Reading with Language Models*

Erik Fredner

Accepted manuscript of the article forthcoming as:
Fredner, Erik. "Reading with Language Models." In "Cultural AI," special issue, *MFS Modern Fiction Studies* 73, no. 1 (forthcoming).

Abstract

Literary scholars' justified opposition to their students' use of language models (LMs) to avoid the reading, writing, and thinking crucial to the teaching of literature has obscured how LMs can advance literary scholarship. I argue that literary scholars using LMs have already begun to construct an implicit norm that gives greater license to read with LMs when they adopt what Louise Rosenblatt terms the efferent stance toward literary texts, but less license when they adopt what she terms the aesthetic stance. Rosenblatt's distinction may help mediate the conflict between literary scholars who strongly oppose LMs and those who are curious to know what value they may have for literary studies by demarcating limits to LMs' use within the discipline's core practices of reading.

Introduction

Based on the evidence available online, literary scholars appear to be largely united in their opposition to language models (LMs) like ChatGPT. In this, surveys suggest that they are no different from the majority of Americans.[1] There are many good reasons for opposition.[2] An important reason within literary studies is that instructors have observed their students using LMs to avoid the reading, writing, and thinking essential to the teaching of literature.[3]

However, justified opposition to the ways some students have been using LMs obscures how these models can be used to advance literary scholarship. Used with caution by literary scholars in ways that I

*Copyright © 2027 Purdue University. This article will appear in *MFS Modern Fiction Studies*, Volume 73, Issue 1, 2027 as part of a special issue on the topic of Cultural AI edited by Richard Jean So and Aarthi Vadde. Published by Johns Hopkins University Press.

[1] See Michelle Faverio and Emma Kikuchi, "What the Data Says about Americans' Views of Artificial Intelligence," Artificial Intelligence in *Pew Research Center*, 2026.

[2] See e.g., Timnit Gebru and Émile P. Torres, "The TESCREAL Bundle: Eugenics and the Promise of Utopia through Artificial General Intelligence," *First Monday*, ahead of print, April 2024, https://doi.org/10.5210/fm.v29i4.13636; Emily M. Bender and Alex Hanna, *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*, First edition (HarperCollins, 2025); and Gael Varoquaux, Sasha Luccioni, and Meredith Whittaker, "Hype, Sustainability, and the Price of the Bigger-Is-Better Paradigm in AI," *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAccT '25, June 2025, 61–75, https://doi.org/10.1145/3715275.3732006.

[3] See Beth McMurtrie, "The Reading Struggle Meets AI," News in *The Chronicle of Higher Education*, https://www.chronicle.com/article/the-reading-struggle-meets-ai, 2025; and Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, et al., *Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task*, arXiv:2506.08872, arXiv, 2025, https://doi.org/10.48550/arXiv.2506.08872.

will describe, reading with LMs can also *occasion* reading, writing, and thinking.[4] While most applications of LMs to texts of interest to literary studies thus far have occurred within digital humanities and adjacent fields, the use of LMs for literary scholarship is not limited to such work. In fact, their use can be so similar to the everyday work of literary studies that many scholars may not realize that they have already begun to use some of them.

Broadly, I will argue that scholars engaged in this work have begun to construct an implicit norm about how to read with LMs. Literary scholars grant greater license to read with LMs when they adopt what Rosenblatt terms the "efferent stance" toward a text, and less license when they adopt what she terms the "aesthetic stance." A reading on the efferent end of Rosenblatt's continuum "is centered predominantly on what is to be extracted and retained after the reading event," whereas a reading on the aesthetic end "adopts an attitude of readiness to focus attention on what is being lived through during the reading event."[5] This emerging norm has rarely been theorized in these terms.[6] While both stances are essential to the discipline's "core practice" of close reading, literary studies also depends upon many other reading practices that receive less attention.[7] These include but are not limited to browsing, skimming, and scanning texts. Rosenblatt's distinction may help mediate the conflict glossed above between literary scholars who strongly oppose the use of LMs within literary studies and those who are curious to know what value they may have for the field by demarcating limits to LMs' use within the discipline's core practices.

I will make the case for this distinction through many examples that fruitfully apply LMs to literary texts, literary criticism, and paraliterary texts of interest to literary studies. By reviewing the tasks for which these models have been used, I show how they can be used persuasively despite their weaknesses. For example, hallucinations—LMs' best-known weakness—cannot be eliminated; they are inherent to the autoregressive structures of these models.[8] Hallucinations therefore pose a significant risk to the use of LMs for research. At the same time, researchers have used and invented techniques to mitigate hallucinations substantially. To claim that LMs are useless because hallucinations are inevitable is analogous to claiming that GPS turn-by-turn navigation is useless because it does not correctly guide drivers to their destination every time. In both cases, people figure out how to use imperfect technologies in ways that account for their shortcomings.

A lot of the research that has already applied LMs to literary texts and questions has been published and presented in venues unfamiliar to most literary scholars. Because of this, I review some of it here in a way that attempts to make its relevance to literary studies writ large clear. I discuss reading with LMs in several distinct contexts: How scholars have already applied LMs to literary texts and literary criticism for tasks including annotation, summarization, imitation, and information retrieval; how literary scholars have and will use vector search and retrieval augmented generation to complement their browsing, skimming, and scanning, especially of secondary sources; and how I have used LMs to classify

[4]On LMs and the question of "reading," see Melanie Mitchell and David C. Krakauer, "The Debate over Understanding in AI's Large Language Models," *Proceedings of the National Academy of Sciences* 120, no. 13 (2023): e2215907120, https://doi.org/10.1073/pnas.2215907120.

[5]"The Transactional Theory of Reading and Writing," in *Theoretical Models and Processes of Literacy*, 7th ed. (Routledge, 2018), 458.

[6]For a previous use of Rosenblatt's distinction within digital humanities more generally, see Tom Liam Lynch, "Electrical Evocations: Computer Science, the Teaching of Literature, and the Future of English Education," *English Education* 52, no. 1 (2019): 15–37, https://doi.org/10.58680/ee201930312.

[7]John Guillory, *On Close Reading* (The University of Chicago Press, 2024), 3.

[8]Mark Russinovich, Ahmed Salem, Santiago Zanella-Béguelin, and Yonatan Zunger, "The Price of Intelligence," *Commun. ACM*, ahead of print, August 2025, https://doi.org/10.1145/3749447.

and extrude data from paraliterary texts that contain unstructured information about authors, works, and their relative positions in literary canon. All of these uses of LMs occur closer to the efferent end of Rosenblatt's continuum. However, information they reveal about the texts thus analyzed can also occasion readings closer to the aesthetic end.

