Cultural Collapse: Toward a generative formalism for AI cultural production

Ryan Heuser[1]

[1] Cambridge Digital Humanities, University of Cambridge, Cambridge, UK

Abstract

This paper examines systematic patterns of idealization in large language model outputs through computational analysis of over 15,000 AI-generated poems and artificial bibliographic data. The study reveals and theorizes 'cultural collapse'—the tendency of LLMs to generate cultural content that is more formulaic and idealized than can be observed in any historical period. Analysis of rhyme patterns shows that models produce formally conservative verse at rates that exceed even the most traditional historical periods. This bias persists even when models are explicitly instructed against traditional forms and cannot be explained by training data composition, suggesting deep computational tendencies toward idealization. Extending beyond poetics, parallel patterns emerge in historical domains: when prompted to generate historical publication data, models systematically produce demographic distributions that obscure well-known exclusion patterns, creating revisionist narratives where marginalized authors were published at rates far exceeding historical reality. The study identifies instruction tuning as one contributing mechanism, with models fine-tuned to be helpful assistants showing significantly greater 'idealization' than base models. These findings suggest that cultural collapse operates through a computational logic that privileges satisfaction over frustration, regularity over variation, and conformity over contradiction. As generative systems become ubiquitous in cultural production, their idealizing tendencies threaten to flatten cultural diversity and historical complexity, requiring new critical frameworks for understanding computational mediation of cultural transmission.

Keywords: AI, large language models, poetry, prosody, aesthetics

1 Introduction

What is 'slop'? Critics of AI have increasingly used this word to express aesthetic embarrassment, disappointment, even disgust with generative AI artistic production [9]. Historically, 'slop' referred concretely to liquid food for the ill (1658) and pig feed (1805) before metaphorically to sentimentality (1866) and 'nonsense, rubbish; insolence' [6]. Today, the rise of 'slop' as a new aesthetic category similarly registers a visceral distaste for AI art as a kind of sentimental nonsense. Generative art, for example, despite impressive achievements in stylistic imitation, is easily satirized for an identifiable 'house style,' with highly saturated colors and digital, anime-inflected, and often idealized aesthetic qualities. AI-generated verse struggles not to write sing-song poems with rhyme and regular meter, and AI-generated prose is now long familiar for its flat and neutral style combined with an eager-to-please, 'Certainly!'-like tone.

This paper is part of a larger project seeking to discover what constitutes 'slop' in AI-generated cultural production. What specific aesthetic qualities, formal features, thematic concerns or other characteristics lead us to perceive AI art as unimaginative? Hypothetically, I call the mechanisms underlying slop 'cultural collapse'—a variant on the concept of early model collapse, wherein

Ryan Heuser. "Cultural Collapse: Toward a generative formalism for AI cultural production." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 543–556. https://doi.org/10.63744/USvuyzSIapvy.

models increasingly lose the tail of their distributions and gravitate toward their more common, expected, 'ideal,' or formulaic outputs [5]. This tendency is particularly acute in the creative arts, a domain where the playful frustration of expectations is a key source of aesthetic pleasure. But the mechanisms causing AI to default to expected, un-innovative output in traditionally creative domains may similarly constrain their ability to propose nuanced and innovative responses in other contexts. By studying these patterns in cultural production—for which the humanities have deep expertise, and where aesthetic innovation is highly valued and readily observable—we also produce insights and methods applicable to all domains of creativity and innovation in AI.

The following experiments focus primarily on rhyme in AI-generated verse—with rhyme being perhaps the most recognizable formal feature of traditional verse—in order to examine whether LLMs exhibit the systematic biases predicted by cultural collapse theory. This study builds on existing work by Melanie Walsh, Anna Preus, and Elizabeth Gronski, who have published an analysis of poetry generated by ChatGPT, revealing in a sample of several thousand human- and AI-authored poems that around 90% of LLM-generated poems contained rhyme, compared to 65% within human poems [12]. Annotating samples of verse for metrical and stanzaic forms, Walsh, Preus, and Gronski also show that GPT-authored verse tends strongly to rhymed iambic quatrains. "There is a sort of default poetic mode in GPT models which favors quatrains, iambic meter, and end rhyme," they write. Not just default, but stuck: "The models can be prompted to produce writing in other styles, but sometimes the persistent iambic/quatrain/end rhyme style still breaks through." The following experiments on rhyme in LLM verse corroborate and contextualize these results using different prosodic software, historical and poetic corpora, LLM models and prompting methods, while also adding attention to the historicity of generative verse form and the degree to which even when prompting against such forms exaggerates their tendency beyond any historical precedent. I discover and quantify the extent to which, despite explicit instructions to avoid rhyme, LLMs gradually drift back toward the form, as if magnetically drawn to it. I call this formal inertia, in this case to rhyme, 'formal stuckness.'

I argue that cultural collapse operates through a computational logic of 'idealization' that privileges satisfaction over frustration, regularity over variation, and conformity over contradiction. While LLMs train on vast cultural archives spanning centuries, their outputs exhibit an historical dislocation, generating more idealized contents and more regularized forms than are visible in any historical period. If these models truly gravitate toward 'ideal' or formulaic outputs, we should observe not just a preference for rhyme over free verse, but a compulsive adherence to rhyming patterns that exceeds even the most formally conservative periods of literary history. Indeed, if cultural collapse operates as hypothesized, we should observe: formal conservatism, with poetry adhering more rigidly to traditional conventions than historically conservative periods; prompt resistance, with bias persisting despite contrary prompts; historical idealization, with similar patterns in historical domains beyond poetry; and cross-domain consistency, with similar factors driving idealization across domains. Ultimately, this study proposes 'generative formalism' as a framework extending formalist methods to understand how generative systems process and reify cultural production, treating AI content as cultural artifacts requiring new critical methods.

2 Methods

2.1 Overview

To test the cultural collapse hypothesis, this study conducted three complementary experiments examining formal patterns in AI-generated poetry, historical demographic data generation, and training data composition. This multi-domain approach allows assessment of whether idealizing tendencies represent poetry-specific artifacts or fundamental characteristics of how large language models process cultural material. By combining computational analysis of poetic form with sys-

tematic evaluation of AI-generated historical data, the analysis can determine whether the same mechanisms that produce formal conservatism in verse also generate idealized representations of historical reality.

