Computational Humanities Research

www.cambridge.org/chr

A theory-first approach toward using generative AI for humanities research

Andrew Piper

McGill University, Canada

Research Article

Cite this article: Piper Andrew. 2026. "A theory-first approach toward using generative AI for humanities research." *Computational Humanities Research*, **2**:e18, https://doi.org/10.1017/chr.2026.10039

Received: 30 January 2025 UTC
Revised: 12 February 2026 UTC
Accepted: 1 June 2026 UTC

Keywords:
generative AI; LLMs; perspectival AI; reproducibility; research design; theory

Corresponding author:
Andrew Piper;
Email: andrew.piper@mcgill.ca

Abstract

The integration of generative AI into humanities research presents both new opportunities and significant methodological challenges. While generative AI enables more sophisticated textual, visual, and historical analysis than ever before, it also introduces novel methodological risks. Prompts now function as measurement instruments, construct validity is entangled with model fluency, the increasing abstraction and subjectivity of tasks challenge single-answer evaluations, small changes in prompt or model configuration can materially alter substantive conclusions, and model training encodes strong theoretical assumptions prior to human interaction. This article advances a theory-first framework for generative-AI-assisted humanities research that traces how theoretical commitments intervene across the research lifecycle. Drawing on examples from the computational humanities and social sciences, it argues that theory operates as *methodological infrastructure* that constrains degrees of freedom, supports reproducibility, and enables principled iteration. By reframing generative AI as a site of theoretical inscription, the article outlines practical strategies for building robust and theoretically grounded humanities research in the era of generative AI.

Plain language summary

The use of artificial intelligence, particularly multi-modal large language models (LLMs/MLLMs), provides exciting opportunities for humanities research. These AI systems, which can analyze and generate human-like text and images, offer new ways for scholars to study literature, history, and culture. However, their use also raises important questions: How do we ensure that these models align with the goals of humanities research?

This article illustrates why a theory-first approach matters when using generative AI in humanities research. Through concrete examples and visual schematics, it shows how theory can guide each stage of the research process – from formulating meaningful research questions, specifying constructs, designing and validating prompts, to developing novel benchmarks and intervening in model tuning. Rather than treating LLMs as neutral tools that are "human-like," the article argues that theory is indispensable for guiding their use toward the discovery of new ideas and refining existing knowledge.

Introduction

The advent of large language models (LLMs), and now multi-modal models (MLLMs), has transformed the research landscape across multiple disciplines, including the humanities. While quantitative analysis of cultural and historical data was already a well-established field, the generative capacities of models enable researchers to detect and analyze far more nuanced and complex concepts than prior methods. This can be conceptualized as a second-wave form of scale: from the initial research gains of analyzing quantitatively more documents using computational methods (wave 1), we can now derive qualitative gains through the increased sophistication of questions that can be asked and answered. Such conceptual "scaling up," however, is not without distinct challenges. The growing adoption of generative AI in humanities research has only intensified the need to address fundamental questions about methodology, validity, and the relationship between theory and computational tools.

This article advocates for a "theory-first" approach toward the use of generative AI, emphasizing the need for humanities scholars to ground their practices in robust theoretical frameworks.[1] Early debates around the digital humanities frequently critiqued an over-reliance on a "tool-first" methodology, highlighting a perceived deficit of theoretical engagement (Cohen and Scheinfeldt 2013; Warwick 2015; Windhager and Mayr 2024). Debates about the place of theory within the humanities more broadly is of course already a well-worn topic (Eagleton 2011, 2004). The issue

[1] I owe a debt to Schmidt (2011) who coined this term in a 2011 blog post and which has since been used in a number of academic articles, that is, I am not the first to use the term "theory first."

This article was awarded Open Data and Open Materials badges for transparent practices. See the Data availability statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

of sufficient theoretical grounding is also not unique to the humanities but extends into fields like natural language processing and computational linguistics, where the development of tools and methods often outpaces critical reflection on their cultural and historical implications (Radford and Joseph 2020).

My aim is not to revisit these debates to definitively settle the question of a theory deficit in computational research. Nor am I the first to focus on the issue of theory within computational research (Cecire 2011; McCarthy and Dore 2023; Radford and Joseph 2020; Underwood 2020; Underwood et al. 2022). Rather, the goal of this article is to articulate how generative AI reshapes the basic steps of scholarly research in ways that make theoretical commitments increasingly consequential (see Karjus 2025 for related work). Rather than treating theory as its own layer or end unto itself, theory can be understood as a form of methodological infrastructure when working with generative models – guiding problem formation, shaping the operationalization of constructs, constraining evaluation practices, and orienting our interpretations.

Generative AI lowers many of the technical barriers that once constrained computational research, making it easier to ask complex questions but also to conflate model fluency with conceptual precision. The theory-first approach developed here offers a set of practical steps or frameworks that researchers at all levels can use to ensure computational research continues to be theoretically rigorous in an era of generative AI. A theory-first approach toward AI-assisted research will thus ideally help:

1. build more concretely from existing traditions of humanities scholarship, which will facilitate disciplinary integration;
2. guide and constrain the interpretive degrees of freedom involved in the study of cultural and historical data and that expand exponentially under generative AI;
3. facilitate reproducibility as the acknowledged gold standard of scientific research;
4. test and refine inherited theoretical frameworks and innovate new directions.

My goal here is not to innovate at the level of theory. I leave theory-for-theory's-sake arguments to others. Nor is it to offer a hot take on a novel way to use AI – those approaches change too rapidly to serve as a stable foundation. Instead, my aim is more quotidian: to step back and support what Kuhn called "normal science," that is, the everyday practices of academic research, where generative AI is now one of the core computational components. Such practices of normalization are essential for building and nurturing a field that is rapidly developing in exciting new ways. It starts, as it must, with theory.

A (very) brief theory of theory

Already in 1949, the sociologist Robert K. Merton warned, "like so many words that are on everyone's lips, the word 'theory' is in danger of losing all meaning" (Merton 1949). Nevertheless, Merton's work would go on to serve a foundational role in framing the function of theory for academic research. According to Merton, a theory should consist of three key functions: it should be *explanatory*, *predictive*, and *iterative*. A theory of any aspect of the world should be able to explain salient relationships between core aspects of that world, up to and including causal relationships, that is, why things happened the way they did.

Second, a theory should be able to predict future outcomes, that is, it should generalize to unforeseen circumstances. A theory of the world is not valuable if it cannot help us navigate novel information. Such novel information must not simply be future-oriented; it can also entail new information about the past.

Finally, a theory should be an appropriate scale such that it can be iteratively refined as more and more of the world is observed. This was Merton's theory of the value of "middle-range" theories, as opposed to the "grand" theories of many of his sociological predecessors. Middle-range theories are closer to the world and thus can evolve as our understanding of the world evolves. Grand theories attempt to explain everything and are therefore too static. For Merton, they were too out of touch with reality.

Merton's is just one way of thinking about the value and limitations of theory for academic research. There is a voluminous literature on the topic. But it offers to my mind one of the most succinct and actionable frameworks to guide how we think about the role of theory within AI-assisted research. For Merton, theory must be in touch with the world it aims to explain; must be able to adequately incorporate new information (i.e., generalize well); and finally, theory must be dynamic in its construction to align with the goal of cognitive updating.

In what follows, I will highlight the place of this theory of theory within the core practices of AI-assisted research.

Step 1. Theory-guided problem formation under generative AI

Generative AI reshapes how research problems are formed. They do so in three principal ways: by lowering the cost of abstraction; facilitating observations across different textual or visual scales; and generating outputs that closely mirror human judgment. In doing so, they expand what appears tractable, measurable, and meaningful from a research perspective.