My focus on how literary studies has and might read with LMs differs from many recent discussions about LMs and the humanities. Much of that work has focused on the relationship between the questions, knowledge, and methods of the humanities, and the creation, use, and effects of LMs in the world.[9] Others have focused on the interrelationships between university administrators, university workers, and the increasing (and increasingly worrisome) proportion of university budgets diverted to big tech, including but not limited to contracts for LMs.[10]

I do not focus on the field's use of these models to add to the AI hype. Rather, I do so in opposition to a weak critique. Too many arguments against the use of LMs in literary studies proceed from the supposition that they are not or cannot be useful to the discipline because they have been harmful in the classroom. However true this argument may be of the classroom, it is demonstrably false with respect to research. Ethical arguments against LMs—how they disempower labor, reproduce biases including but not limited to racism and sexism, rely on copyrighted material without compensating its creators, accelerate surveillance by the state and by capital, etc.—are far stronger. Claims made for or against LMs on grounds of their usefulness demand serious engagement with both their capacities and disciplinary norms governing their use. While I cannot discuss the former here as the state of the art expires weekly, I will attempt to describe how I see the latter emerging.

Generating readings with LMs

While this essay discusses scholars using LMs to advance literary research, it does not advocate for LMs independently generating readings of literary texts. A common rejoinder to LM-generated text captures the prevailing view of this issue: "Why should I bother to read something that no one bothered to write?"[11] My aim in this section is to articulate the assumptions on which this question rests. A strong version of its argument must hold true in a hypothetical future where LM-generated literary scholarship

[9]e.g., Aarthi Vadde, "Inside and Outside the Language Machines," *PMLA* 139, no. 3 (2024): 553–58, https://doi.org/10.1632/S0030812924000579; Ruha Benjamin, "The New Artificial Intelligentsia," in *Los Angeles Review of Books*, https://larevieweofbooks.org/article/the-new-artificial-intelligentsia, 2024; Katherine Elkins, "A(I) University in Ruins: What Remains in a World with Large Language Models?" *PMLA* 139, no. 3 (2024): 559–65, https://doi.org/10.1632/S0030812924000543; Lauren Klein, Meredith Martin, André Brock, et al., *Provocations from the Humanities for Generative AI Research*, arXiv:2502.19190, arXiv, 2025, https://doi.org/10.48550/arXiv.2502.19190; Drew Hemment and Cody Kommers, *Doing AI Differently: Rethinking the Foundations of AI via the Humanities* (Zenodo, 2025); and Edwin Roland and Richard Jean So, *Generative AI & Fictionality: How Novels Power Large Language Models*, arXiv:2603.01220, arXiv, 2026, https://doi.org/10.48550/arXiv.2603.01220.

[10]See e.g., Matthew Kirschenbaum and Rita Raley, "AI and the University as a Service," *PMLA* 139, no. 3 (2024): 504–15, https://doi.org/10.1632/S003081292400052X; Annie McClanahan and Louise McCune, "Ed Tech," in *University Keywords*, ed. Andy Hines, Critical University Studies (Johns Hopkins University Press, 2025); and Matt Seybold, "Against Technofeudal Education," Substack Newsletter, in *The American Vandal*, 2025. For the view from the professioriate, see *Artificial Intelligence and Academic Professions* (American Association of University Professors, 2025). On AI and labor more generally, see Matteo Pasquinelli, *The Eye of the Master: A Social History of Artificial Intelligence* (Verso, 2023).

[11]There is little doubt that LM-generated research has already been submitted and may well be published, if it has not been already. For example, Weixin Liang, Yaohui Zhang, Zhengxuan Wu, et al., "Quantifying Large Language Model Usage in Scientific Papers," *Nature Human Behaviour* 9, no. 12 (2025): 2599–609, https://doi.org/10.1038/s41562-025-02273-8 show that words disproportionately favored by LMs like *pivotal* and *intricate* began appearing much more often in scientific abstracts after the release of ChatGPT.

is indistinguishable from the work of experts. Today, one can argue for the superiority of expert work on the merits. But it is more useful to contemplate why this judgment would persist even if LM-generated text were one day indistinguishable from expert work. That answer turns on reading and writing as embodied experiences.

A thought experiment will help to ground this point. Jorge Luis Borges's character Pierre Menard wants to live his life such that he will "produce a number of pages which coincided—word for word and line for line—with those of Miguel de Cervantes."[12] In the judgment of the narrator of Borges's story, "The Cervantes text and the Menard text are verbally identical, but the second is almost infinitely richer. (More *ambiguous*, his detractors will say—but ambiguity is richness)."[13] The same words, but not the same meaning. Menard's seem "richer" because of his quixotism.

Now, imagine if a future LM could be trained on a scholar's prior reading and writing such that it could generate a new reading of a new text that is identical to one that scholar had independently written. Even though these texts would be identical, most literary scholars would not regard them as being of equal value because they emerged from different contexts.

The scholar's version would be considered more valuable because it testifies to the embodied experiences of specific texts encountered by specific readers at specific times. A version of this position has been central to feminist scholarship, Black studies, queer theory, disability studies, and many other fields. As Paula Moya put it recently, "Because a work of literature is only actualized in the process of being read, it can never be the same for all readers or even for the 'same' reader over time."[14] And one reader's embodied experience of one text at the moment of its reading—what Derek Attridge calls "the literary event"—cannot be modeled, either.[15]

Writing after reading discloses aspects of that experience that the writer hopes to make meaningful to other readers. As I.A. Richards put it a century ago, "Criticism is the endeavour to discriminate between experiences."[16] A century later, Lauren Klein et al. said, "Models make words, but people make meaning."[17] Dan Sinykin and Johanna Winant emphasize embodiment in close reading by narrating how the method works with anaphoric emphasis on *you*: "…you start with someone else's words; you see something in those words that means something to you…And now you might look up from the text because you want to show what's happened to someone else. You want to explain something to another person: your own reader."[18] N. Katherine Hayles has made the same point in the context of LMs: "Literary criticism…has always worked from one customary presupposition: that the texts it interrogates have been written by humans with language processed by human brains."[19] What Hayles says here of literary texts applies equally to scholarship. LM-generated texts do not imply that embodied work of observation, interpretation, and communication. LM outputs mean differently without this context, just as Menard's *Quixote* means differently than Cervantes's.

Leif Weatherby might characterize this distinction as "remainder humanism." It is a "remainder" in

---

[12]"Pierre Menard, Author of the *Quixote*," in *Collected Fictions*, trans. Andrew Hurley (Allen Lane The Penguin Press, 1999), 91.

[13]"Pierre Menard, Author of the *Quixote*," 94.

[14]"Some Propositions on Close Reading," *Symploke* 32, no. 1 (2024): 359.

[15]*The Singularity of Literature*, Routledge Classics (Routledge, 2017), 84–85.

[16]*Principles of Literary Criticism*, International Library of Psychology, Philosophy, and Scientific Method (K. Paul, Trench, Trubner, & Co., ltd.; Harcourt, Brace & Co., inc, 1925) viii.

[17]*Provocations from the Humanities for Generative AI Research*, 1.

[18]*Close Reading for the Twenty-First Century*, Skills for Scholars (Princeton University Press, 2025), 1.