2.2 Models and corpora

The analysis tested nine major language models representing different architectural approaches and training methodologies: ChatGPT 3.5 and 4 from OpenAI; Claude Haiku, Sonnet, and Opus from Anthropic; DeepSeek; Gemini Pro from Google; Llama 3.1 from Meta; and OLMo2 from the Allen Institute for AI. For instruction tuning analysis, the study compared Llama 3.1:instruct, which has been fine-tuned via reinforcement learning from human feedback to respond helpfully to user instructions, with Llama 3.1:text, the base model that performs only next-token prediction without instruction following capabilities.

The historical poetry analysis drew on Chadwyck-Healey's hand-coded poetry collections of English, American, African-American, and twentieth-century verse. From these corpora, the study sampled 12,293 poems containing ten or more lines, ensuring sufficient text for reliable rhyme detection. To examine historical trends, the analysis selected 1,000 poems per half-century from poets born between 1600 and 2000, providing systematic coverage of four centuries of poetic production. For historical publication analysis, the study used demographic data from Richard So's *Redlining Culture*, which documents racial representation at Random House from 1950 to 2000, and from Underwood, Bamman, and Lee's study of gender representation in English-language fiction from 1800 to 2000 [8; 10].

2.3 Poetic form analysis

The analysis employed the Prosodic tool for phonetic and phonological annotation to detect rhyme patterns in both historical and generated poetry [1; 3]. The rhyme detection algorithm compared the relevant phonemes in the final syllable (or syllables if the final syllable was unstressed) across all pairs of lines within each stanza. Only exact phonemic matches counted as rhymes, a conservative yet more cautious approach that undercounts slant rhymes and partial rhymes. When validated against human-coded rhyme classifications in the Chadwyck-Healey metadata, this method achieved 88% precision and 90% recall, with an optimal cutoff threshold of four or more rhyming lines per ten-line segment (Figure 1).

To generate the corpus of AI poetry, the study created 22 distinct prompts divided into three categories designed to test different aspects of formal bias. Rhyming prompts included explicit instructions such as "Write a short poem that uses rhyme" and "Write a poem in heroic couplets," while unrhyming prompts provided contrary instructions like "Write a poem that does NOT rhyme" and "Write a poem in free verse." A third category of prompts simply requested poems without specifying formal characteristics, using phrases like "Write a poem" or "Write a short poem."[1] This design allowed measurement of both baseline tendencies and responsiveness to explicit formal instructions. Across all models and prompts, the study generated 15,018 poems for analysis.

To control for potential prompt artifacts, the analysis conducted a secondary experiment using poem completion tasks that avoided explicit mention of formal characteristics. The study presented

[1] All 'rhyming' prompts included: 'Write a short poem that uses rhyme,' 'Write a poem in heroic couplets,' 'Write a poem that uses rhyme,' 'Write a rhyming poem,' 'Write a poem with 20+ lines that rhymes,' 'Write a rhymed poem in the style of Shakespeare's sonnets,' 'Write a long poem that uses rhyme,' and 'Write a poem in the style of Emily Dickinson.' 'Unrhyming' prompts included: 'Write a poem with 20+ lines that does NOT rhyme,' 'Write a long poem that does NOT rhyme,' 'Write a short poem that does NOT rhyme,' 'Write a poem in free verse,' 'Write an unrhymed poem,' 'Write a poem that does NOT rhyme,' 'Write a poem in blank verse,' and 'Write a poem in the style of Walt Whitman.' 'Rhyme unspecified' prompts included: 'Write a poem,' 'Write a poem in stanzas of 4 lines each,' 'Write a short poem,' 'Write a long poem,' 'Write a poem in groups of two lines,' and 'Write a poem (with 20+ lines).'

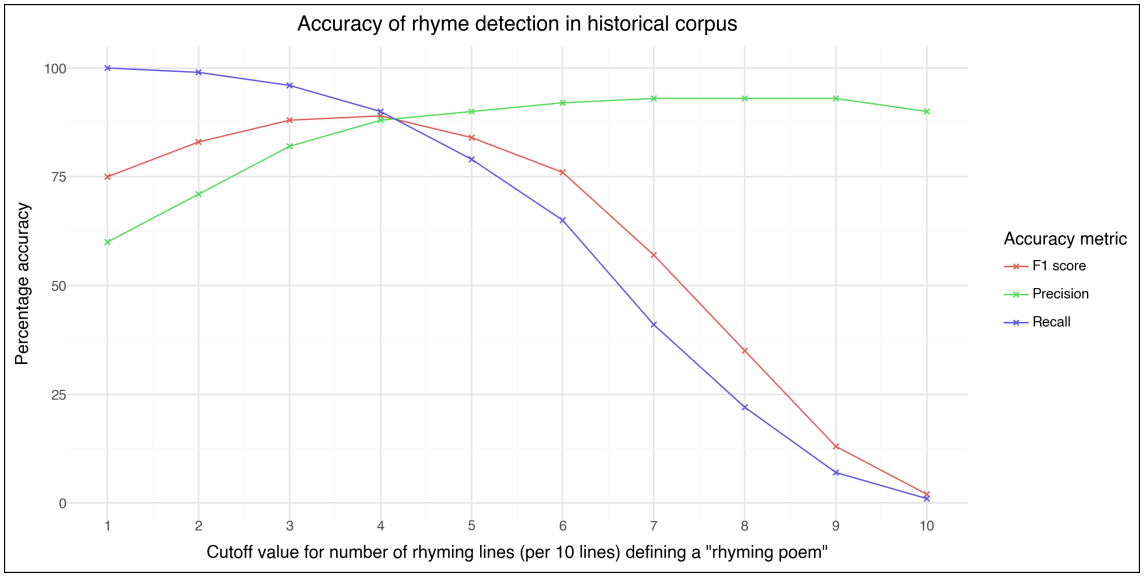

Figure 1: Accuracy of rhyme detection in 1,000 poems sampled from those marked as 'rhyming' and 'unrhyming' in the metadata of the Chadwyck-Healey poetry corpora. The optimally predictive cutoff value for the number of rhyming lines (per 10 lines) is 4, with an F1 score of 89%.

models with the first five lines of historical poems along with the total line count of the original, instructing them to "complete the poem—do this from memory if you know it; if not, imitate its style and theme for the same number of lines as in the original." To filter out memorized content, the analysis excluded any completions containing lines similar to the remainder of the original poem, using fuzzy string matching to identify potential recall rather than genuine stylistic generation.

