpretational machines. *AI & Soc.* 35, 73–86.

Rosen, S. (2003). *Hermeneutics as politics*. New Haven, CT: Yale University Press.

Schlangen, D. (2021). "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, vol. 2, 670–674.

Schleiermacher, F. (1998). *Schleiermacher: Hermeneutics and criticism: and other writings*. Cambridge, UK: Cambridge University Press.

Sharma, A., Rushton, K., Lin, I. W., Nguyen, T., and Althoff, T. (2024). "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, 1–29.

Simpson, L. C. (2020). *Hermeneutics as critique: Science, politics, race, and culture*. New York City, NY: Columbia University Press.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C.- p. M., et al. (2024). "Position: A roadmap to pluralistic alignment" in Proceedings of the 41st international conference on machine learning, 46280–46302.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A., Abid, A., Fisch, A., et al. (2023). Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *Transactions Machine Learning Research*.

Staufer, L., Yang, M., Reuel, A., and Casper, S. (2025). Audit cards: contextualizing ai evaluations. *arXiv preprint arXiv*.

Stoltz, D. S., and Taylor, M. A. (2021). Cultural cartography with word embeddings. *Poetics* 88:101567. doi: 10.1016/j.poetic.2021.101567

Szondi, P. (1995). *Introduction to literary hermeneutics*. Cambridge, UK: Cambridge University Press.

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., et al. (2024). "The metacognitive demands and opportunities of generative AI" in Proceedings of the 2024 CHI conference on human factors in computing systems, 1–24.

Tao, Y., Viberg, O., Baker, R. S., and Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3:pgae346. doi: 10.1093/pnasnexus/pgae346

Tomaszewska, P., and Biecek, P. (2024). Position: do not explain vision models without context. *Proc. Mach. Learn. Res.* 235.

Tseng, Y. S. (2023). Assemblage thinking as a methodology for studying urban AI phenomena. *AI & Soc.* 38, 1099–1110.

Turney, P. D., and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* 37, 141–188. doi: 10.1613/jair.2934

Underwood, T., Nelson, L. K., and Wilkens, M. (2025). Can language models represent the past without anachronism? *arXiv preprint arXiv*.

Varimalla, N. R., Xu, Y., Saakyan, A., Wang, M. F., and Muresan, S. (2025). VideoNorms: benchmarking cultural awareness of video language models. *arXiv preprint arXiv:2510.08543*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.

Veselovsky, V., Argin, B., Stroebl, B., Wendler, C., West, R., Evans, J., et al. (2025a). Localized cultural knowledge is conserved and controllable in large language models. *arXiv preprint arXiv*.

Veselovsky, V., Stroebl, B., Bencomo, G., Arumugam, D., Schut, L., Narayanan, A., et al. (2025b). Hindsight merging: diverse data generation with language models. In The 41st Conference on Uncertainty in Artificial Intelligence.

Weidinger, L., Mellor, J., Pegueroles, B., Marchal, N., Kumar, R., Lum, K., et al. (2024). "Star: sociotechnical approach to red teaming language models" in Proceedings of the 2024 conference on empirical methods in natural language processing, 21516–21532.

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., et al. (2023). Sociotechnical safety evaluation of generative AI systems. *arXiv preprint arXiv*.

Yadav, A., Patel, A., and Shah, M. (2021). A comprehensive review on resolving ambiguities in natural language processing. *AI Open* 2, 85–92.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., et al. (2023). Diffusion models: a comprehensive survey of methods and applications. *ACM Comput. Surv.* 56, 1–39.

Yong Cao, Y., Li Zhou, L., Seolhwa Lee, S., Laura Cabello, L., Min Chen, M., and Daniel Hershcovich, D. (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)* (Dubrovnik, Croatia: Association for Computational Linguistics), 53–67.

Zhao, Y., Zhang, R., Li, W., and Li, L. (2025). Assessing and understanding creativity in large language models. *Mach. Intell. Res.* 22, 417–436.

Zhou, N., Bamman, D., and Bleaman, I. L. (2025). "Culture is not trivia: sociocultural theory for cultural NLP" in *Proceedings of the 63rd annual meeting of the Association for Computational Linguistics*, vol. *1* (Vienna, Austria: Association for Computational Linguistics), 25869–25886.