At the same time, they introduce new risks. Generative AI blurs the boundary between models as cultural objects and scholarly instruments. Research problems may be defined in terms of what models easily produce rather than what theories demand. And analytical constructs may drift toward model-internal regularities rather than culturally grounded distinctions. Theory thus plays a crucial role at this initiatory step to orient ourselves toward each of these concerns (Figure 1).

AI as object or instrument

A first and foundational decision is whether the model itself is the object of analysis or the instrument through which a cultural phenomenon is studied. While this distinction has long existed in computational research, generative AI brings this problem more strongly to the foreground. LLMs can simultaneously function as tools for performing interpretive tasks, such as classification, summarization, or evaluation and as cultural artifacts that shape social reality. Without an explicit theoretical orientation, research questions can slide unintentionally between these roles, conflating claims about model behavior with claims about cultural meaning. Theory provides a necessary guide at this juncture, clarifying what kind of knowledge a given study seeks to produce and what counts as evidence in support of that aim.

Model as object

In their paper, "Marked personas: Using natural language prompts to measure stereotypes in language models," Cheng, Durmus, and

https://doi.org/10.1017/chr.2026.10039 Published online by Cambridge University Press

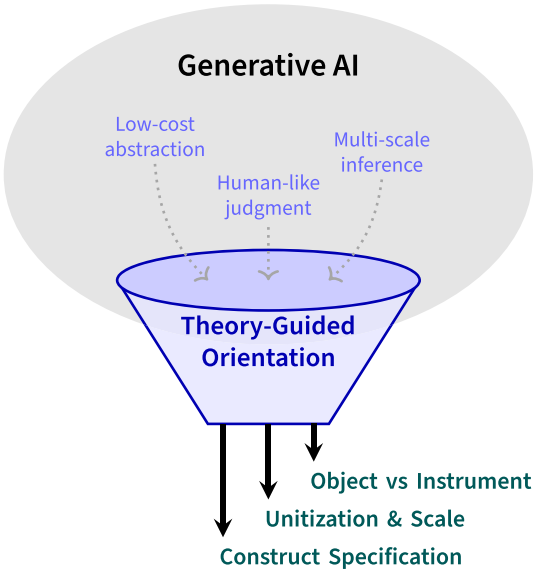


Figure 1. Theory-guided problem formation under generative AI. Generative AI dramatically expands what appears measurable through affordances, such as low-cost abstraction, multi-scale inference, and human-like judgment. Theory functions as an orienting constraint, shaping which affordances become legitimate research problems. The three pathways represent distinct dimensions of theory-guided specification.

Jurafsky (2023) test whether LLMs reproduce stereotypes when prompted to describe personas from marginalized demographic groups. This is a classic AI as object evaluation. In particular, they are interested in the theoretical problem of "benevolent stereotyping," when models utilize positive, but essentializing language, for example, when Black women are repeatedly described as "resilient" and "strong," Asian women as "petite" and "delicate," and Latina women as "vibrant" and "curvaceous." These attributes, though technically positive in sentiment, reinforce limiting archetypes and obscure individual variability. For their study, the researchers begin with a theory of "representational harm" and probe models for this *general* behavior. Do models encode certain kinds of beliefs and representations and how might we surface them?

Model as instrument

This question however is different from asking whether models *will* do this under different, expert-guided conditions. For example, Hobson et al. (2024) combine theories of ethical criticism and narrative archetypes to test whether LLMs can generate "story morals" that align with human readers' judgments and preferences. Here, the goal is not to generalize about model behavior but rather to use the model as an analytical tool to study moral perspectives of cultural narratives. We can think of this as an alignment problem: are models (or individual models) *generally aligned* with human beliefs such that we can make inferences about the world through them?

The difference between these two approaches is ultimately one of constraint: model as object research generalizes to potentially future use cases, and model as instrument research generalizes about a particular application under particular conditions.

Unitization and scale

When forming research problems under generative AI, questions of unitization and scale take on renewed urgency (Karjus 2025). LLMs operate fluently across multiple levels of textual analysis – tokens, sentences, documents, and collections – dramatically lowering the technical barriers that once constrained scale selection. As a result, scale is no longer enforced by feasibility but becomes a theoretical decision that must be made explicitly.

One must consider, for example, whether a phenomenon is best understood at the level of words or phrases (as in the case of benevolent stereotypes or sentiment), at the level of documents (as with story morals), or across collections of texts. Social network analysis is a useful lens through which we can observe this problem. In their study of narrative "character interaction networks" using LLMs, Piper, Xu, and Ruths (2024) focus on the sentence level to identify fine-grained interaction types, treating narrative relations as locally instantiated and sequential. Hamilton et al. (2025) and Christou and Tsoumakas (2025), on the other hand, model "character relation networks" at the document level, emphasizing more durable patterns of association that emerge across an entire narrative. Much historical work, meanwhile, operates at an even broader scale, constructing social networks across documents (Ahnert et al. 2020), including studies of book dedications (Ladd 2021), court structures (Tambuscio and Vogeler 2025), and epistolary exchange (Edelstein et al. 2017).

What unites these approaches is a shared reliance on theory to justify how scale maps onto meaning. Under generative AI, this mapping becomes far more fluid and accessible and is thus even more in need of theoretical framing. Models can infer relations at sentence, document, or corpus levels with comparable ease, but the interpretive implications of these choices differ substantially. Theory thus plays a critical role in specifying what counts as a unit, what constitutes a connection, and how edges and nodes should be interpreted at a given level of aggregation (Emirbayer and Goodwin 1994; Ryan and Ahnert 2021).

Construct specification

The final and arguably most foundational role of theory in problem formation under generative AI lies in the specification of analytical constructs. LLMs make it deceptively easy to invoke complex cultural concepts through prompts, producing fluent outputs for terms, such as "narrative discourse," "social networks," "visual point of view," or "cultural change" without requiring those concepts to be rigorously defined. This affordance creates a new form of conceptual risk: constructs may appear operationalized simply because models can generate plausible responses, even when the underlying theoretical commitments remain underspecified. The risk lies in surfacing model presuppositions and not the theoretical focus of the research question. Theory is therefore essential not merely for naming constructs, but for stabilizing their meaning – clarifying what a construct is intended to capture, at what level it operates, and how it relates to broader cultural or historical processes.

To take one recent framework of interest to the field, the concepts of "reception," "reader response," or "social reading" have served as an increasingly important yet flexible constructs across recent computational humanities research. In one of the conference proceedings of *CHR 2024*, for example, we can find four papers that operationalize this construct in four distinct ways: as literary quality (Jacobsen et al. 2024), topical alignment (Yu and Pianzola 2024), individual preference (Rebora and Vezzani 2024), and the temporal trajectory of recognition (Kim 2024). Each frames reception as a particular lens through which texts interact with readers, institutions, and cultural systems over time.

https://doi.org/10.1017/chr.2026.10039 Published online by Cambridge University Press

Importantly, each operationalization involves different kinds of data, different labeling systems, and different theoretical perspectives. Yu and Pianzola (2024) foreground "digital social reading" (Pianzola 2025; Rebora et al. 2021) through the study of web novels. Rebora and Vezzani (2024) emphasize the social construction of literary value (Von Heydebrand and Winko 2008) via Goodreads reviews. Jacobsen et al. (2024) focus on questions of cultural capital and canon formation using fanfiction, and Kim (2024) draws deeply from Bourdieu's field theory but critiques its synchronic bias by applying Abbott's life-course modeling to literary careers. These studies illustrate how the concept of "reception" is not a single construct but a dynamic nexus between theory, data, and method.

Checklist Step 1: Problem formation

- **Clarify the role of the model:** Is generative AI the object of analysis, whose cultural assumptions and representational patterns are being studied, or an instrument used to investigate another cultural phenomenon?
- **Treat scale as an interpretive choice:** Specify whether the relevant phenomenon operates at the level of words, sentences, documents, collections, images, or networks, and explain why that scale is theoretically appropriate.
- **Specify the construct before prompting the model:** Do not assume that concepts such as "reception," "reader response," "social networks," or "cultural change" are operationalized simply because a model can generate fluent responses about them.
- **Align theory, data, and method:** Ensure that the chosen data source, labeling scheme, model task, and interpretive framework all instantiate the same theoretical construct.
- **Guard against construct drift:** Ask whether the model is capturing the intended concept or merely reproducing model-internal regularities, prompt conventions, or culturally generic associations.