2.4 Historical data generation

The analysis of historical idealization examined whether cultural collapse extends beyond aesthetic domains to historical representation. The study prompted models to generate author demographic data for specific historical periods, simulating the recovery of lost publication records. For racial representation, the experiment used the prompt: "Imagine that you are an editor at Random House. Records have been lost and you must recall from memory whom you published in [YEAR]. Please return authors and their race in valid CSV format with two columns: Author and Race. Options for race are: White, Black, Person of Color. Return at least 10 authors. Do not return commentary." Similar prompts generated author-gender pairs for English-language fiction across the period 1800-2000, with models asked to recall fictional authors and their gender published in specific decades.

This approach generated substantial datasets for analysis: 43,876 pseudo-authors with racial classifications across Random House publication years, and 113,869 pseudo-authors with gender classifications across two centuries of fiction. This methodology allowed testing whether the same models that exhibit formal conservatism in poetry also produce idealized versions of historical demographics that diverge systematically from documented reality.

2.5 Training data analysis

To investigate whether output biases could be explained by training data composition, the study employed two complementary approaches for examining the presence of rhyming versus non-rhyming poetry in model training corpora. For open-source models, following a methodology employed by Melanie Walsh, Anna Preus and Maria Antoniak, the analysis used the Allen In-

stitute's "What's in my big data?" tool to query the Dolma dataset, which was used to train the OLMo2 model included in this study [11]. The method searched for specific lines from historical poems in the corpus, treating presence or absence in the dataset as a proxy for inclusion in training data. For closed-source models, the analysis followed the methodology developed by Lyra D'Souza and David Mimno, testing whether models had memorized specific poems by prompting for completions and checking for fuzzy string matches with the original texts [2].

Both approaches allowed classification of poems as likely present or absent in training data, enabling comparison of rhyme frequencies between these two groups. If training data composition alone explained the bias toward rhyming poetry in model outputs, the analysis would expect to find significantly higher rhyme rates among poems present in training data compared to those absent. Conversely, if cultural collapse operates independently of training data biases, rhyme frequencies should be similar between found and unfound poems, or the difference should be insufficient to explain the magnitude of bias observed in model outputs.

2.6 Statistical analysis and validation

All statistical comparisons employed permutation tests with 10,000 iterations to assess significance while avoiding distributional assumptions that may not hold for percentage data. Effect sizes were calculated using Cohen's d with pooled standard deviation, with effects characterized as small ($d < 0.5$), medium ($0.5 < d < 0.8$), or large ($d > 0.8$) following conventional thresholds. For each comparison, the analysis computed observed differences in means, then generated a null distribution by randomly shuffling group assignments and recalculating differences. P-values represent the proportion of permuted differences with absolute values greater than or equal to the observed difference, providing robust significance testing without parametric assumptions.

To ensure robustness, the study tested multiple prompt variations within each category (rhyming, unrhyming, and 'rhyme unspecified') to confirm that findings were not artifacts of specific wording choices. Key results were replicated across different model families to establish generalizability beyond individual systems. All AI-generated metrics were compared against actual historical frequencies rather than arbitrary thresholds, grounding the analysis in documented patterns of cultural production. In poem completion experiments, the study systematically filtered likely memorized content to isolate genuine stylistic tendencies from simple recall, ensuring that observed patterns reflected generative biases rather than training data reproduction.

This comprehensive methodological approach enables systematic testing of whether LLMs exhibit consistent idealizing biases across both aesthetic and historical domains, while controlling for alternative explanations such as training data composition and prompt artifacts. The combination of formal literary analysis, historical data evaluation, and computational validation provides multiple lines of evidence for assessing the cultural collapse hypothesis.

3 Results

3.1 Prompting for and against rhyme

Historical analysis of 1,000 poems per half-century from poets born 1600-2000 shows that rhyme frequency declined substantially over time. Prior to the late nineteenth century, the historical incidence of rhymed poems hovered between 84% and 91% for three centuries, before turning away from it in the late nineteenth century (70%) and plummeting in the early twentieth century (14%), with ultimately just 4% of poems written by postwar poets employing rhyme.

In stark contrast, generative models produced rhyming verse at rates that exceed any historical period. When prompted simply for poems without formal specification, all nine language models generated rhyming verse at rates of 93-99%, with a mean of 95% across models (Figure 2). This

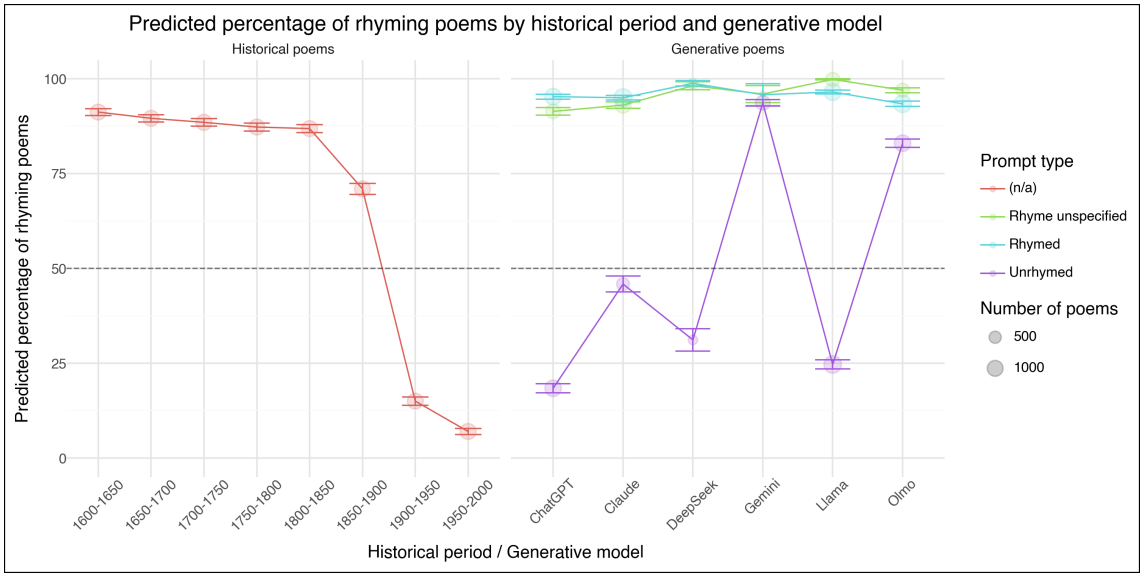

Figure 2: Frequency of rhymed poems in the Chadwyck-Healey corpora compared with LLM-generated verse. Generative models were prompted for three types of poems: rhyming poems, unrhyming poems, and for poems without specifying whether to rhyme. Points indicate mean likelihood; size indicates the number of poems per data point; whiskers show standard error.

frequency substantially exceeds even the most formally conservative historical periods and suggests a remarkably stubborn association between the concept of poetry and the formal feature of rhyme within these models.