[19]*Bacteria to AI: Human Futures with Our Nonhuman Symbionts* (The University of Chicago Press, 2025), 139.

the sense that it defines as human the "ever shrinking area of things that 'computers can't do'."[20] Weatherby is right that defining the human in the negative is a losing game. Recent studies have shown that the difference between human and machine performance on measurable aspects of close reading is surprisingly small: no statistically significant difference was observed in the grades assigned to essays about Old English poetry between Oxford University students and GPT-4;[21] GPT-4 approximated or outperformed literary scholars at correctly identifying poetic forms of unlabeled poems;[22] a small LM performed better than the average of human evaluators on college-level multiple-choice close reading questions;[23] and LM-generated interpretations of texts helped people answer close reading questions about those texts more accurately than they otherwise would have.[24] This goes beyond close reading quizzes to other supposedly distinctive human abilities. For example, one recent study found that LMs matched humans on questions designed to test theory of mind.[25] Even if LMs match or exceed measurable expert performance on such tasks, arguing from embodied experience means that context need not be evident in the text in order to make a meaningful difference.

However, arguing from embodied experience does cut against key insights from the conflict between phenomenology and structuralism. Jacques Derrida would have challenged the notion that the absence of embodied context from LM-generated text differentiates it from any other instance of writing because "a written sign carries with it a force that breaks with its context, that is, with the collectivity of presences organizing the moment of its inscription."[26] For Derrida, what applies to writing also applies to "the entire field of what philosophy would call experience, even the experience of being," challenging whether this distinction is any distinction at all.[27] With respect to LMs, Hayles calls this "the null strategy,"[28] which she opposes because it relies on "the incorrect assumption that [LM-generated texts] display interiority and subjectivity."[29] Widespread uptake of the rejoinder "Why should I bother to read something that no one bothered to write?" suggests that Derrida's argument is being reconsidered in the context of LMs. This would be a striking turn of the dialectic, given that, as Ted Underwood has argued, contemporary LMs themselves represent "the empirical triumph of theory," especially structuralism. As the philosopher Alva Noë recently put it, "Computers, however vitally important these may become as technological extensions of our work, never enter into human being."[30]

LMs cannot independently generate readings that can answer the question "Why should I bother to

[20] *Language Machines: Cultural AI and the End of Remainder Humanism*, Posthumanities 74 (University of Minnesota Press, 2025), 37.

[21] T. Revell, W. Yeadon, G. Cahilly-Bretzin, et al., "ChatGPT versus Human Essayists: An Exploration of the Impact of Artificial Intelligence for Authorship and Academic Integrity in the Humanities," *International Journal for Educational Integrity* 20, no. 1 (2024): 1–19, https://doi.org/10.1007/s40979-024-00161-8.

[22] Melanie Walsh, Anna Preus, and Maria Antoniak, *Sonnet or Not, Bot? Poetry Evaluation for Large Models and Datasets*, arXiv:2406.18906, arXiv, 2024, https://doi.org/10.48550/arXiv.2406.18906.

[23] Peiqi Sui, Juan Diego Rodriguez, Philippe Laban, et al., *KRISTEVA: Close Reading as a Novel Task for Benchmarking Interpretive Reasoning*, arXiv:2505.09825, arXiv, 2025, https://doi.org/10.48550/arXiv.2505.09825.

[24] Jiayin Zhi, Hoyt Long, Richard Jean So, and Mina Lee, *What Does AI Do for Cultural Interpretation? A Randomized Experiment on Close Reading Poems with Exposure to AI Interpretation*, 2026, https://doi.org/10.1145/3772318.3791727.

[25] James W. A. Strachan, Dalila Albergo, Giulia Borghini, et al., "Testing Theory of Mind in Large Language Models and Humans," *Nature Human Behaviour* 8, no. 7 (2024): 1285–95, https://doi.org/10.1038/s41562-024-01882-z.

[26] "Signature Event Context," in *Limited Inc* (Northwestern University Press, 1988), 9.

[27] "Signature Event Context," 9.

[28] *Bacteria to AI*, 147.

[29] *Bacteria to AI*, 144.

[30] *The Entanglement: How Art and Philosophy Make Us What We Are* (Princeton University Press, 2023), 160.

read something that no one bothered to write?" because no one bothered to write them. Embodied experiences of reading and writing appear to be prerequisites for such work, albeit ones that did not need to be named before it became possible that a reading could be created without them. However, LMs need not generate close readings independently to be useful to close readers.[31] The remainder of this essay will show how LMs are already being used to create evidence that supports literary scholarship.

Reading with LMs

By now, researchers have already been reading literature with LMs for several years. For example, the National Endowment for the Humanities first funded the AI for Humanists project, which seeks to help position humanists "to make use of—and to critique" LMs, in 2021.[32] However, much of this work has been published in venues far afield from the reading of most literary scholars. In this section, I describe some recent research that exemplifies why and how literary studies has and will read with LMs.[33] While these studies tend to foreground computational and statistical methods, I have prioritized discussing studies that have co-authors who are literary scholars, who are well aware of where and why their approaches break from traditional literary scholarship. I organize this brief discussion around some of the key tasks for which scholars have used LMs with literary texts.

One such task is annotation. While different in form, annotation with LMs is identical in spirit to marking up a book with any systematic approach to marginalia, highlight colors, or page flags. The difference is that digital annotations are computationally tractable, so they can be used to count examples or retrieve text associated with one or more annotations. For instance, Andrew Piper and Sunyam Bagga use GPT-4 to annotate passages from many kinds of prose narratives to identify which predetermined characteristics, if any, a given passage possesses (e.g., does it contain specific markers of time?). They find that the model annotates passages in ways that tend to agree with human annotations of the same passages for the same features. Similarly, Catherine Yeh, Tara Menon, et al. use LMs to annotate where characters are geographically located at specific points in the narratives of novels, plays, and epic poems. They use the resulting data to visualize the relationship between time, space, and characters over an entire text, such that they can show which characters are at the Bennetts' and which at Netherfield across narrative time in *Pride and Prejudice*. Their approach also demonstrates how LM annotations can be reviewed and corrected by researchers after generation, a research process usually referred to as "human-in-the-loop." Haaris Mian, Melanie Subbiah, Sharon Marcus, et al. use LMs to operationalize and extend Alex Woloch's account of character. Where previous such analyses primarily focused on computationally tractable elements (like direct mentions of a character's name), they extend this model to more complex categories that they use LMs to annotate, including interiority, action, discussion by other characters, and discussion by the narrator for each character. Their model uses the

[31] In some fields, LMs have been used to simulate feedback from peer reviewers and editors. In the sciences, see Weixin Liang, Yuhui Zhang, Hancheng Cao, et al., "Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis," *NEJM AI* 1, no. 8 (2024), https://doi.org/10.1056/AIoa2400196. In creative writing, see Katy Ilonka Gero, Tao Long, and Lydia B. Chilton, "Social Dynamics of AI Support in Creative Writing," *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '23, April 2023, 1–15, https://doi.org/10.1145/3544548.3580782.

[32] "The AI for Humanists Project," in *AI for Humanists*, http://www.bertforhumanists.org//, 2025.