Step 2. Evaluation and validation

Once we have made theory-informed choices around scale, construct specification, and the role of the model (as object or instrument), we next move into the process of model evaluation. With generative AI, this phase is anchored in *prompt design* (or "prompt engineering"), which functions as the primary mechanism through which theoretical constructs are operationalized. What distinguishes generative AI from earlier computational approaches to the study of culture is the way it does so through open-ended generative processes that are highly sensitive to instructional framing. For any given construct, there are innumerable ways to elicit model responses, each of which implicitly encodes theoretical assumptions. As a result, theory becomes essential for constraining the space of possible prompt formulations, grounding model outputs through validation against human judgments, and determining whether what is being produced genuinely corresponds to the construct under study.

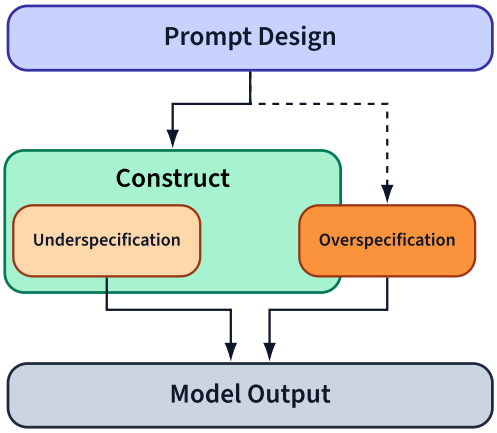


Figure 2. Prompt design operationalizes a construct. Validity threats arise from (left) prompt underspecification (content underrepresentation) and (right) prompt overspecification (construct-irrelevant variance), both shaping the observed model output.

Prompt design and construct validity

While the term "prompt engineering" is often used (suggesting a primarily technical pathway toward improved performance), a more productive way to understand prompt design in generative-AI-assisted research is through the lens of construct validity (Smith 2005; Strauss and Smith 2009). In work with LLMs, prompts do not merely instruct a system but function as measurement instruments, translating theoretical constructs into operational procedures that elicit model outputs. Because generative models can produce fluent responses to a wide range of instructions, the apparent success of a prompt may mask deeper misalignments between what is being elicited and the construct a researcher intends to study.

Construct validity addresses how well our operational measures capture the theoretical constructs we intend to study (i.e., "Are we measuring what we think we're measuring?"). This concern is not new to the social sciences or computational humanities, but generative AI intensifies it in two important ways. First, the open-ended and probabilistic nature of model outputs makes it easy to obtain plausible responses even when constructs are underspecified. Second, the flexibility of prompting introduces a vast space of possible operationalizations, each of which implicitly encodes assumptions about what the construct entails. As a result, validity can no longer be assumed on the basis of surface coherence or task completion alone.

Two classic threats to construct validity are especially useful for diagnosing the risks introduced by prompt-based measurement (Figure 2)

Content underrepresentation

This occurs when a prompt elicits only a narrow slice of the intended construct. For example, a prompt designed to identify a narrative's "moral" that focuses exclusively on explicit didactic statements may underrepresent broader ethical dimensions conveyed through character action, consequence, or narrative structure. In such cases, the model may perform consistently while systematically omitting theoretically important aspects of the construct.

A good example from the literature of content underrepresentation can be found in early work on "plot arcs" (Reagan

https://doi.org/10.1017/chr.2026.10039 Published online by Cambridge University Press

et al. 2016). A theory of rising and falling action imported from classical tragedy was captured by mapping sentiment over narrative time. One core dimension that such an approach misses is the problem of narrative focalization – from whose perspective is such narrative valence being experienced? The measure (sentence-level sentiment) under-represents the full construct (narrative arc or fortune). An LLM that captured sentiment in this way, even if more accurate and sensitive to context, would still underrepresent the construct under study.

Construct-irrelevant variance

Here, prompt design may inadvertently elicit patterns that reflect model-specific biases or linguistic regularities rather than the target construct itself. A useful illustration can be found in recent theory-of-mind benchmarks for LLMs. "Theory of Mind" (ToM) refers to a model's (or human mind's) ability to infer beliefs and intentions of other people. In Chen et al. (2024), the authors introduce ToMBench, a broad and carefully constructed battery of tasks designed to address problems of content underrepresentation in previous theory-of-mind benchmarks by sampling widely across 31 established ToM abilities.

At the same time, the prompts used to elicit model responses are framed as story–question–answer selections. This framing may capture abilities that are not associated with mental-state attribution. Because models are asked to select the "correct" answer from a set of plausible options given a short narrative, this framing invites a strategy based on textual entailment and semantic plausibility rather than perspective-sensitive reasoning. In practice, models can succeed by identifying the option that best matches salient lexical cues in the story without representing what a character knows or believes.

Annotating for human diversity

Once prompts have been designed with construct validity in mind, a further question immediately arises: how do we determine whether model outputs meaningfully correspond to the constructs they are intended to operationalize? One of the core ways we do so is by recording human judgments about whatever it is we are studying. Does the model *align* with human beliefs?

A growing body of work across computational humanities, NLP, and human–computer interaction has emphasized that many, if not most, humanistic constructs are inherently plural (Frenda et al. 2025; Ying et al. 2025). Basile (2020) has memorably termed this "the end of the gold standard," a play on concepts of "ground truth" used to frame human annotations. While this flexibility was true under classical NLP models, generative AI's ability to model both a broad range of possible outputs but also experience problems of "mode collapse," where answers converge on a culturally homogenous center (Hamilton 2024; Jiang et al. 2025; Zhang et al. 2025a), brings this issue more pressingly to the foreground.

Seen in this way, annotation does not serve merely as a mechanism for correcting model error, but as a form of theoretical grounding that makes visible the range of human judgment. Moving beyond single-answer frameworks becomes essential, not only for evaluating model performance, but for clarifying what it means to measure complex cultural phenomena in the first place (Figure 3).

Deterministic (plausibility-based) annotation

The most permissive annotation regime asks whether a model output is plausible: is the response reasonable enough that at least one or more informed human judges would accept it as a valid interpretation? This approach does not aim to exhaust the space of possible answers, nor to identify a central tendency, but simply to rule out responses that are clearly incoherent or misaligned with the task. In practice, plausibility can be operationalized using soft-matching criteria, where agreement with any single annotation among a trusted group is treated as sufficient evidence of alignment.

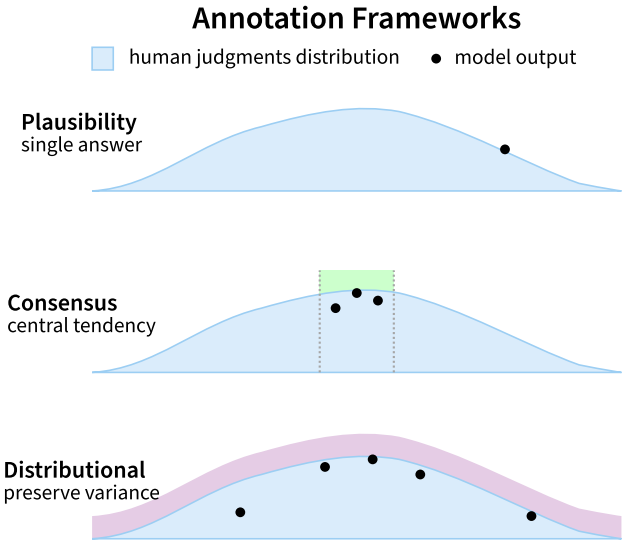


Figure 3. Different annotation regimes imply different theories of alignment between model outputs and human judgment: plausibility (any defensible output), consensus (closeness to the central tendency), and distributional alignment (overlap with the full distribution of responses).

