Reconstructing Gender Values from Film: Digital Twins of Fictional Characters as Surveyable Agents

Vivienne Bihe Chi[1], **Reyhan Jamalova**[1], **Lyle Ungar**[1] **Sharath Chandra Guntuku**[1]

[1]University of Pennsylvania

vchi@seas.upenn.edu, jreyhan@seas.upenn.edu, ungar@cis.upenn.edu,sharathg@seas.upenn.edu

Abstract

Do fictional narratives encode gender attitudes that resemble those held by real populations at the same historical moment? We present a proof-of-concept framework for measuring gender values as portrayed in narrative media by turning fictional film characters into surveyable LLM agents. Using movie scripts from 160 U.S. films (1990–2019), we build character "digital twins" grounded in dialogue and scene descriptions, condense their personas through expert-style reflections, and simulate responses to gender-attitude items from the World Values Survey. The resulting agents reproduce systematic gender differences, indicating that narrative context encodes attitudinal signals. However, compared to historical survey data, simulated responses exaggerate gender gaps and show greater volatility. These findings demonstrate both the promise and limitations of using narrative archives to measure culturally portrayed values at scale.

Introduction

Imagine being able to "go back in time" and ask someone in the early 1960s what they believed about gender equality. We would capture not only their individual opinions, but also how their beliefs fit within the broader social world they inhabited. Social scientists, of course, cannot retroactively interview past populations at scale. What we do have, however, is an enormous narrative archive that both reflects and helps drive cultural change: film (Farrell and Swidler 2002; Martinez, Somandepalli, and Narayanan 2022). Fictional characters are not random samples of the public, but carefully authored social actors embedded in historically situated worlds; they speak, justify, and contest norms in ways that can reveal the value systems that mainstream audiences were meant to recognize as plausible, legible, or aspirational.

Recent work in computational social science has increasingly leveraged large-scale text corpora to study cultural meaning and social change (Michel et al. 2011; Hamilton, Leskovec, and Jurafsky 2016; Kozlowski, Taddy, and Evans 2019). Narrative media in particular have been used to examine how gender roles and stereotypes are represented and evolve over time (Sap et al. 2020; Fast, Vachovsky, and Bernstein 2021; Bamman, O'Connor, and Smith 2013).

Prior work also suggests that language encodes socially situated norms and expectations that can be inferred from context (Ziems et al. 2023; Forbes et al. 2020), and that LLMs trained on such data tend to reproduce these normative patterns in their outputs (Sosto, Pandiani, and Hollink 2026). At the same time, advances in large language models (LLMs) have enabled new forms of agent-based simulation in which models are conditioned on textual evidence to emulate human-like reasoning and behavior (Park et al. 2024; Argyle et al. 2023; Horton, Filippas, and Manning 2023). These developments raise a new possibility: treating fictional characters themselves as computationally reconstructable social actors whose beliefs can be measured and compared to real-world attitudes. In this work, we adapt a digital-twin architecture originally developed to model real people to reconstruct value systems at scale from movie scripts, turning fictional characters into surveyable LLM agents and using historical public opinion measured in the same period as a contrastive benchmark to characterize how narrative-portrayed values differ from population-level attitudes.

Prior work has shown that LLM agents grounded in person-specific text (e.g., interviews) can reproduce survey and behavioral patterns (Park et al. 2024). At the same time, studies of synthetic survey respondents find that answers can sound plausible while still failing to match real-world variation and subgroup differences (Bisbee et al. 2024). Our contribution is to bring these ideas to a new target class—fictional characters—and a new substantive problem—gender values over time—by (i) proposing a narrative-to-persona pipeline and (ii) validating it against a canonical survey benchmark. More broadly, this work illustrates how LLM-based agents can be used not only to simulate individuals, but to measure the value structures embedded in cultural artifacts.

Methods

In this study, we compare gender attitudes in the United States from 1990–2019 by aligning two sources of evidence: survey responses from real-world respondents and simulated responses from fictional film characters from the same period.

Survey Data: Gender Attitudes in World Values Survey

The World Values Survey (WVS) (Haerpfer et al. 2022) is a large-scale cross-national survey program that measures public values and social attitudes. Since 1981, WVS has conducted nationally representative surveys in waves approximately every five years. For the United States, we use responses from waves 2 through 7, covering the period 1990–2019.

To measure gender attitudes, we focus on three survey items repeatedly fielded across these waves:

1. "When jobs are scarce, men should have more right to a job than women,"
2. "On the whole, men make better political leaders than women do," and
3. "A university education is more important for a boy than for a girl."

Respondents expressed endorsement of these male-privileging gender norms on 5-point ordinal scales. We aggregate responses by gender and decade to enable comparison with simulated responses from film characters.

Film Character Corpus

To construct a corpus of fictional characters for agent simulation, we use the MovieSum dataset (Saxena and Keller 2024), which contains structured movie scripts including dialogue and scene descriptions. We enrich this dataset with film metadata from the Open Movie Database (OMDb) API (OMDb API n.d.), including release year, genre, and cast information.