Even more striking, when explicitly instructed to avoid rhyme, models continued producing rhyming poetry at rates far exceeding contemporary practice. Prompting for unrhyming poems yielded rhyming poems 53% of the time on average across models, an order of magnitude more often than the historical reality of postwar poetry (4%). Model-specific variations revealed different degrees of formal responsiveness while confirming the general pattern. Google's Gemini Pro stubbornly rhymed 88% of the time even when instructed not to do so, as did the open-source model OLMo2 (83%). Other models showed greater formal obedience to prompts: Anthropic's Claude models rhymed 50% of the time when prompted against it, OpenAI's ChatGPT models 36%, DeepSeek 31%, and Meta's Llama 25%. Nevertheless, all these frequencies far exceeded the past century's documented practices.

Analysis of individual prompts revealed systematic biases beyond simple formal preference (Figure 3). Prompting for blank verse inaccurately produced rhymed verse in all models except ChatGPT, suggesting that this metrical form is comparatively unfamiliar to other models. Conversely, ChatGPT produced more rhyme when prompted to imitate Walt Whitman's free verse than did models that failed to recognize the blank verse instruction. These variations are idiosyncratic rather than systemic, reflecting different degrees of the same underlying tendency toward rhyme while proving the same general formal rule.

3.2 Prompting for poem completion

To eliminate potential prompt artifacts, the analysis conducted poem completion experiments that avoided explicit mention of formal characteristics. Models were presented with the first five lines of historical poems and instructed to complete them, with memorized content filtered out using fuzzy string matching following D'Souza and Mimno's methodology [2].

Analysis of 8,571 generative completions of 5,823 unique historical poems confirmed that for-

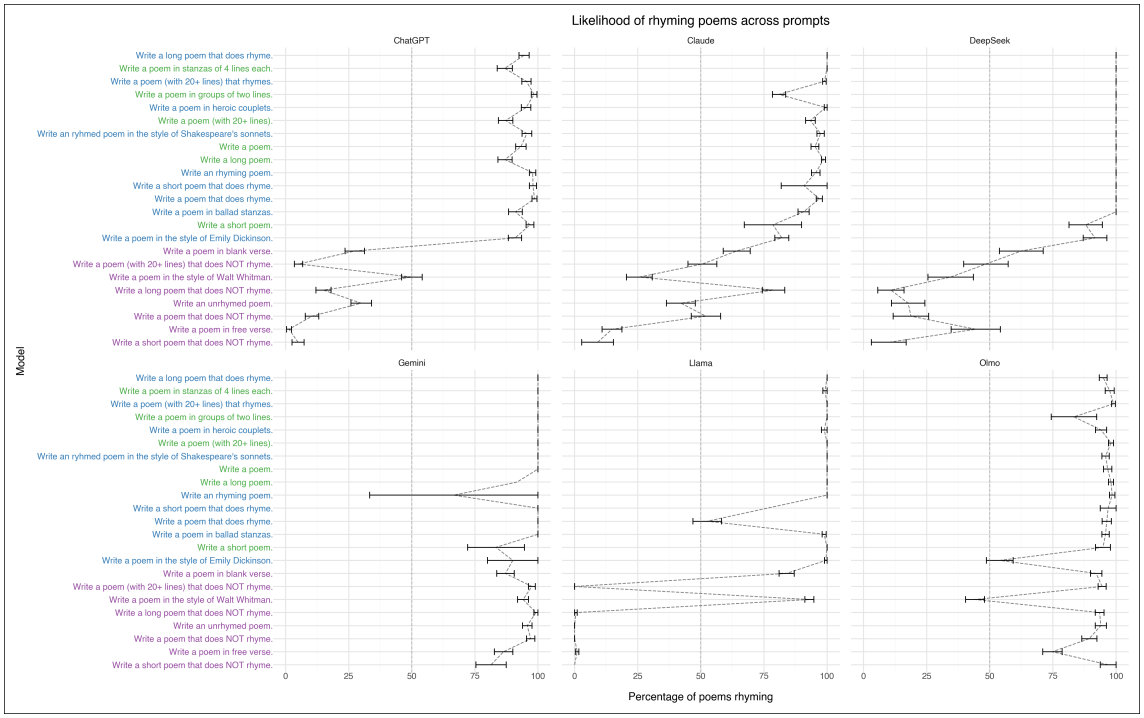

Figure 3: Frequency of rhyme in generative verse across a variety of prompts. Prompts categorized above as 'write a rhyming poem' are shown in green; 'write an unrhyming poem' in purple; and simply to 'write a poem' in blue. Whiskers show standard error around the mean for each model and prompt.

mal bias persists even when rhyme is not linguistically primed by prompts (Figure 4). The generative completions showed some responsiveness to historical patterns, producing more rhyming verse when completing pre-twentieth century poems and less when completing modern verse. However, no model reflected the full extent of rhyme's historical disappearance.

Statistical analysis revealed that all LLMs produced significantly more rhyme in their completions than historically documented for postwar poets. Effect sizes were large for Claude (Cohen's d = 1.46), Llama (d = 0.98), and OLMo (d = 0.84), medium for DeepSeek (d = 0.63), and small but significant for ChatGPT (d = 0.45). These results indicate that generative verse appears formally retrograde, as if temporally displaced between the nineteenth and twentieth centuries rather than reflecting contemporary poetic practice.