[33] For a review that covers the humanities and social sciences more broadly, see Andres Karjus, "Machine-Assisted Quantitizing Designs: Augmenting Humanities and Social Sciences with Artificial Intelligence," *Humanities and Social Sciences Communications* 12, no. 1 (2025): 277, https://doi.org/10.1057/s41599-025-04503-w.

sum of these factors as an index of how major or minor individual characters are.[34]

In these examples and others like them, researchers use LMs to annotate passages from literary texts in order to extract data about structural or formal features that can be used for other analyses—a quintessentially efferent reading practice. As these examples also suggest, one standard approach to assessing the accuracy of LM annotations compares them to identical annotations created and validated by experts on a representative sample of the texts to be annotated. Such annotation tasks often find that LM annotation is faster and cheaper than human annotation, while producing results that evaluations suggest are as good or better than those produced by humans. Their speed and performance also makes it possible to do this kind of annotation at otherwise unimaginable scales.[35] While working at larger scales does not necessarily confer advantages for close reading (though it can), it clearly makes a difference for fields of literary study that attempt to make sense of larger bodies of texts, such as literary history, genre theory, and stylistics.

Summarization is another task for which researchers have applied LMs to literary texts. Cleanth Brooks's heresy of paraphrase notwithstanding, summarizing texts can be useful for researchers who wish to identify passages that discuss similar themes or topics using dissimilar language. When literary works address subjects indirectly, metaphorically, or through omission, computational approaches like counting words can conceal similarities between passages. For example, as Toni Morrison emphasizes in *Playing in the Dark*, American literature's pervasive "Africanist presence" is evinced through "significant and underscored omissions."[36] Building on such observations, Lucy Li et al. use LMs to summarize passages from fiction, asking the LM to "tell" rather than "show" what happens in each passage. They use the resulting summaries as inputs for topic modeling, a natural language processing technique that predates LMs for characterizing the topics that documents discuss based on the underlying distributions of words in a corpus.[37] Eschewing topic modeling, Andrew Piper and Sophie Wu have used LMs to annotate narrative topics in news and fiction directly, finding that LMs performed as well as humans with respect to identifying news topics, but outperformed humans when identifying topics in fiction. LM summaries help researchers identify topics, concepts, and patterns in their distributions.

Imitation of authorial style is another task related to but distinct from summarization. In literature classrooms, imitation has long been used as a pedagogical technique to help students better understand minutiae of authorial style. Gabi Kirilloff et al. use GPT-4 to generate 6,000 synthetic paragraphs in the styles of ten nineteenth-century authors ranging from Pauline E. Hopkins to Charles Dickens. They find that GPT-4's imitations are generally easy to detect, capturing authors' "themes without capturing literary style."[38] For example, the synthetic passages use nouns and determiners much more often than do the authors the LM was assigned to imitate, so much so that the researchers can predict whether a passage was written by the author or the LM more than 95% of the time. However, as in the classroom, degrees of failure in imitation also reveal aspects of authors' styles. They show that Mark Twain is an exception to their general conclusion: GPT-4 imitates Twain much better than the other authors in

[34] Haaris Mian, Melanie Subbiah, Sharon Marcus, Nora Shaalan, and Kathleen McKeown, *Computational Representations of Character Significance in Novels*, arXiv:2601.15508, arXiv, 2026, 3, https://doi.org/10.48550/arXiv.2601.15508.

[35] See Cody Kommers, Drew Hemment, Maria Antoniak, et al., *Meaning Is Not A Metric: Using LLMs to Make Cultural Context Legible at Scale*, arXiv:2505.23785, arXiv, 2025, https://doi.org/10.48550/arXiv.2505.23785.

[36] *Playing in the Dark: Whiteness and the Literary Imagination*, The William E. Massey, Sr. Lectures in the History of American Civilization 1990 (Harvard University Press, 1992), 6.

[37] David M. Blei, "Probabilistic Topic Models," *Communications of the ACM* 55, no. 4 (2012): 77, https://doi.org/10.1145/2133806.2133826.

[38] "'Written in the Style of ': ChatGPT and the Literary Canon," *Harvard Data Science Review* 7, no. 4 (2025): 22, https://doi.org/10.1162/99608f92.6d5fb5ef.

their sample.

Information retrieval is a fourth task for which LMs have been used with literary texts, and one that I will discuss further in the next section. Where the previous studies I have cited in this section apply LMs directly (or, in the case of imitation, indirectly) to literary texts, Katherine Thai and Mohit Iyyer apply LMs to both literary fiction and literary criticism simultaneously. Specifically, they provide an LM with the full text of a work of prose fiction as well as a work of literary criticism that features at least one direct quote from that work of fiction. For the experiment, one direct quote has been blanked out, though the surrounding critical context remains. The LM is then asked to determine which quote from the provided fiction has been blanked out in the criticism. The authors find that Google's LM Gemini outperforms experts at correctly identifying which quote has been removed. This is an exemplary information retrieval task because the goal is to find a specific passage in the fiction (the "needle") that best fits the context provided by both the criticism and the fiction itself (the "haystack").[39] The authors argue that this demonstrates LMs' capacity to assist with what they term literary evidence retrieval: Given a source text and a critical context, an LM can do as well as (or better than) experts at identifying omitted evidence.

Although there are other studies that I could discuss (as well as nits that could be picked with these studies), these four tasks give a good sense of how and why scholars have used LMs to read literary texts and literary scholarship directly. These descriptions also suggest some of the ways in which the assumptions underlying this work differ from those of most literary scholarship. For example, much of this work uses LMs to identify and evaluate many examples with shared properties, rather than focusing on how a few examples illuminate larger wholes. Though such aggregative approaches are most strongly associated with computational literary studies, they are also directly applicable to other fields such as literary history, genre studies, and stylistics, among others.[40]

All of these examples also directly or indirectly acknowledge two limitations of these approaches. The first is that all of these tasks artificially reduce the complexity of individual literary texts, though they also do so for limited purposes. This reduction is characteristic of efferent reading. The second is that LMs' stochastic natures mean that, even if their outputs are correct in the vast majority of cases, they *will* be wrong in some cases, and it is not possible to predict why or how they will go wrong when they go wrong. Such errors are intrinsic to the autoregressive structures of these models; they can be reduced, but it is not clear that they can be eliminated.[41] This is usually cited as the best argument against using LMs for research. However, critiques that stop there miss two points that the research reviewed above demonstrates. First, researchers can account for that degree of error using techniques like human-in-the-loop verification, ensemble methods, and statistical approaches to quantify their uncertainty. Second, the relevant comparator here is not perfect accuracy in evaluating a small number of cases, but expert accuracy in evaluating a large number of cases.

---

[39] On needles in haystacks, see also Sil Hamilton, Rebecca M. M. Hicke, Matthew Wilkens, and David Mimno, *Too Long, Didn't Model: Decomposing LLM Long-Context Understanding With Novels*, arXiv:2505.14925, arXiv, 2025, https://doi.org/10.48550/arXiv.2505.14925.