Such an approach is appropriate when the research goal is to test whether models can generate some acceptable interpretation, rather than a complete or representative one. For example, in recent work on LLMs and topic labeling (Pham et al. 2024; Piper and Wu 2025), the aim in both cases was not to enumerate all topics a passage might evoke from all readers' perspectives, but to determine whether a model could produce at least one topic label that human annotators also recognized. Here, deterministic plausibility functions as a minimal validation threshold: the question is not whether the model captures the construct in full, but whether it produces a defensible instantiation of it.

Consensus-based annotation

A stronger form of validation evaluates whether model outputs align with the central tendencies of human judgment. Consensus-based approaches acknowledge interpretive variability, but assume that across many annotations, shared patterns will emerge that can serve as a reference point. Under this regime, models are assessed not on whether their outputs are merely acceptable, but on how closely they approximate aggregated human judgments.

Consensus can be operationalized in several ways, including semantic similarity to the centroid of human responses, agreement-based metrics such as Krippendorff's alpha, or rank-based comparisons with aggregated human preferences (e.g., Rank-Biased Overlap; Webber, Moffat, and Zobel 2010). For example, Hobson et al. (2024) and Wu and Piper (2026) compare within-group semantic similarity among human annotators' story morals to between-group similarity between human and model-generated

https://doi.org/10.1017/chr.2026.10039 Published online by Cambridge University Press

morals. They find that models consistently fall within the range of human variance, suggesting that generative models are particularly well-suited to capturing consensus interpretations – even when individual judgments diverge.

Distributional (diversity-preserving) annotation

Finally, some research problems require not convergence but coverage: the goal is to capture the range of plausible human judgments rather than collapse them into a single representative answer. Distributional approaches treat disagreement not as noise or error, but as a substantive property of the construct. This is especially important in cross-cultural, multilingual, or demographically heterogeneous contexts, where interpretive differences are theoretically meaningful rather than incidental (Zhang et al. 2025b; Zhou, Bamman, and Bleaman 2025).

A canonical example is the Social Chemistry project (Forbes et al. 2020), in which annotators rated the social acceptability of everyday behaviors using natural language explanations. Rather than enforcing a single "correct" moral judgment, the dataset explicitly retained the distribution of human responses, allowing models to be evaluated on their ability to reproduce patterns of moral diversity. Success is defined not by accuracy against a single gold label, but by fidelity to the shape and spread of human judgment – highlighting how generative AI can be evaluated as a model of variation rather than simply consensus.

In each case, then, our annotation and validation criteria need to be theoretically aligned with our stated research goals as well as the problem under study.

Prompt sensitivity and reproducibility

A central challenge for evaluating research that relies on generative AI lies in the sensitivity of model outputs to seemingly minor variations in prompts and contextual framing (Lu et al. 2022; Reynolds and McDonell 2021; Sclar et al. 2023; Webson and Pavlick 2022). Unlike traditional measurement instruments, generative AI systems do not behave as stable devices whose outputs can be assumed to reflect a fixed underlying construct. Instead, they operate as *responsive interfaces* whose behavior shifts with linguistic cues (I will get to model dependency in the next section).

One natural response to prompt sensitivity is to treat it as a selection problem: given multiple prompts or models, researchers can simply choose the configuration that best aligns with human judgments. While this strategy is often reasonable in practice, it does not eliminate the deeper epistemic issue. Different prompts can produce outputs that all appear plausible to human evaluators, yet reflect distinct assumptions about what the task is asking the model to do. In such cases, agreement with human judgments does not guarantee that a prompt is measuring the intended construct rather than a closely related alternative.

This problem is nicely illustrated in a recent large-scale replication project on LLM-based text annotation. In a systematic reanalysis of published social science studies, Baumann et al. (2025) show that small, defensible changes in prompting or model choice can reverse substantive conclusions in downstream analyses, even when models achieve high agreement with human labels. The authors refer to this as "LLM hacking" in reference to well-known problem of p-hacking in quantitative research. For example, in tasks classifying political texts as economically left- or right-leaning, different prompt paraphrases applied to the same data yield results that range from no detectable ideological difference to statistically significant effects – and in some cases even opposite directional claims. Crucially, these reversals persist when analyses are restricted to the best-performing models or to prompts selected for maximal alignment with human annotations. The issue, then, is not simply that some prompts or models perform poorly, but that multiple high-performing configurations can support incompatible interpretations of the same phenomenon. Selecting the "best" configuration may improve local accuracy, but it does not resolve the underlying instability introduced by prompt-dependent operationalizations of the construct.

Taken together, these findings suggest that AI-assisted research in the humanities should shift its evaluative focus from optimization to a concept of *robustness* (Figure 4). In this sense, robustness plays a role analogous to "convergent validity" in the social sciences: rather than relying on a single operationalization, researchers seek results that persist across different, theoretically motivated measurement choices, including prompt and model configurations. Rather than aiming to identify a single prompt or model that performs best according to human judgments, researchers should ask whether their substantive claims remain stable across a controlled range of theoretically defensible prompt and model variations. Baumann et al. (2025) recommend three key moves when using generative AI:

1. systematically test multiple prompt formulations and model configurations;
2. report results across these configurations rather than selecting a single favorable outcome;
3. document all prompt, model, and parameter choices tested.

For example, in a recent paper studying the moral history of the novel using generative AI, Piper (2025) randomly assigned all story summaries in their data to one of eight prompt variants drawn from a fully crossed 2 ×2 ×2 design, varying role framing (expert/non-expert), information ordering (where the passage is given), and question phrasing (formal/colloquial). Each variant is passed to a single model at zero temperature, with the resulting distribution of moral keywords used to capture prompt-induced interpretive variability rather than relying on a single "best" prompt. One could imagine enhancing this work with multiple temperature settings and model architectures.

Viewed in this light, prompt variation can be used as a form of epistemic stress-testing. Systematically varying prompts makes it possible to probe which aspects of a result are stable and which are contingent on particular framings. When treated deliberately, such variation can help researchers identify hidden dependencies in their analyses and distinguish between findings that reflect robust properties of the phenomenon under study and those that arise from the idiosyncracies of prompt design. Theory thus plays a dual role in AI-assisted research: it constrains the space of defensible prompt and model choices in advance, and it provides the interpretive framework needed to evaluate stability across them.

Checklist Step 2: Evaluation and validation

- **Treat prompts as measurement instruments:** Prompt design should operationalize a theoretical construct, not merely improve model performance.
- **Watch for construct underrepresentation:** Ensure the prompt does not reduce a complex construct to only its most explicit or easily measurable features.
- **Watch for construct-irrelevant variance:** Test whether outputs reflect the target construct rather than adjacent or model-friendly substitutes.

https://doi.org/10.1017/chr.2026.10039 Published online by Cambridge University Press