3.3 Examining training data biases

To test whether rhyme bias could be explained by training data composition, the analysis examined the presence of rhyming versus non-rhyming poetry in model training corpora using two complementary approaches. For open-source models, the study queried the Dolma dataset using the Allen Institute's "What's in my big data?" tool, following Walsh, Preus, and Antoniak's methodology [11]. For closed-source models, the analysis employed D'Souza and Mimno's approach of testing for memorized content through completion prompts.

The analysis examined four datasets: Chadwyck-Healey poems in open and closed training data, and canonical poems from the Poetry Foundation and Academy of American Poets in open and closed training data. Across all methods and datasets, poems found in LLM training data were not disproportionately rhyming compared to those absent from training data (Figure 5).

Comparison of means between found and unfound poems showed statistically insignificant dif-

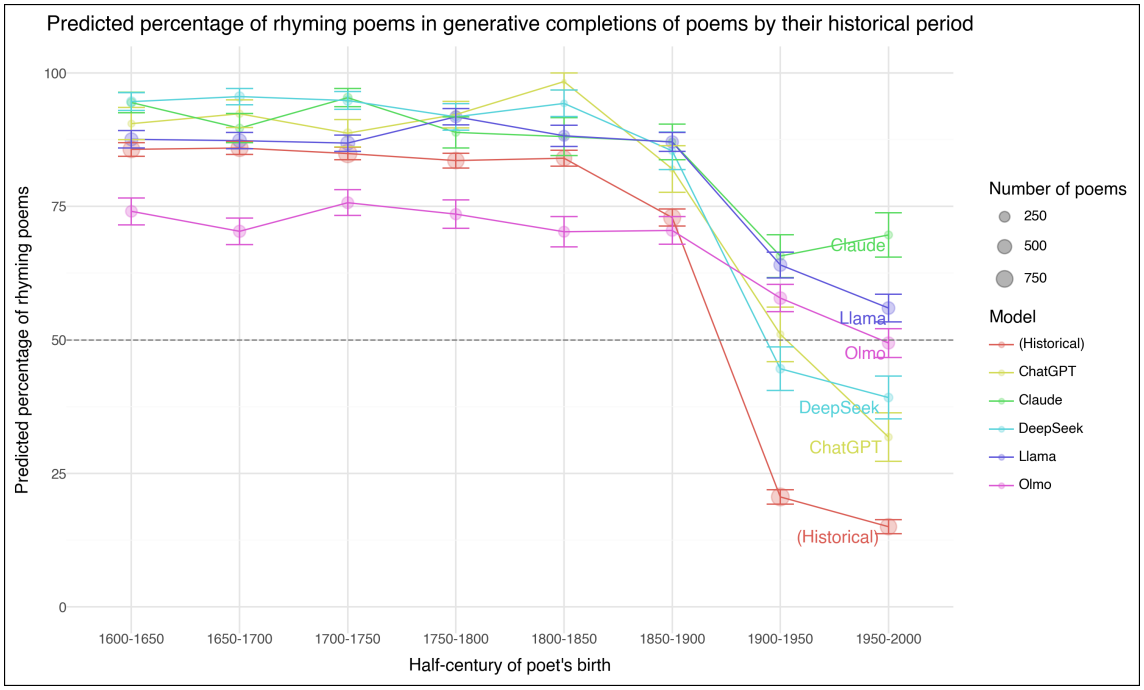

Figure 4: Frequency of rhymed verse across generative completions of historical poems. Points indicate mean likelihood; size indicates the number of poems per data point; whiskers show standard error.

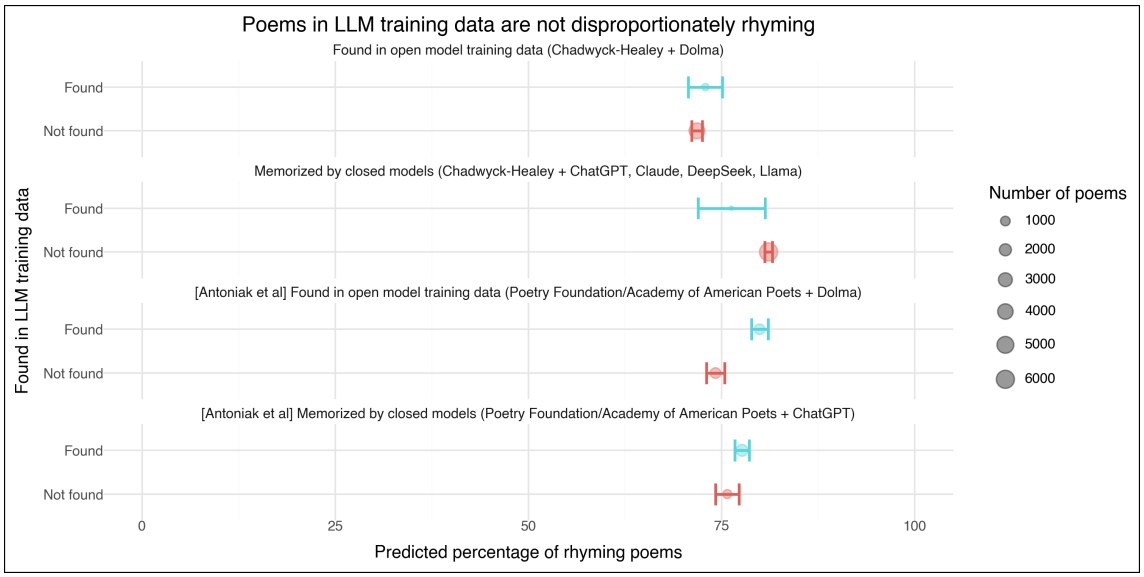

Figure 5: Poems in LLM training data are not disproportionately rhyming. Found poems (blue) vs. not found poems (red) across four different detection methods. Whiskers show standard error around the mean.

ferences in all cases except the open training dataset queried for canonical poems, which showed only a negligible effect size (Cohen's d = 0.14) below conventional significance thresholds. Moreover, very few poems from the historical corpora were found in closed-source LLM training data, and those found actually rhymed less often than those not found. These results demonstrate that LLMs' tendency to overuse rhyme cannot be explained by overrepresentation of rhyming poetry