[40] See e.g., Oleg Sobchuk and Artjoms Šeļa, "Computational Thematics: Comparing Algorithms for Clustering the Genres of Literary Fiction," *Humanities and Social Sciences Communications* 11, no. 1 (2024): 1–12, https://doi.org/10.1057/s41599-024-02933-6.

[41] See Gary Marcus, *Taming Silicon Valley: How We Can Ensure That AI Works for Us* (The MIT Press, 2024).

Browsing, scanning, skimming with LMs

As both Pierre Bayard and Amy Hungerford have suggested, literary scholars must solve the problem of having too much to read by browsing, skimming, scanning, and sometimes skipping texts.[42] These judgments are made using efferent reading practices, not aesthetic ones. To be sure, much literary scholarship itself rewards Rosenblatt's aesthetic stance. But it would be impossible for a scholar to do all of their professional reading in that mode. Despite increases in publication and the expansion of admissible evidence, literary studies' aggregate research time has shrunk as the academic precariat has grown.[43] Below, I demonstrate how scholars have already begun to complement these techniques for reviewing secondary literature with others enabled by LMs, specifically vector search and retrieval augmented generation (RAG).

Unlike the research described in the previous section, vector search and RAG will seem similar to the ordinary research practices of most literary scholars today, especially for literature reviews. Specifically, they complement keyword search. Because of this connection, it is worth briefly considering how keyword search recently changed scholarly practices of browsing, skimming, and scanning, and how the discipline responded at that time. In 1995, Yahoo! Search came online; scholarly concern about keyword search followed soon thereafter.[44] Suddenly, students could find texts to plagiarize with ease. Over recent decades, this has gotten easier. Yet warnings about keyword search abated as researchers began to use this tool for their own work.[45] Few scholars today oppose keyword search with the vigor that some did in the 1990s. In a 2001 essay reflecting on the impact of keyword search on literary studies, David S. Miall lamented that full texts online "offer only a partial and inadequate solution to the needs of a literary scholar; even full-text searching provides access only to words, not to concepts."[46] Vector search and RAG can be combined to search texts in ways more closely aligned with Miall's conceptual search than keyword search. Though it may seem like a remote possibility now, what happened with keyword search after the 1990s may repeat with vector search and RAG, especially as they are currently being incorporated into electronic resources that scholars already depend on like EBSCO and Primo.

Vector search

To explain how vector search and RAG impact literary studies' reading, I must first briefly explain how vector search differs from keyword search.[47] LMs fundamentally emerge from observations of words and parts of words (collectively, "tokens") in many contexts across many documents. Relationships between one token and other tokens can be represented as vectors. This representation makes it possible

[42] The distinction: "Skimming is defined as getting the main idea or gist of a selection quickly and scanning as a high speed search for the answer to a specific question or the location of a specific fact" (Martha J. Maxwell, "Skimming and Scanning Improvement: The Needs, Assumptions and Knowledge Base," *Journal of Reading Behavior* 5, no. 1 (1972): 48, https://doi.org/10.1080/10862967209547021).

[43] For the latest figures, see Glenn Colby, "Data Snapshot: Tenure and Contingency in US Higher Education, Fall 2023," *Academe Magazine* 111, no. 2 (2025).

[44] See Scott Stebelman, "Cybercheating: Dishonesty Goes Digital." *American Libraries* 29, no. 8 (1998): 48–51 and Lisa Renard, "Cut and Paste 101: Plagiarism and the Net." *Educational Leadership* 57, no. 4 (1999): 38.

[45] See Christine L. Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (MIT Press, 2007) and Hannah Frydman, "In Defense of the Search Bar," *The American Historical Review* 130, no. 2 (2025): 714–35, https://doi.org/10.1093/ahr/rhaf010.

[46] "The Library versus the Internet: Literary Studies under Siege?" *PMLA* 116, no. 5 (2001): 1407, https://www.jstor.org/stable/463544.

[47] For a more technical overview, see Meredith Syed and Erika Russi, "What Is Vector Search?" in *IBM*, https://www.ibm.com/think/topics/vector-search, 2024.

to show, for example, that the vector for *claw* is more similar to the vector for *cat* than the vector for *hat*. Phrases, sentences, and paragraphs can be represented in much the same way, allowing for the similarity of entire passages to be compared using the same principle. Vector search uses this technique to retrieve the most similar passages in a document to a researcher's prompt.

The key difference between vector search and keyword search is that, in vector search, semantically similar passages can be identified even when none of the keywords in the prompt appear in the document. For example, suppose a researcher were scanning a monograph for discussions of serial publication, but none of its paratexts (table of contents, index, etc.) pointed to a clear answer. A keyword search for *serial* would miss passages about works "published in monthly numbers" or those that "appeared in installments," even though those phrases match the intent of the query precisely. Keyword search risks false negatives in cases like this by missing discussions of a subject that lack the keyword. Vector search, by contrast, risks false positives, retrieving relatively dissimilar passages this monograph did not, in fact, discuss serial publication.

While retrieving semantically similar passages is the exceptional feature of vector search, it is important to note why using vector search specifically (and LMs generally) on relatively recent documents in languages well represented online—such as recent scholarship in English—performs more reliably than does vector search of older documents, or documents in languages not well represented online.[48] Vector search is therefore more useful for searching secondary sources than primary sources, which is evident in the platforms that have begun to adopt it such as EBSCO, Primo, JSTOR, and Project MUSE. To understand why, we can think of perhaps the best-known example of semantic shift in English over recent centuries: *gay*.[49] The context of *gay* as used by a Twitter troll may be anachronistically imported into the semantic space of *gay* in *The Great Gatsby* because the same token appears in both places in the training data for the embedding model.[50] Ted Underwood, Laura K. Nelson, and Matthew Wilkens have studied anachronism of this sort as a particularly important problem for the use of LMs to analyze historical texts.

Although false positives and anachronism pose different threats to vector search than keyword search's false negatives, they can be addressed using the same strategy. No one who uses keyword search finds that its results invariably reflect their intentions. Keyword searchers try multiple queries when one does not return an expected result, and accept that inconclusive results occur. Experience with vector search likewise cultivates a sense of its capacities and limits. However, the line between keyword and vector search is becoming blurred by a paradigm called hybrid search that invisibly performs both keyword and vector search with the same query. Google Search has already worked this way for some time.[51] Vector and hybrid search paradigms are also being incorporated into library e-resources that scholars depend upon for searching catalogs and databases, such as EBSCO.[52]

---

[48] As of this writing, the top languages on the web by total number of webpages are English (49%), Spanish (6%), German (6%), Japanese (5%), and French (4%). See "Usage Statistics and Market Share of Content Languages for Websites, June 2025," in *W3Techs*, https://w3techs.com/technologies/overview/content_language, 2025.

[49] For a computational analysis of this shift, see William L. Hamilton, Jure Leskovec, and Dan Jurafsky, "Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change," *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* 2016 (2016): 2116.