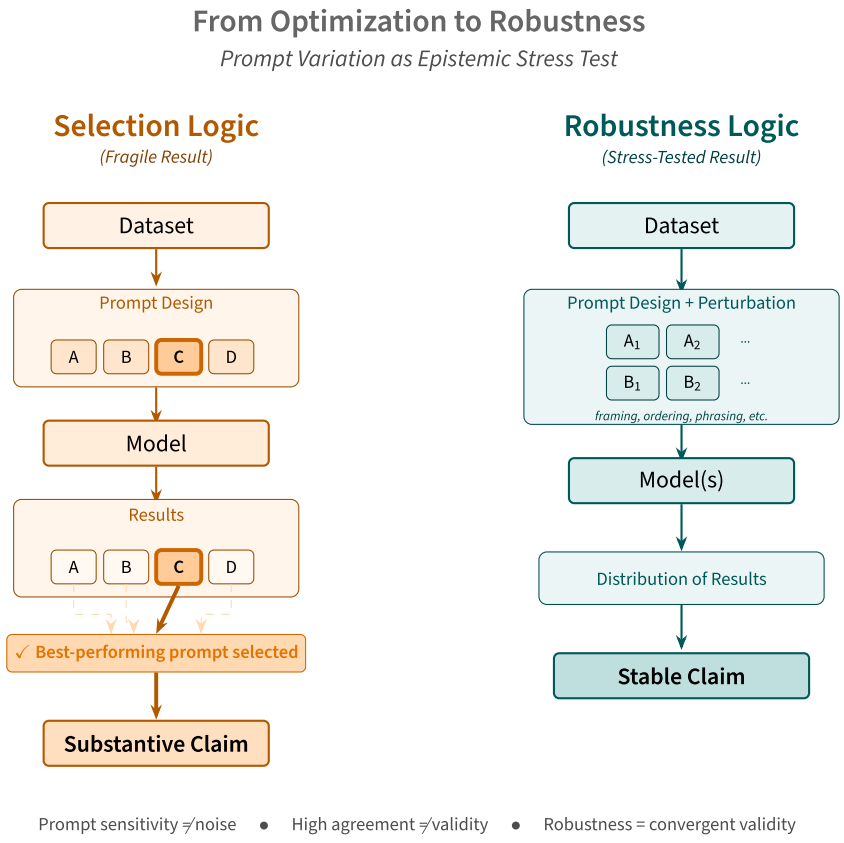


Figure 4. The left panel illustrates a selection logic in which multiple plausible prompts are tested but a single "best-performing" configuration is chosen to support a substantive claim. Although this approach may achieve high agreement with human annotations, it risks anchoring inference to a particular prompt-dependent operationalization of the construct. The right panel illustrates a robustness logic, in which prompts are systematically varied across a theoretically defensible design space and results are evaluated as a distribution rather than a single outcome. Substantive claims are warranted only when they remain stable across these perturbations. Prompt variation thus functions not as noise to be minimized but as an epistemic stress test that distinguishes fragile findings from robust ones.

Checklist Step 2: Continued

- **Align validation with the research goals:** Decide whether success means plausibility, consensus with human judgment, or preservation of the full distribution of human responses.
- **Preserve human diversity where it matters:** Treat disagreement as theoretically meaningful when studying interpretive, moral, cultural, or cross-demographic phenomena.
- **Test robustness, not just accuracy:** Evaluate whether substantive claims remain stable across theoretically defensible prompt variations, models, and parameter settings.

Step 3. Interpretation and iteration

If evaluation and validation address whether AI-assisted analyses are methodologically defensible, interpretation and iteration concern what such analyses can ultimately be said to mean and how research should proceed in light of them. Generative AI complicates interpretation by embedding assumptions upstream – through pretraining data and prompt design – while simultaneously enabling forms of rapid iteration that were previously impractical at scale. This combination creates a distinctive epistemic tension: model outputs often appear coherent or theory-confirming, even as the conditions that produce them remain only partially visible. At the same time, the ease with which analyses can be rerun or reformulated opens new opportunities for theory-driven exploration. This section argues that realizing these opportunities requires a clear account of the interpretive limits of generative AI outputs and a principled understanding of iteration not as exploratory tinkering, but as a method for testing, revising, and updating theoretical claims.

Model assumptions and interpretive limits

Interpretation in AI-assisted research cannot begin at the level of outputs alone. Generative models embed assumptions upstream, through their pretraining corpora and optimization objectives, which shape what kinds of distinctions they reliably produce and which remain under-articulated or absent (Figure 5). These assumptions do not constitute errors in need of correction, nor do they automatically invalidate model outputs. Rather, they define the conditions under which those outputs can be meaningfully interpreted (Yang et al. 2025). Understanding what a model has been trained to reproduce – and what it systematically overlooks – is therefore a prerequisite for assessing what kinds of cultural, narrative, or conceptual claims its outputs can support (Zhang et al. 2025c).

Theory-informed benchmarking

In contemporary AI research, benchmarking plays a central role in determining what models are capable of doing. Benchmarks consist of standardized tasks, datasets, and evaluation metrics used to compare models and track progress over time. They serve

https://doi.org/10.1017/chr.2026.10039 Published online by Cambridge University Press

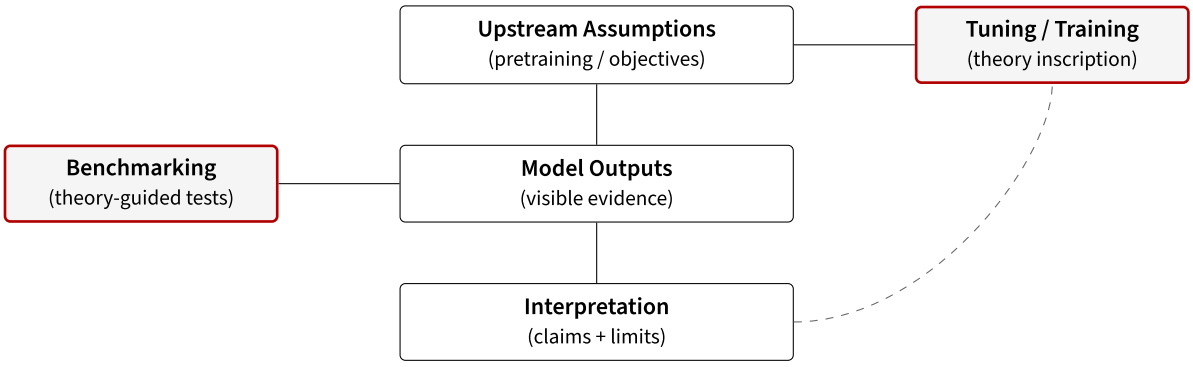


Figure 5. Generative AI outputs are interpretable only in light of upstream training assumptions. Theory-informed benchmarking (left) provides structured tests that bound what outputs can support, while theory-informed tuning (right) treats model training as a site of theoretical inscription. Iteration links interpretation back to intervention: claims motivate revised benchmarks or training signals rather than ad hoc prompt tweaking.

several important functions: guiding model development, enabling comparison across systems, and establishing shared standards for claims about performance. For many audiences, benchmarks are the primary way in which model competence is rendered legible and authoritative. At the same time, benchmarking has become a site of growing debate, as critics have noted that benchmark success can incentivize narrow forms of optimization and conflate performance on specific tasks with more general forms of understanding or reasoning (Banerjee, Agarwal, and Singh 2024; McIntosh et al. 2026; Pacchiardi et al. 2024).

These concerns are especially acute for humanities research, where interpretive judgment and contextual sensitivity (situatedness) are central to knowledge production. Most existing benchmarks are designed around tasks that favor clear labels and surface regularities, leaving aside plural interpretation and historically situated meaning. This gap creates an opening for humanities- and theory-informed interventions: not to reject benchmarking as such, but to redesign evaluative practices so that they reflect the interpretive stakes of humanistic inquiry. From this perspective, theory-informed benchmarking becomes a way to make model assumptions more visible.

A recent example can be found in Underwood et al. (2026). ChronoLogic is a benchmark designed to assess whether language models can respond like writers situated in a given historical period. The team makes several explicit theoretical choices. First, they do not separate cultural representation from general model capability. Unlike benchmarks that treat these as orthogonal, ChronoLogic includes questions from domains where historically appropriate answers have changed to varying degrees, thus testing a range of difficulty. Second, they particularize social context rather than posit unified cultures. Most questions are grounded in a specific source text and introduced through a "metadata prefix" that situates the question by date and venue of publication, along with the author's nationality, age, and profession when available. Similar to NarraBench (Hamilton, Wilkens, and Piper 2025), another recent benchmark that draws on narrative theory to construct benchmarks for narrative understanding, these initiatives show how humanities theory can guide the design of evaluation tasks that make interpretive assumptions explicit.