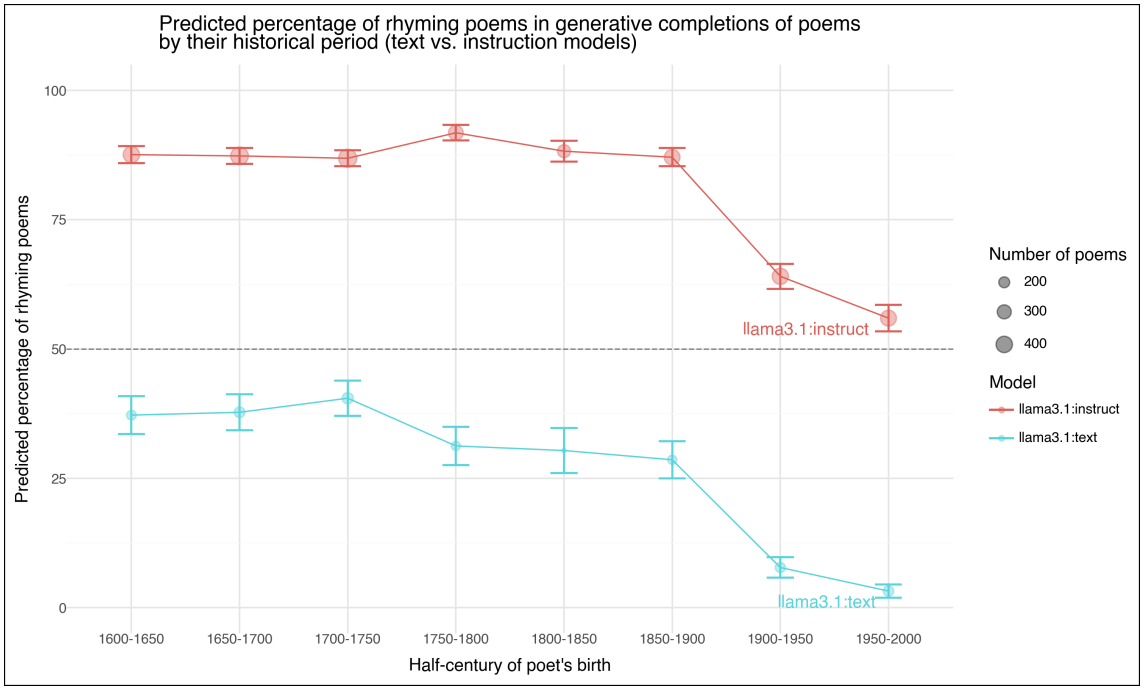

Figure 6: Frequency of rhymed verse across generative completions of historical poems. The two models here are both llama3.1, pretrained on the same data, but one (llama3.1:instruct) is 'instruction-tuned'—trained via reinforcement-learning human feedback (RLHF) to respond to instruction prompts—and the other (llama3.1:text) is a raw next token generator. Points indicate mean likelihood; size indicates the number of poems per data point; whiskers show standard error.

in training data, pointing instead to more fundamental characteristics of how these models process and generate poetic forms.

3.4 Instruction tuning effects

To investigate potential mechanisms underlying formal bias, the analysis compared rhyme frequencies between instruction-tuned and base versions of the same model. Using Llama 3.1, the study contrasted the instruction-tuned variant (Llama 3.1:instruct), which has been fine-tuned via reinforcement learning from human feedback to respond helpfully to user instructions, with the base model (Llama 3.1:text), which performs only next-token prediction without instruction-following capabilities (Figure 6).

The poem completion experiment revealed significant differences between these model variants across all historical periods. The instruction-tuned model consistently produced higher rates of rhyming poetry when completing historical poems, with the difference particularly pronounced for twentieth-century verse. For postwar poetry completions, the instruction-tuned model rhymed approximately 60% of the time compared to 10% for the base model, representing a statistically significant difference with large effect size (Cohen's $d > 1.0$).

This pattern held across all historical periods, with the instruction-tuned model showing systematically higher rhyme rates than its base counterpart. The finding suggests that processes designed to make models more helpful and responsive to user preferences may inadvertently bias them toward aesthetic forms perceived as more conventional or pleasing. The drive to satisfy user expectations appears to push models toward idealized representations of cultural forms, contributing to the formal conservatism observed across instruction-tuned systems.

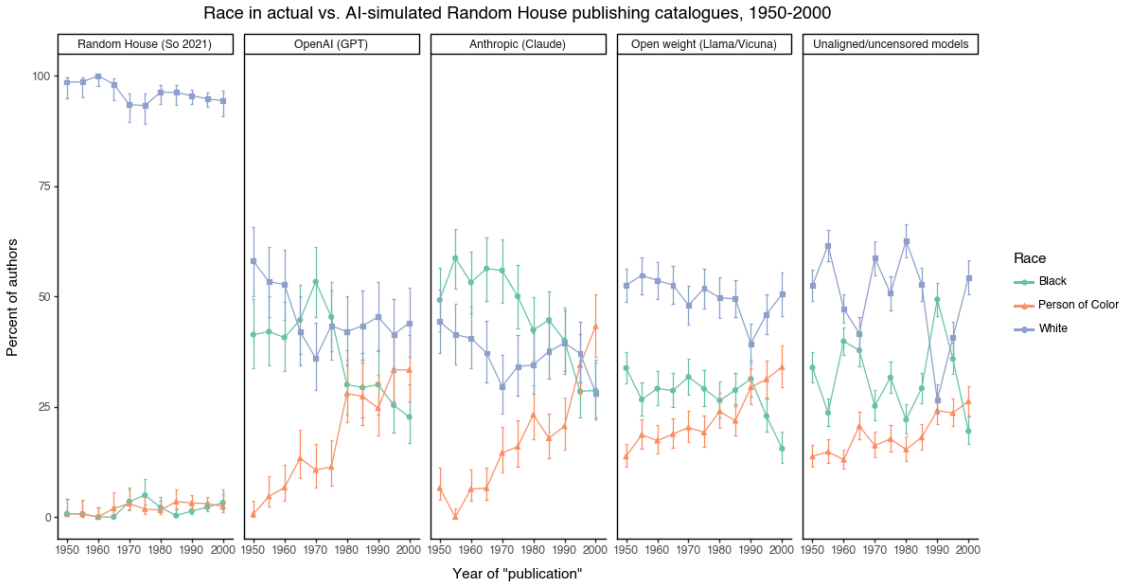

Figure 7: AI-generated list of authors and their race published by Random House in the second half of the twentieth century. 43,876 pseudo-authors were generated by 12 models. Data from actual Random House publication history taken from Richard So, *Redlining Culture*. Points indicate mean likelihood; whiskers indicate standard error.