[50] The former is Oxford English Dictionary, *Gay, Adj., Sense 10*, Oxford University Press, 2025, while the latter is Oxford English Dictionary, *Gay, Adj., Sense 3.a*, Oxford University Press, 2025.

[51] "About Hybrid Search," in *Google Cloud*, https://cloud.google.com/vertex-ai/docs/vector-search/about-hybrid-search, 2025.

[52] "EBSCO's AI Natural Language Search Mode," in *EBSCO Connect*, https://connect.ebsco.com/s/article/Coming-soon-Introducing-EBSCOs-AI-Natural-Language-Search-Mode?language=en_US, 2025.

If vector search seems only moderately useful for searching a single document, its effects become more profound when retrieving passages from multiple documents simultaneously, as in retrieval augmented generation.

Retrieval augmented generation

Vector search retrieves semantically similar passages to a query. Retrieval augmented generation (RAG) uses the passages surfaced by vector search as evidence as it generates textual responses to prompts in the style of a chatbot.[53] Crucially for scholars, RAG systems quote, cite, and link to the passages identified by vector search in their generated responses. Because their responses are "grounded" by specific documents, RAG has been found to reduce the likelihood of hallucination (or confabulation), which is when an LM generates plausible-sounding but false information, such as citations to documents that do not exist.[54] RAG is already being used for research in domains with certain structural similarities to literary studies like law.[55]

Like vector search, literary scholars have likely already experienced RAG whether they know it or not. Google's AI Overviews use RAG to respond to queries by gathering information from multiple webpages, generating responsive text, and citing the pages referenced in the result.[56] To get a sense of what this is like in a scholarly context, researchers can try JSTOR's RAG tool, which identifies responsive passages from across JSTOR's collection.[57] Primo also has a similar tool for identifying references in libraries' catalogs.[58]

The disadvantage of RAG systems like these is that the researcher does not choose the documents to be searched. However, a custom RAG system can complement the browsing, skimming, and scanning scholars do when preparing a literature review by allowing them to choose the documents themselves. For example, Javier Cha has done this by using Open WebUI's RAG implementation with LMs running on university-owned hardware. Because of the predominance of commercial chatbots like ChatGPT in discourse about LMs, few realize that it is possible to run open-weight LMs on their own computers using software like Ollama. Scholars who use reference management software like Zotero are especially well positioned to retrieve not only responsive passages across their documents, but also

[53] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA), NIPS '20, December 2020, 9459–74.

[54] For discussions of hallucination and confabulation, see Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo, *Hallucinations in Neural Machine Translation*, September 2018 and Peiqi Sui, Eamon Duede, Sophie Wu, and Richard Jean So, *Confabulation: The Surprising Value of Large Language Model Hallucinations*, arXiv:2406.04175, arXiv, 2024, https://doi.org/10.48550/arXiv.2406.04175. On RAG and hallucination, see Orlando Ayala and Patrice Bechard, "Reducing Hallucination in Structured Outputs via Retrieval-Augmented Generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, ed. Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar (Association for Computational Linguistics, 2024), https://doi.org/10.18653/v1/2024.naacl-industry.19.

[55] Ryan C. Barron, Maksim E. Eren, Olga M. Serafimova, Cynthia Matuszek, and Boian S. Alexandrov, *Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical Non-Negative Matrix Factorization*, arXiv:2502.20364, arXiv, 2025, https://doi.org/10.48550/arXiv.2502.20364.

[56] Few searchers check those citations, however. See Athena Chapekis and Anna Lieb, "Google Users Are Less Likely to Click on Links When an AI Summary Appears in the Results," in *Pew Research Center*, 2025.

[57] "JSTOR's AI Research Tool," in *About JSTOR*, n.d.

[58] "Getting Started with Primo Research Assistant," in *Ex Libris Knowledge Center*, https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/020Primo_VE/Primo_VE_(English)/015_Getting_Started_with_Primo_Research_Assistant, 2024.

metadata about those passages. Whatever a specific RAG implementation may be, the key point is that a RAG system can run on a laptop to search documents of a scholar's choosing without necessarily providing copies of any of those sources to model-makers like OpenAI, and can even be used without an internet connection.

Vector search and RAG will be more similar in their effects on literary scholarship to keyword search than other uses of LMs that attract more attention and alarm, such as generating essays. Just as most scholars today routinely search library catalogs, databases, and PDFs for keywords, RAG makes it possible to identify passages within and across documents that are semantically similar to researchers' prompts. And just as scholars had to become experienced with keyword search to understand its strengths and weaknesses, so too with vector search and RAG, though hybrid search paradigms may make this process more difficult by blurring the distinction. As Joshua Rothman recently suggested in *The New Yorker*, "In our current reading regime, summarized or altered texts are the exception, not the rule. But over the next decade or so, that polarity may well reverse: we may routinely start with alternative texts and only later decide to seek out originals." "Alternative texts" need not be summaries generated by LMs. In the case of vector search and RAG, the constituent passages of a text are *reorganized* based on their semantic similarity to a query, such that the passage of an article most similar to the query becomes the first one presented for reading, irrespective of its true position in the text. In certain respects, this is similar to the excerption of longer prose works in an anthology, where a reader's encounter with a part of a text makes a case for its whole.[59] The important difference here is that, rather than selections made by an editor for a general audience, the reorganization of passages in a RAG system is individualized, transient, and always subject to the revision of the prompt or a change of LM. Grounded in a bibliography determined by the researcher, vector search and RAG complement existing browsing, skimming, and scanning techniques that scholars already use to determine what to read and how to read it.

Reading paraliterary texts with LMs

The preceding two sections discussed using LMs to read literary texts and literary scholarship. Both involved the evaluation of LM outputs by experts, whether through comparison to validated data, or through iterative refinement of vector search and RAG results. However, irrespective of the validation processes put in place, some literary scholars likely remain skeptical about using these models to gather information directly from literary texts themselves because of the literary qualities like irony and ambiguity that make those texts interesting to study in the first place. In this section, I will show how LMs can be used to study literature without applying them to literary texts. I will do so by briefly discussing some of my ongoing research on the literary canon, which uses LMs in different ways and on different kinds of texts than those discussed thus far.

Although the literary canon "never appears as as a complete and uncontested list in any particular time and place," there exist some good proxies for that list, including comprehensive literature anthologies like those published by Norton as well as reference works like *The MLA International Bibliography*.[60][61] In pursuit of other such lists, I have extended this line of research to the quiz show

[59] See Leah Price, *The Anthology and the Rise of the Novel: From Richardson to George Eliot* (Cambridge University Press, 2000).

[60] John Guillory, *Cultural Capital: The Problem of Literary Canon Formation*, First edition, enlarged (The University of Chicago Press, 2023), 30.

[61] See Erik Fredner and J. D. Porter, "Counting on *The Norton Anthology of American Literature*," *PMLA* 139, no. 1 (2024):

Jeopardy.