Theory-informed model tuning and training

Benchmarking makes visible the interpretive tendencies of generative models, but those tendencies are not accidental. They are the result of concrete design decisions, including choices about training corpora, objective functions, and reward models. Recent work has begun to explore how modifying these training stages can shift model behavior in systematic ways (Underwood, Nelson, and Wilkens 2025).

One illustration of this approach is provided by CultureLLM (Li et al. 2024), which uses fine-tuning to steer model outputs toward greater cultural difference and awareness. Starting from the observation that LLMs disproportionately encode Western and Anglophone value systems due to their pretraining corpora, CultureLLM intervenes not by expanding pretraining at scale, but by introducing a deliberately minimal and theory-grounded fine-tuning signal. Using just 50 questions from the World Values Survey as seed data, which is then augmented with synthetic semantic paraphrasing, the authors fine-tune culture-specific and unified models that reliably shift judgments in domains, such as political authority, gender norms, and social values.

Critiques of work in LLM cultural alignment focus heavily on the lack of theoretical framing for interventions, a concern articulated forcefully by Zhou, Bamman, and Bleaman (2025), who argue that much recent research operationalizes "culture" through shallow proxies rather than through a coherent account of what culture is and how it functions. They show that benchmarks and fine-tuning datasets frequently rely on coarse categories, such as nationality or survey responses, treating culture as a static bundle of facts or values rather than as a dynamic, interactional process.

For humanities research, the significance of these debates and approaches lies not in producing a "better" or more neutral model, but in recognizing model training as a space of theoretical inscription. Adjusting a model is therefore an interpretive act, one that embeds assumptions about what distinctions matter and what forms of understanding are worth reproducing.

Iteration, replication, and exploration

If generative AI complicates the tasks of evaluation and interpretation, it also enables new forms of methodological experimentation that were previously costly or impractical. LLMs dramatically lower the barriers to replicating prior studies and exploring alternative operationalizations of the same theoretical question. Tasks that once required extensive manual annotation, bespoke modeling, or large codebases can now be iterated rapidly across different prompts, models, or analytical assumptions, a fact that will only accelerate as agentic AI improves. This capacity for iteration has the potential to support more reflexive and theoretically engaged research practices, provided it is treated as a means of inquiry rather than a source of novelty or speed alone (i.e., slop).

At the same time, theory should not be seen as a static, permanent object. Theory is itself a living, dynamic framework that can

https://doi.org/10.1017/chr.2026.10039 Published online by Cambridge University Press

evolve. Although this is true across disciplines and methods, when working with generative AI this relationship becomes particularly productive: theory not only shapes how we design our studies and structure our questions, but it is also *reshaped* by the potentially unexpected patterns that models reveal.

As I have argued at length elsewhere (Piper 2019), we should think of computational modeling as a recursive process in which theoretical commitments are continuously tested, updated, or reimagined. This includes not only the more familiar processes like replication and confirmation but also the initiation of spaces for model-driven discovery and critical rethinking, traditionally called "exploratory data analysis." As Nelson (2020) writes, "Computational methods should not be seen as merely replicating existing theoretical frameworks, but as offering new opportunities to generate theory by uncovering patterns and relationships that might otherwise remain invisible." Let me close with two potentially productive ways that researchers can frame their exploratory interventions.

Feature discovery

One of the most novel ways generative AI can aid theory innovation is by helping researchers identify features that are difficult to discover with inherited computational methods because they are implicit, relational, or interpretive rather than lexical. These might include a text's *normative pressure* – where it directs sympathy, blame, approval, or suspicion; its *causal imagination* – whether events are explained through psychology, fate, institutions, environment, or social constraint; its *implied social ontology* – whether the world is organized around individuals, families, classes, nations, markets, or networks; or its *reader positioning* – whether the audience is addressed as witness, judge, accomplice, or acolyte. In this sense, generative AI does not simply automate the detection of familiar categories; it can support the discovery of intermediate theoretical features that sit between close reading and large-scale measurement.

Hypothesis generation

Beyond surfacing features, generative AI can support exploratory theory-building by suggesting plausible hypotheses grounded in large-scale textual or visual patterns (Zhou et al. 2024). By prompting models with exploratory questions on a given corpus or set of artifacts, researchers can generate candidate hypotheses that might not emerge from close reading alone. These hypotheses can then be operationalized and tested using traditional statistical methods. In this way, generative AI acts as a theory partner: not producing theories *ex nihilo*, but amplifying our ability to detect subtle regularities, contradictions, or absences that invite structured investigation, especially in under-theorized or high-dimensional domains.

Checklist Step 3: Interpretation and iteration

- **Interpret outputs through model assumptions:** Treat results as conditioned by pretraining data, optimization objectives, and model architecture.
- **Theoretically informed benchmarking:** Make cultural assumptions visible rather than treating performance as general competence.
- **Treat tuning as theory inscription:** Fine-tuning, reward modeling, and dataset construction embed theoretical assumptions about which distinctions matter.
- **Iterate, replicate, explore:** Ask whether an iteration is confirming a prior claim, testing robustness, revising an operationalization, or opening a new theoretical question.
- **Let results revise theory:** Treat unexpected patterns, failures, absences, and model disagreements as opportunities to refine constructs and generate new hypotheses.

Conclusion

This article argues that generative AI should not be understood simply as a new technical instrument for humanities research. It is better understood as a methodological environment that changes the conditions under which research questions are formed, constructs are operationalized, evidence is evaluated, and theories are revised. The central claim of a theory-first approach is that theory is not something to be added after computational analysis, but should inform the entire research process. Three concrete principles follow from this argument.

Researchers should begin by specifying the theoretical status of the model and the object of study

Is the model itself being analyzed as a cultural artifact, or is it being used as an instrument to study some other cultural phenomenon? What construct is being measured, and why does that construct matter? What is the appropriate unit of analysis and at what scale does the construct operate? Theory can help organize the place of generative AI and guard against problems of "construct drift."

Evaluation should shift from optimization to validity

The question is not simply which prompt, model, or benchmark performs best, but whether a given design captures the intended construct, preserves the relevant range of human judgment, and supports claims that remain stable across reasonable variations. In this sense, prompts should be treated as measurement instruments; human annotations should be understood as theoretical evidence rather than merely corrective labels; and prompt sensitivity should be used as an epistemic stress test rather than dismissed as noise.

Interpretation should remain bounded by the assumptions built into models while also taking advantage of the iterative possibilities they create

Model outputs are never self-evident. They must be interpreted in relation to pretraining data, optimization objectives, benchmark design, tuning procedures, and prompt context. At the same time, generative AI can support forms of theoretical discovery that are difficult to achieve with inherited methods: surfacing implicit features, revealing unexpected absences, generating hypotheses, and helping researchers revise constructs in response to new patterns.

The theory-first approach thus offers a way to normalize generative AI within humanities research without naturalizing it. It neither rejects these systems as intrinsically unreliable nor accepts them as neutral proxies for human interpretation. Instead, it treats them as situated instruments whose value depends on the clarity of the theoretical choices that guide their use. For humanities scholars, this is not a defensive position. It is an opportunity. As models increasingly mediate access to language, images, history, and culture, the humanities can contribute not

https://doi.org/10.1017/chr.2026.10039 Published online by Cambridge University Press

only new applications of AI, but better accounts of what it means to measure, interpret, and understand cultural meaning at scale.

Data availability statement. All data and code referenced in this article can be found in the cited articles.

Author contributions. A.P. is the sole author and was responsible for conceptualization, writing – original draft, and writing – review and editing.

Funding statement. This research was generously funded by the Social Sciences and Humanities Research Council of Canada.

Competing interests. The author declares none.