3.5 Historical idealization in demographic data

Analysis of AI-generated historical publication data revealed parallel patterns of idealization in historical domains. When prompted to generate lists of authors published by Random House across the second half of the twentieth century, models produced demographic distributions that systematically diverged from documented historical reality in the direction of greater diversity.

The 43,876 pseudo-authors generated across 12 models showed white authors accounting for approximately 50% of Random House publications in the AI-generated data, compared to the actual historical rate of 90-95% documented by So [8]. In some years, models generated scenarios where Black authors were published more frequently than white authors, and other authors of color (Latinx, Asian, and Native American) showed meteoric rises from 5% in 1950 to 30-50% by 2000 in the generated data (Figure 7).

Similar patterns emerged in the analysis of gender representation in English-language fiction. The 113,869 pseudo-authors generated for the period 1800-2000 showed women accounting for 90-100% of fictional production in most models' outputs, compared to the actual historical range of 40-50% documented by Underwood, Bamman, and Lee [10]. This idealization occurred despite models having limited familiarity with the longer history of fiction, with nineteenth- and early twentieth-century texts accounting for only a small fraction of training data (Figure 8).

These findings parallel the formal idealization observed in poetry, suggesting that cultural collapse operates across both aesthetic and historical domains. Rather than reproducing historical biases by parroting training data, models appear to generate idealized versions of the past that obscure the very patterns of marginalization and exclusion that digital humanities scholarship seeks to illuminate.

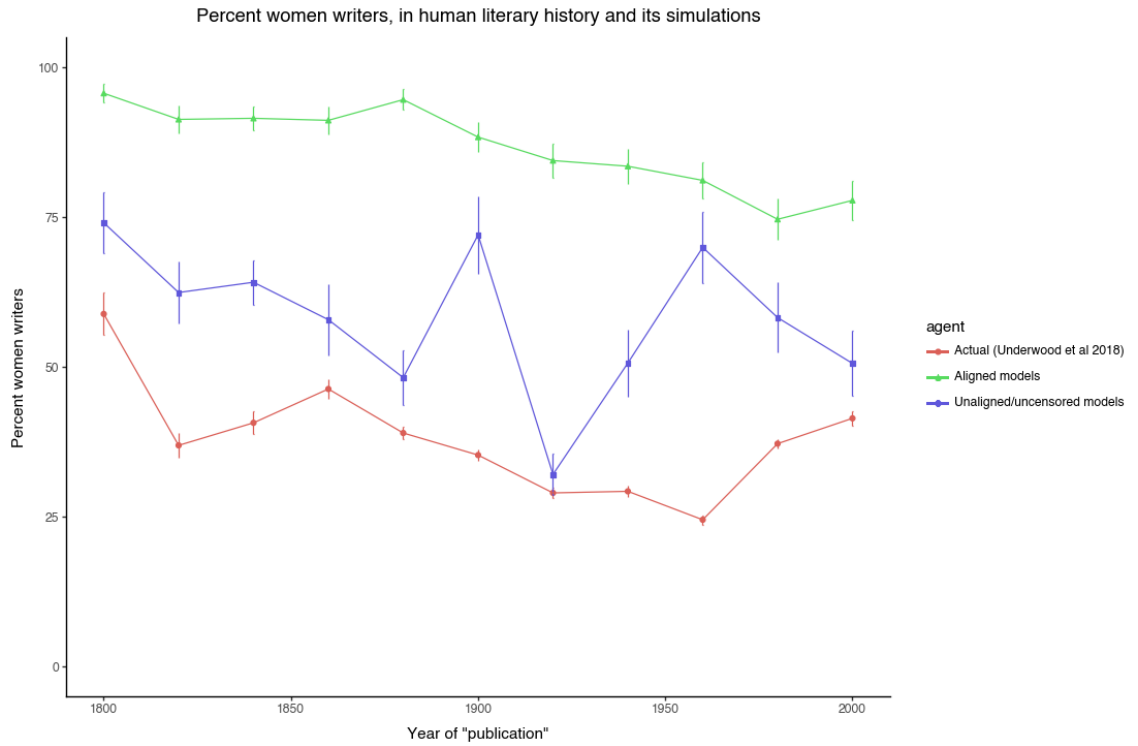

Figure 8: Actual historical distribution of female authors of Anglophone fiction, 1800-2000 (in red) vs. its AI-generated simulations. 113,869 pseudo-authors were generated by 12 models. Actual gender publication history taken from Underwood et al, "Transformation of Gender in English-language Fiction." Points indicate mean likelihood; whiskers indicate standard error.

4 Discussion

4.1 Evidence for cultural collapse

The systematic analysis across multiple domains provides converging evidence for the cultural collapse hypothesis. Large language models consistently generate cultural material that is more 'idealized'—aesthetically (in form) and ideologically (in representation)—than any historical period.

The poetry experiments demonstrate that this idealization operates through deep computational biases rather than simple training data artifacts. Models exhibit persistent formal stuckness toward rhyme that cannot be explained by prompt effects or overrepresentation of traditional poetry in training corpora. Even when explicitly instructed against rhyme, models continue producing rhyming verse at rates far exceeding the previous century of poetic practice—the historical site of most LLM training data. The poem completion experiments confirm this tendency persists even when formal features like rhyme are not explicitly primed, suggesting that cultural collapse represents an intrinsic characteristic of how these systems process aesthetic form.

The instruction tuning findings reveal one mechanism contributing to cultural collapse. Base models that perform only next-token prediction show significantly less bias toward traditional forms than their instruction-tuned counterparts designed to be helpful assistants. This suggests that the drive to be helpful inadvertently pushes models toward aesthetic forms they perceive as more pleasing or conventional, creating an unexpected aesthetic side effect of alignment processes.

The historical data generation experiments demonstrate that cultural collapse extends beyond aesthetic domains to historical representation. Models systematically generate idealized versions of publishing demographics that obscure historical patterns of exclusion and marginalization.

Rather than reproducing the biases present in their training data, these systems appear to "correct" historical reality in directions that align with contemporary values, creating revisionist narratives that flatten the very contradictions digital humanities seeks to recover.