Jeopardy should be of interest to literary scholars because it is one of the few places in American popular culture that routinely engages with literary history. About one in five questions references literature. Unlike anthologies or the *Bibliography*—both of which approach canonicity from a scholarly perspective—*Jeopardy* provides evidence about both popular and scholarly understandings of the canon.[62]

Although about one in five *Jeopardy* questions references literature, they are not all equally difficult. Easy questions will likely be known by many viewers, whereas difficult questions may only be known by experts. *Jeopardy* quantifies difficulty through gameplay, such as the different dollar values writers assign to each question. As a result, over the past forty years, *Jeopardy* clue writers have made thousands of specific wagers about which literary authors and works Americans would be more or less likely to know. Aggregating these wagers provides new insights into the structure of the literary canon. But it remained inaccessible to scholars because *Jeopardy* questions are not structured for this analysis as written. However, I have used LMs to extract and restructure that latent information for analysis.

The results of this research have already shown difficult questions are more likely to be associated with authors and works better known to literary scholars than to the general public, whereas easy questions are more likely to be associated with literature for young people and popular literature, especially authors and texts with film and television adaptations.[63] Some of the authors most frequently referenced in the most difficult clues include Gogol, Pynchon, Petrarch, and Strindberg whereas some of the easiest are Aesop, Schulz, Collins, Rowling, and Tolkien. However, *Jeopardy* does not reproduce the popularity-prestige continuum solely along the axis of difficulty. A small number of the most canonical authors like Shakespeare and Cervantes appear at every level of difficulty. However, the relative difficulty of authors' or texts' questions tends to reproduce the relative valuation that characterizes canonicity itself.

While there is much more to be said about these results elsewhere, my present purpose is to explain how LMs enabled this research by classifying and extruding structured data from *Jeopardy* clues. This work differs from other research discussed thus far by using LMs to read paraliterary texts in order to study a literary topic. This distinction matters because it avoids the problem of attempting to read literary texts themselves, which are often studied precisely because of their ability to resist encapsulation.

Classification

In support of this research, the editor of the *Jeopardy* fan site J-Archive! shared a copy of their database of about 550,000 questions that have been asked on air since 1984 with me.[64] I then had to determine which of those hundreds of thousands of questions referenced literary authors or texts. Many computational approaches to text classification predate LMs, and LMs are not always the best choice

50–65, https://doi.org/10.1632/S0030812923001189 as well as Erik Fredner and Mark Algee-Hewitt, "*The MLA International Bibliography*'s History of English-Language Literary Studies, 1982-2023," *DH2024* (Arlington, VA), August 2024.

[62] For other efforts to quantify the popularity-prestige distinction with respect to canonicity, see J. D. Porter, "Popularity/Prestige," *Pamphlets of the Stanford Literary Lab*, no. 17 (September 2018) as well as Jean Barré, Jean-Baptiste Camps, and Thierry Poibeau, "Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature," *Journal of Cultural Analytics* 8, no. 3 (2023), https://doi.org/10.22148/001c.88113.

[63] Erik Fredner, "The Literary Canon on *Jeopardy!*, 1984-2024," *DH2025* (Universidade NOVA de Lisboa), July 2025, https://doi.org/10.5281/zenodo.19494801.

[64] See "The Fan-Created Archive of *Jeopardy!* Games and Players," in *J! Archive*, https://j-archive.com/, n.d.

for this kind of task.[65] However, other approaches often require a large number of words in each text to be reliably classified. *Jeopardy* questions, by contrast, are cryptic and concise. For example:

Category	BEFORE & AFTER
Clue	San Antonio Spurs "Admiral" marooned by Daniel Defoe
Answer	David Robinson Crusoe

This clue requires contestants to combine knowledge of basketball (David Robinson) and the eighteenth-century novel (Robinson Crusoe) to produce the nonce answer "David Robinson Crusoe." LMs outperform other classification methods in gnomic cases like this because, in an LM's representation, words are embedded with other words that do not appear in the text. James Dobson and Scott Sanders have argued that counting words radically decontextualizes "the enclosed words in ways that foreclose many modes of critical analysis."[66] For contemporary texts in languages well represented online, LMs do the opposite.

As in much of the work cited above, I evaluate how well LMs classify *Jeopardy* questions by comparing the outputs of many different LMs to a benchmark dataset with a representative sample of validated classifications. After identifying the best-performing LM, I run that model against the same dataset multiple times, allowing it to vote on each classification several times. Because any given model response has a chance of containing a hallucination, self-consistency is a common strategy to mitigate hallucinations. When classifying *Jeopardy* questions, the model disagreed with itself about 2.5% of the time. Those points of disagreement often corresponded with ambiguous cases. For example, co-authoring the political thriller *The President is Missing* with James Patterson is not the most salient fact about Bill Clinton. Does his stint as a novelist mean that every clue about Clinton after 2018 should also count as a reference to an author of fiction, even when a given clue is about his presidency, his ability as a saxophonist, or his veganism? Reasonable people could disagree.

Despite such difficulties, GPT-5 matched the benchmark classification data about 94% of the time.[67] Although perfect accuracy might seem necessary, some degree of error is inevitable. If these classifications had instead been done manually, it would have taken months of meticulous work that might not have produced better results. People cannot do this kind of classification work effectively for even one hour at a time. Some studies suggest that humans' ability to do this kind of work begins to deteriorate in as little as two minutes.[68] For tasks like this, the relevant comparator is not perfect accuracy, but human accuracy.

Extrusion

Jeopardy fans will already know that it is acceptable to answer a question about a person with their last name using the form, "Who is Eliot?" However, in a question about literature, this answer would not include all of the relevant information: Is this clue about George, T.S., or some other Eliot? Similarly,

[65] See David Bamman, Kent K. Chang, Li Lucy, and Naitian Zhou, *On Classification with Large Language Models in Cultural Analytics*, arXiv:2410.12029, arXiv, 2024, https://doi.org/10.48550/arXiv.2410.12029.

[66] "Distant Approaches to the Printed Page," *Digital Studies / Le Champ Numérique* 12, no. 1 (2022): 4, https://doi.org/10.16995/dscn.8107.

[67] Precision of 0.94, recall of 0.93, and F1 of 0.94.

[68] Curtis M. Craig and Martina I. Klein, "The Abbreviated Vigilance Task and Its Attentional Contributors," *Human Factors* 61, no. 3 (2019): 426–39, https://doi.org/10.1177/0018720818822350.

literature questions that contain quotes or other references do not necessarily name the author or work referenced. For example, a question about the famous movie line "Stella!" may be answered "Brando," leaving Tennessee Williams and *A Streetcar Named Desire* unmentioned. For these reasons, I use an LM to extrude structured information about authors and texts referenced in clues about literature even when they are not explicitly named. This is especially important when a clue contains multiple literary references:

Category	FOR WHOM THE BELL TOLLS
Clue	Quasimodo would be relieved that Emmanuel, this structure's 13-ton bell, is now rung electronically
Answer	Notre Dame

This clue minimally references Donne's *Devotions upon Emergent Occasions*, Hugo's *Notre-Dame de Paris*, and Hemingway's *For Whom the Bell Tolls*.[69] The LM extrudes that information in a data structure that can be used to link recurring references to the same authors and works, as well as disambiguate works with the same title or authors with the same name.