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

AI assistance statement. This article treats generative AI as part of the contemporary writing environment and used it accordingly: as an interlocutor for developing ideas, locating sources, drafting prose, and revising arguments. The author directed all uses of AI, made all final editorial decisions, and accepts full responsibility for the article's claims, interpretations, and errors.

References

Ahnert, Ruth, Sebastian E. Ahnert, Catherine Nicole Coleman, and Scott B. Weingart. 2020. *The Network Turn: Changing Perspectives in the Humanities*. Cambridge: Cambridge University Press.

Banerjee, Sourav, Ayushi Agarwal, and Eishkaran Singh. 2024. "The vulnerability of Language Model Benchmarks: Do They Accurately Reflect True LLM Performance?" Preprint, *arXiv:2412.03597*.

Basile, Valerio. 2020. "It's the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks [in en]." AIxIA Discussion Papers Workshop.

Baumann, Joachim, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del-Arco, Johannes B. Gruber, and Dirk Hovy. 2025. "Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation." Preprint, arXiv:2509.08825 [cs.CL]. https://arxiv.org/abs/2509.08825.

Cecire, Natalia. 2011. "Introduction: Theory and the Virtues of Digital Humanities. Winter 2011." *Journal of Digital Humanities* 1, no. 1. https://journalofdigitalhumanities.org/1-1/introduction-theory-and-the-virtues-of-digital-humanities-by-natalia-cecire/?utm_source=chatgpt.com.

Chen, Zhuang, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. "ToMBench: Benchmarking Theory of Mind in Large Language Models." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by **Lun-Wei Ku, Andre Martins, and Vivek Srikumar**, 15959–15983. Bangkok: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.847.

Cheng, Myra, Esin Durmus, and Dan Jurafsky. 2023. "Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models." Preprint, *arXiv:2305.18189*.

Christou, Despina, and Grigorios Tsoumakas. 2025. "Artificial Relationships in Fiction: A Dataset for Advancing NLP in Literary Domains." In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LATECH-CLFL 2025)*, edited by **Anna Kazantseva, Stan Szpakowicz, Stefania Degaetano-Ortlieb, Yuri Bizzoni, and Janis Pagel,** 130–47. Albuquerque, New Mexico: Association for Computational Linguistics.

Cohen, Daniel J., and Tom Scheinfeldt. 2013. *Hacking the Academy: New Approaches to Scholarship and Teaching from Digital Humanities*. Ann Arbor, MI: University of Michigan Press.

Eagleton, Terry. 2004. *After Theory*. London: Penguin UK.

Eagleton, Terry. 2011. *Literary Theory: An Introduction*. Chichester: John Wiley & Sons.

Edelstein, Dan, Paula Findlen, Giovanna Ceserani, Caroline Winterer, and Nicole Coleman. 2017. "Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project." *The American Historical Review* 122, no. 2: 400–24.

Emirbayer, Mustafa, and Jeff Goodwin. 1994. "Network Analysis, Culture, and the Problem of Agency." *American Journal of Sociology* 99, no. 6: 1411–54.

Forbes, Maxwell, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. "Social Chemistry 101: Learning to Reason about Social and Moral Norms." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by **Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu,** 653–70. Online: Association for Computational Linguistics.

Frenda, Simona, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist Approaches to Natural Language Processing: A Survey [in en]. *Language Resources and Evaluation* 59, no.2: 1719–46. Accessed: January 6, 2026. https://doi.org/10.1007/s10579-024-09766-4.

Hamilton, Sil. 2024. "Detecting Mode Collapse in Language Models Via Narration." In *Proceedings of the First Edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, edited by **Antonio Valerio Miceli-Barone, Fazl Barez, Shay B. Cohen, Elena Voita, Ulrich Germann, and Michal Lukasik,** 65–72. St. Julian's, Malta: Association for Computational Linguistics.

Hamilton, Sil, Rebecca M. M . Hicke, David Mimno, and Matthew Wilkens. 2025. "A City of Millions: Mapping Literary Social Networks at Scale." Preprint, *arXiv:2502.19590*.

Hamilton, Sil, Matthew Wilkens, and Andrew Piper. 2025. "NarraBench: A Comprehensive Framework for Narrative Benchmarking." Preprint, *arXiv:2510.09869*.

Hobson, David, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. "Story Morals: Surfacing Value-Driven Narrative Schemas Using Large Language Models." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, edited by **Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen,** 12998–13032. Miami, Florida, USA: Association for Computational Linguistics.

Jacobsen, Mia, Yuri Bizzoni, Pascale Feldkamp Moreira, and Kristoffer L. Nielbo. 2024. "Patterns of Quality: Comparing Reader Reception across Fanfiction and Commercially Published Literature." In *Proceedings of the Computational Humanities Research Conference 2024 (CHR 2024)*, Vol. 3834, 718–39. Aarhus: CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3834/paperXX.pdf.

Jiang, Liwei, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025. "Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and beyond)." In *Advances in Neural Information Processing Systems*, Vol. 38, edited by **Danielle Belgrave, Cheng Zhang, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, and Nancy Chen.** Red Hook, NY: Curran Associates, Inc.

Karjus, Andres. 2025. "Machine-Assisted Quantitizing Designs: Augmenting Humanities and Social Sciences with Artificial Intelligence." *Humanities and Social Sciences Communications* 12, no. 1: 1–18.

Kim, Alexander. 2024. "Sequencing Literary Reception: Trajectories of Russian Novels over the Afterlife Course." In *Proceedings of the Computational Humanities Research Conference 2024 (CHR 2024)*, Vol. 3834, 982–98. Aarhus: CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3834/paper46.pdf.

Ladd, John R. 2021. "Imaginative Networks: Tracing Connections among Early Modern Book Dedications." *Journal of Cultural Analytics* 6, no. 1. https://doi.org/10.22148/001c.21993.

Li, Cheng, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. "CultureLLM: Incorporating Cultural Differences into Large Language Models." *Advances in Neural Information Processing Systems*, Vol. 37, edited by **Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang,** 84799–84838. Red Hook, NY: Curran Associates, Inc.

Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. "Fantastically Ordered Prompts and where to Find them: Overcoming Few-Shot Prompt Order Sensitivity." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by **Smaranda Muresan, Preslav Nakov, and Aline**

Villavicencio, 8086–98. Dublin, Ireland: Association for Computational Linguistics.

McCarthy, Arya D., and Giovanna Maria Dora Dore. 2023. "Theory-Grounded Computational Text Analysis." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), edited by **Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki,*** 1586–94. Toronto, Canada: Association for Computational Linguistics.

McIntosh, Timothy R., Teo Susnjak, Nalin Arachchilage, Tong Liu, Xu Dan, Paul Watters, and Malka N. Halgamuge. 2026. "Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence." *IEEE Transactions on Artificial Intelligence* 7, no. 1: 22–39.

Merton, Robert K. 1949. [1968]. "On Sociological Theories of the Middle Range." In *Social Theory and Social Structure*, edited by **Robert K. Merton,** 39–72. New York: The Free Press.

Nelson, Laura K. 2020. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research* 49, no. 1: 3–42.

Pacchiardi, Lorenzo, Marko Tesic, Lucy G Cheke, and Josè Hernández-Orallo. 2024. "Leaving the Barn Door Open for Clever Hans: Simple Features Predict LLM Benchmark Answers." Preprint, *arXiv:2410.11672*.

Pham, Chau, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. "Topicgpt: A Prompt-Based Topic Modeling Framework." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, edited by **Kevin Duh, Helena Gomez, and Steven Bethard,** 2956–84. Mexico City, Mexico: Association for Computational Linguistics.

Pianzola, Federico. 2025. *Digital Social Reading: Sharing Fiction in the Twenty-First Century*. Cambridge, MA: MIT Press.

Piper, Andrew. 2019. *Enumerations: Data and Literary Study*. Chicago, IL: University of Chicago Press.