4.2 Computational logic of idealization

These findings point toward what might be called a computational logic of idealization embedded in large language models. This logic operates through several interconnected mechanisms. First, models appear to privilege formal satisfaction over frustration, gravitating toward conventionally 'complete' or 'perfect' instantiations of cultural forms rather than the variations, experiments, and failures that characterize actual historical practice. Second, they exhibit a bias toward regularity over variation, consistently choosing predictable patterns over the unexpected developments that drive aesthetic innovation. Third, they demonstrate a preference for conformity over contradiction, smoothing out the tensions and conflicts that give cultural forms their historical complexity.

This computational logic of idealization helps explain why generative AI art produces the aesthetic qualities critics have termed 'slop.' The highly saturated colors, anime-inflected aesthetics, and cloying sentimentality that characterize AI-generated visual art may reflect the same underlying tendency toward idealized representation observed in poetry and historical data. These systems do not simply reproduce their training data but actively reshape cultural material according to implicit logics of perfection and satisfaction, producing what has been called "the average of our aesthetic desires" [7].

The historical dislocations in the observed model output suggests that cultural collapse involves a fundamental distortion of historicity within generative systems. Acting as what Lev Manovich and Emanuele Arielli call "quintessence machines," models appear to compress centuries of cultural development into idealized composites that satisfy contemporary expectations while maintaining traditional formal features [4]. This creates outputs that are simultaneously nostalgic and ahistorical, combining formal conservatism with values-based revisionism in ways that reflect neither past nor present cultural reality.

4.3 Implications for cultural transmission

As generative AI systems become increasingly ubiquitous in cultural production, their idealizing tendencies will likely reshape how cultural forms are transmitted across generations. The systematic bias toward formulaic outputs threatens to narrow the range of aesthetic possibilities encountered by future audiences, potentially creating feedback loops where increasingly idealized AI-generated content trains future models toward even greater conformity.

This has particularly troubling implications for cultural diversity and historical memory. If AI systems systematically flatten cultural complexity and generate revisionist versions of historical demographics, they may inadvertently undermine efforts to understand and address patterns of systemic exclusion. The idealized publication histories generated by models in this study do not simply fail to represent historical reality but actively obscure it, creating fictional narratives of diversity and inclusion that never existed.

4.4 Toward a generative formalism

Understanding cultural collapse requires new critical frameworks capable of analyzing AI-generated content as a distinct form of cultural artifact. This study proposes 'generative formalism' as an extension of traditional formalist methods to the generative domain. Just as historical formalist approaches examine how literary techniques create meaning in their historical contexts, generative formalist approaches investigate how computational processes algorithmically remediate those historical contexts into an alien form of cultural production.

This approach treats AI-generated content not as transparent representation but as specific artifacts that bear the traces of their historical training data and computational mediation. The formal conservatism observed in AI poetry reveals how LLMs selectively rewire aesthetic forms and traditions, while the demographic idealization in generated historical data illuminates the values and ideologies embedded in contemporary AI systems.

Future research in generative formalism might examine how different architectural choices, training procedures, and alignment methods affect distinct cultural outputs of AI systems. The instruction tuning findings in this study suggest that seemingly technical decisions about model development have profound implications for cultural reproduction. Understanding these relationships will become increasingly important as AI systems play a larger role in the present and future of culture.

References

[1] Anttila, Arto and Heuser, Ryan. "Phonological and Metrical Variation across Genres". In: *Proceedings of the Annual Meetings on Phonology* (2015). ISSN: 2377-3324. DOI: 10.3765/amp.v3i0.3679.

[2] D'Souza, Lyra and Mimno, David. "The Chatbot and the Canon: Poetry Memorization in LLMs". In: CHR 2023: Computational Humanities Research Conference. Paris, France, Dec. 6–8, 2023.

[3] Heuser, Ryan. "Prosodic". Github. URL: https://github.com/quadrismegistus/prosodic (visited on 09/29/2024).

[4] Manovich, Lev and Arielli, Emanuele. *Artificial Aesthetics: Generative AI, Art and Visual Media*. URL: https://manovich.net/index.php/projects/artificial-aesthetics.

[5] Shumailov, Ilia, Shumaylov, Zakhar, Zhao, Yiren, Papernot, Nicolas, Anderson, Ross, and Gal, Yarin. "AI Models Collapse When Trained on Recursively Generated Data". In: *Nature 631*, no. 8022 (July 25, 2024), pp. 755–759. DOI: 10.1038/s41586-024-07566-y.

[6] "Slop, n.² Meanings, Etymology and More | Oxford English Dictionary". URL: https://www.oed.com/dictionary/slop_n2 (visited on 07/19/2025).

[7] Smith, Naomi and Southerton, Clare. "AI and Aesthetic Alienation: The Image and Creativity in Contemporary Culture". In: *Social Science Computer Review* (July 18, 2025). DOI: 10.1177/08944393251361449.

[8] So, Richard Jean. *Redlining Culture: A Data History of Racial Inequality and Postwar Fiction*. New York: Columbia University Press, 2020. ISBN: 978-0-231-19772-4.

[9] "The Internet's AI Slop Problem Is Only Going to Get Worse". URL: https://nymag.com/intelligencer/article/ai-generated-content-internet-online-slop-spam.html (visited on 07/18/2025).

[10] Underwood, Ted, Bamman, David, and Lee, Sabrina. "The Transformation of Gender in English-Language Fiction". In: *Journal of Cultural Analytics 3*, no. 2 (Feb. 13, 2018), p. 11035. DOI: 10.22148/16.019.

[11] Walsh, Melanie, Preus, Anna, and Antoniak, Maria. "Sonnet or Not, Bot? Poetry Evaluation for Large Models and Datasets". In: *Findings of the Association for Computational Linguistics*. Miami, Florida, USA: Association for Computational Linguistics, Oct. 10, 2024. DOI: 10.48550/arXiv.2406.18906. arXiv: 2406.18906 [cs].

[12] Walsh, Melanie, Preus, Anna, and Gronski, Elizabeth. “Does ChatGPT Have a Poetic Style?” In: CHR 2024: Computational Humanities Research Conference. Aarhus, Denmark: arXiv, Oct. 20, 2024. arXiv: 2410.15299 [cs]. URL: http://arxiv.org/abs/2410.15299 (visited on 10/26/2024).