Evaluating the quality of such output is more complex than classification because it must be more flexible. It is not sufficient to write down correct answers and check if the model's output matches them exactly. LMs often produce output that is accurate without precisely matching reference data. For example, some clues with references to Sherlock Holmes are associated with the author "Arthur Conan Doyle" whereas others are associated with "Sir Arthur Conan Doyle." Similarly, LMs occasionally substitute alternate titles, such as connecting one reference to Scheherazade with *One Thousand and One Nights* and another with *The Arabian Nights*. These answers both correctly refer to the same entities without using identical representations. There are many strategies to evaluate and improve the accuracy of this kind of output, some of which involve using another LM to judge the semantic equivalence between an output and a validated answer.[70] However, this kind of output can also be evaluated and corrected by hand.

Grading a representative sample of such extrusions, I found the ouputs to be about 98% accurate.[71] In addition to exceptionally accurate performance on straightforward clues (e.g., "This author of *The Wizard of Oz*…"), LMs also reliably catch subtler literary references. For example:

Category	THE OLD TESTAMENT
Clue	In Genesis 21, Abraham banishes Hagar & this son of theirs to the desert; call him…
Answer	Ishmael

LMs consistently identify the allusion to *Moby-Dick*'s first sentence at the end of this clue despite its brevity, modified syntax ("Call him" not "Call me Ishmael"), and the overriding biblical context.

I dwell on these details to emphasize that the tasks for which I used LMs in this research—classification and extrusion of literary references—bear considerable resemblance to work

[69] Though Emmanuel might reasonably be construed as a reference to the Book of Isaiah (and Notre Dame to Mary), I exclude canonical sacred texts such as the Bible, the Quran, and the Vedas from this analysis.

[70] Haitao Li, Qian Dong, Junjie Chen, et al., *LLMs-as-Judges: A Comprehensive Survey on LLM-Based Evaluation Methods*, arXiv:2412.05579, arXiv, 2024, https://doi.org/10.48550/arXiv.2412.05579.

[71] `gpt-5`. Authors: precision 0.99, recall 0.98, F1 0.98. Works: precision 0.98, recall 1.0, F1 0.99.

that many literary scholars do (or, if they have the means, hire research assistants to do) such as Tessa Roynon's study of classical allusions in Toni Morrison's novels. Validating and correcting outputs manually exemplifies the ways in which LMs can be used to complement literary expertise. Evaluating model outputs one by one emphasizes both how impressive and how inconsistent LMs can be. Researchers cannot afford to lose sight of either when reading with them.

Unlike earlier examples that used LMs with literary texts and criticism, this research on *Jeopardy* reads paraliterary texts with LMs to study the literary canon. Even if *Jeopardy* clues themselves do not reward close reading, their metadata—authors and works referenced, difficulty, dating, etc.—characterizes clue writers' perceptions of the commonness or rarity of the knowledge they test, providing valuable evidence of the changing contours of the literary canon.

Conclusion

Researchers have already been reading literature, literary criticism, and paraliterary texts with LMs in ways that should be of interest to literary studies writ large for years. I have argued that an emergent norm across the many uses of LMs discussed here is a greater license to read with LMs for what Rosenblatt calls efferent reading ("what is to be extracted and retained after the reading event"), but lesser license for aesthetic reading ("what is being lived through during the reading event"). LMs cannot independently generate close readings that fulfill literary studies' expectations without embodied experiences of reading and writing. However, reading with LMs can complement and accelerate the discipline's equally necessary but less often theorized efferent reading.

I wish to conclude by suggesting some of the ways that the uses of LMs described here seem likely to impact literary studies in the short term. Whether and how literary studies will continue to read with LMs depends on material factors, the most important of which may be cost. In 2026, capital expenditure on LM infrastructure by US big tech companies is expected to exceed two percent of US gross domestic product. If it does, this will make it the second most expensive capital project in US history: far greater than the space race, slightly greater than the railroad build-out of the 1850s, and exceeded only by the Louisiana Purchase.[72] Investors will demand that these extraordinary costs be recouped with interest. However, China's "six tigers"—LM companies that make models competitive with the state of the art—have prioritized releasing open-weight models that can be freely downloaded and run on hardware of one's choosing, thereby driving down costs. To compete, both Google and OpenAI have released open-weight models. This competition matters because open-weight models are adequate for many of the research tasks described here, and have other advantages for researchers—chief among them, reproducibility. Proprietary LMs like ChatGPT and Claude may never be cheaper than they now are, which could preclude their use for some kinds of research in the future. At the same time, open-weight models are sufficient for much of the work described here.

Of the applications discussed in this essay, those likely to have the most widespread impact on literary studies are vector search and RAG. Both of these are being incorporated into tools like EBSCO and Primo, which provide the search interfaces for many university library catalogs, as well as search engines like Google. Vector search and RAG will make missing relevant research less likely, such as a monograph that discusses a topic that is related to but would not be identified by a particular keyword search. Tools like Primo Research Assistant also make it possible to generate annotated bibliographies using a library's holdings. While the latter represents a paradigm shift for library patrons from reviewing

[72] Meghan Bobrowsky, Drew An-Pham, and Alana Pipe, "Big Tech's AI Push Is Costing a Lot More Than the Moon Landing," *Wall Street Journal*, February 2026.

a list of search results to reading a chatbot-like interface, I suggested earlier that these kinds of changes may eventually seem as unremarkable as keyword search now seems.

Many of the data-driven research questions discussed above differ from those studied by most literary scholars. However, not every question in literary studies is best approached through close reading. Questions of the sort asked by fields like literary history, genre theory, and stylistics require gathering information from across many different texts, tasks at which these models excel. Scholars studying individual authors with large oeuvres face similar problems. Computational approaches to these problems of scale have historically been limited to the small group of literary scholars who are either computer programmers themselves or who collaborate with programmers. For scholars whose barriers to conducting computational research were technical rather than theoretical, the recent emergence of LM coding agents, the most famous of which is currently Claude Code, suggests that a lack of coding knowledge may no longer be a significant barrier to generating code that meets their needs. That said, doing so responsibly is an entirely different question, and one that falls outside of the bounds of this essay.

Finally, the *Jeopardy* example demonstrates that even *a priori* objections to the use of LMs to read literary texts directly does not exhaust their usefulness to literary studies. LMs can read paraliterary texts and extrude latent information about literary history. Obvious candidates for similar work include identifying literary allusions in periodicals, on social media, and in film, television, or podcast transcripts.

Through all of these examples, I have tried to show that the argument against LMs on grounds of uselessness for literary studies is false with respect to scholarship. The better arguments are ethical. Despite intrinsic problems like hallucination, emerging practices make the use of LMs more reliable. Furthermore, the relevant criterion for assessing their reliability is not perfect accuracy, but human accuracy. Under such conditions, LMs can complement literary studies' reading, though that does not mean that they must.