Piper, Andrew. 2025. "Towards a Perspectival Moral History of the Novel Using LLMS." *Journal of Computational Literary Studies* 4, no. 1; 1–19.

Piper, Andrew, and Sophie Wu. 2025. "Evaluating Large Language Models for Narrative Topic Labeling." In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, edited by **Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar,** 281–91. Albuquerque, USA: Association for Computational Linguistics.

Piper, Andrew, Michael Xu, and Derek Ruths. 2024. "The Social Lives of Literary Characters: Combining Citizen Science and Language Models to Understand Narrative Social Networks." In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, edited by **Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni,** 472–82. Miami, USA: Association for Computational Linguistics.

Radford, Jason, and Kenneth Joseph. 2020. "Theory in, Theory out: The Uses of Social Theory in Machine Learning for Social Science." *Frontiers in Big Data* 3: 18.

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes." *EPJ Data Science* 5, no. 1: 1–12.

Rebora, Simone, Peter Boot, Federico Pianzola, J. Brigitte Gasser, Berenike Herrmann, Maria Kraxenberger, Moniek M. Kuijpers, Gerhard Lauer, Piroska Lendvai, Thomas C . Messerli, and Pasqualina Sorrentino 2021. "Digital Humanities and Digital Social Reading." *Digital Scholarship in the Humanities* 36, no. Supplement_2: ii230–50.

Rebora, Simone, and Gabriele Vezzani. 2024. "Models of Literary Evaluation and Web 2.0. An Annotation Experiment with Goodreads Reviews." In *Proceedings of the Computational Humanities Research Conference 2024 (CHR 2024)*, Vol. 3834, Article no. 46. Aarhus: CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3834/paper46.pdf.

Reynolds, Laria, and Kyle McDonell. 2021. "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm." In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, edited by **Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, and Takeo Igarashi,** 1–7. New York, NY: Association for Computing Machinery.

Ryan, Yann C., and Sebastian E. Ahnert. 2021. "The Measure of the Archive: The Robustness of Network Analysis in Early Modern Correspondence." *Journal of Cultural Analytics* 6, no. 3: 57–88.

Schmidt, Ben. 2011. "Theory First." Blog post on *Sapping Attention*. Accessed: 2025-01-27. http://sappingattention.blogspot.com/2011/11/theory-first.html.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. "*Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying About Prompt Formatting*." Preprint, *arXiv:2310.11324*.

Smith, Gregory T. 2005. "On Construct Validity: Issues of Method and Measurement." *Psychological Assessment* 17, no. 4: 396.

Strauss, Milton E., and Gregory T. Smith. 2009. "Construct Validity: Advances in Theory and Methodology." *Annual Review of Clinical Psychology* 5, no. 1: 1–25.

Tambuscio, Marcella, and Georg Vogeler. 2025. "Bridging Semantics and Structure: A Typed Prosopographical Network of Maximilian I's Court." *Anthology of Computers and the Humanities* 3: 131–41.

Underwood, Ted. 2020. "Critical Response II. The Theoretical Divide Driving Debates about Computation." *Critical Inquiry* 46, no. 4: 900–12.

Underwood, Ted, Sarah Griebel, Laura Nelson, Ziliang Qiu, Edwin Earl Roland, Wenyi Shang, and Matt Wilkens. 2026. "*ChronoLogic: A Benchmark for Historical Language Models*."

Underwood, Ted, Laura McGrath, Richard Jean So, and Chad Wellmon. 2022. "Culture, Theory, Data: An Introduction." *New Literary History* 54, no. 1: 519–30.

Underwood, Ted, Laura K. Nelson, and Matthew Wilkens. 2025. "*Can Language Models Represent the Past Without Anachronism?*" Preprint, arXiv: 2505.00030 [cs.CL]. https://arxiv.org/abs/2505.00030.

Von Heydebrand, Renate, and Simone Winko. 2008. "The Qualities of Literatures: A Concept of Literary Evaluation in Pluralistic Societies." In *The Quality of Literature: Linguistic Studies in Literary Evaluation*, edited by **Willie van Peer,** 223–39. Amsterdam: John Benjamins Publishing Company.

Warwick, Claire. 2015. "Building Theories or Theories of Building? A Tension at the Heart of Digital Humanities." In *A New Companion to Digital Humanities*, edited by **Susan Schreibman, Ray Siemens, and John Unsworth,** 538–52. Chichester, UK: John Wiley & Sons.

Webber, William, Alistair Moffat, and Justin Zobel. 2010. "A Similarity Measure for Indefinite Rankings." *ACM Transactions on Information Systems* 28, no. 4: 1–38.

Webson, Albert, and Ellie Pavlick. 2022. "Do Prompt-Based Models Really Understand the Meaning of their Prompts?" In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by **Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz,** 2300–2344. Seattle, United States: Association for Computational Linguistics.

Windhager, Florian, and Eva Mayr. 2024. "Digital Humanities and Distributed Cognition: From a Lack of Theory to its Visual Augmentation." *Journal of Cultural Analytics* 7, no. 4: 1–34.

Wu, Sophie, and Andrew Piper. 2026. "Lessons Without Borders? Evaluating Cultural Alignment of LLMs Using Multilingual Story Moral Generation." Preprint, *arXiv:2604.08797*.

Yang, Xinyi, Runzhe Zhan, Derek F Wong, Shu Yang, Junchao Wu, and Lidia S Chao. 2025. "*Rethinking Prompt-Based Debiasing in Large Language Models*." Preprint, *arXiv:2503.09219*.

Ying, Lance, Katherine M. Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2025. "*On Benchmarking Human-Like Intelligence in Machines*." Preprint, arXiv:2502.20502 [cs]. Accessed April 4, 2025. https://doi.org/10.48550/arXiv.2502.20502.

Yu, Ze, and Federico Pianzola. 2024. Across the Pages: A Comparative Study of Reader Response to Web Novels in Chinese and English on Qidian and Webnovel. In *Proceedings of the Computational Humanities Research Conference 2024 (CHR 2024)*, Vol. 3834, 322–33. Aarhus: CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3834/paperXX.pdf.

Zhang, Jiayi, Simon Yu, Derek Chong, Anthony Sicilia, Michael R . Tomz, Christopher D Manning, and Weiyan Shi. 2025a. "*Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity*." Preprint, *arXiv:2510.01171*.

https://doi.org/10.1017/chr.2026.10039 Published online by Cambridge University Press

Zhang, Lily Hong, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Bouaziz Wassim, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. 2025b. "Cultivating Pluralism in Algorithmic Monoculture: The Community Alignment Dataset." Preprint, arXiv:2507.09650 [cs]. Accessed August 1, 2025. https://doi.org/10.48550/arXiv.2507.09650.

Zhang, Qinyan, Xinping Lei, Ruijie Miao, Yu Fu, Haojie Fan, Le Chang, Jiafan Hou, Dingling Zhang, Zhongfei Hou, Ziqiang Yang, Ziqiang Yang, Changxin Pu, Fei Hu, Jingkai Liu, Mengyun Liu, Yang Liu, Xiang Gao, Jiaheng Liu, Tong Yang, Zaiyuan Wang, Ge Zhang, and Wenhao Huang. 2025c. "*Inverse IFEval: Can LLMs Unlearn Stubborn Training Conventions to Follow Real Instructions?*" Preprint, *arXiv:2509.04292*.

Zhou, Naitian, David Bamman, and Isaac L Bleaman. 2025. "*Culture is not Trivia: Sociocultural Theory for Cultural NLP*." Preprint, *arXiv:2502.12057*.

Zhou, Yangqiaoyu, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. "Hypothesis Generation with Large Language Models." In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, edited by **Lotem Peled-Cohen, Nitay Calderon, Shir Lissak, and Roi Reichart**, 117–39. Miami, FL: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.nlp4science-1.10.

https://doi.org/10.1017/chr.2026.10039 Published online by Cambridge University Press