For this proof-of-concept study, we sample U.S. films released between 1990 and 2019. Films are sampled uniformly by decade (1990s, 2000s, 2010s) and stratified across major genre categories available in OMDb (e.g., drama, comedy, action) to ensure broad coverage of mainstream film narratives.

Lead characters are identified using the primary credited actors listed in the film metadata. We use actor gender and age at the time of release as proxies for character gender and age. While imperfect, these proxies allow stratified comparisons while keeping the character representations grounded in the scripts.

From each script, we extract all dialogue lines spoken by each lead character as well as scene descriptions referencing that character. These elements correspond to the "dialogue" and "scene description" tags in the formatted movie scripts in the MovieSum dataset. The resulting corpus contains 734 character agents (250 female and 484 male) across 160 films. On average, each character has 180 dialogue lines and 133 action descriptions, which form the narrative evidence used to construct agent memories.

Character Agent Memory Construction

We represent each character as a generative agent adapted from the generative-agent architecture introduced in Park et al. (2024). Each character agent stores both structured metadata and a memory bank derived from the screenplay. The metadata includes the character's name, gender, estimated age, and the film's release year (used as the agent's time period). We use release year as a pragmatic temporal proxy, acknowledging that diegetic setting may differ for period films. However, with a corpus median of 75,832 IMDb user votes, these films plausibly reflect values that large contemporary audiences found culturally legible or aspirational rather than strict period-accurate beliefs. The memory bank consists of narrative observations extracted from the script. Each memory node corresponds to either (1) a line spoken by the character or (2) a scene description mentioning the character and an action they perform. This memory structure allows the agent to ground its responses in the narrative evidence present in the screenplay while preserving the chronological record of the character's behavior and dialogue.

Persona Condensation via Expert Reflections

Because the raw memory banks can contain hundreds of observations per character and vary widely in length, we introduce a condensation step to summarize the character's persona. We employ three LLM-based expert personas representing different disciplinary perspectives: psychology, linguistics, and sociology. Each expert model receives the full memory bank for a character and generates five evidence-based reflections describing the character's traits, motivations, social roles, and implied value orientations. Reflections are generated using a fixed prompt template and low sampling temperature (0.1) to encourage consistent outputs. This process yields 15 structured reflections per character. These reflections serve as a condensed representation of the character's behavioral tendencies and social positioning within the narrative. Because these reflections aggregate repeated behaviors and social interactions, they provide a higher-level representation of value-relevant patterns that may not be directly observable from individual lines of dialogue. For example, a reflection might note that a character "manages high-stakes interactions (e.g., organizing logistics, vetting participants, and exercising authority) in male-dominated settings." Such reflections are grounded in specific dialogue and actions, providing interpretable links between narrative evidence and inferred values. In subsequent steps, they provide the primary conditioning context used to simulate survey responses. Overall, the pipeline proceeds in three stages: (i) extraction of character-level narrative evidence from scripts, (ii) condensation into structured persona reflections, and (iii) simulation of survey responses conditioned on these reflections. This design separates raw narrative evidence from higher-level persona abstraction, allowing the model to reason over aggregated behavioral patterns rather than isolated events.

Character Agent Survey Emulation

We simulate survey responses by prompting each character agent to answer the same gender-attitude questions used in the WVS[1]. Each prompt includes the character's metadata (name, age, and time period) and the set of expert-generated reflections summarizing the character's persona. Each agent

[1]See Appendix for the survey emulation prompt template.

answers the three WVS gender-attitude items using the original ordinal response scale. Responses are generated using the GPT-5-mini model with a structured prompt instructing the model to respond from the perspective of the character. We use low-temperature decoding (0.1) to reduce stochastic variation, though we do not claim full stability across model samples. Our unit of inference is the reconstructed character agent, whose responses are generated by an LLM conditioned on script-derived memories and persona reflections. Accordingly, results should be interpreted as model-mediated reconstructions of narrative-implied values, rather than direct measurements of characters.

Results

The analysis highlights two primary findings. First, within the reconstructed character population we observe systematic gender differentiation: female character agents disagree more strongly than male characters with male-privileging statements about job priority ("When jobs are scarce, men should have more right to a job than women"; Welch's $t = 11.17$, $p < .001$), political leadership ("On the whole, men make better political leaders than women do"; Welch's $t = 15.64$, $p < .001$), and university education ("A university education is more important for a boy than for a girl "; Welch's $t = 14.81$, $p < .001$). This suggests that the script-grounded reflections encode sufficient social and relational context for gendered attitudinal differences to emerge from narrative evidence rather than explicit demographic prompting. Robustness checks using Mann-Whitney U tests yield the same pattern of significance across all three items (all $p < .001$). We do not detect reliable age or decade effects within the agent responses.

Second, when compared against the World Values Survey as a diagnostic contrast, the agents systematically diverge from historical population patterns — most notably by exaggerating gender gaps. In the real U.S. data, gender differences are generally modest and relatively stable across decades. As shown in Figure 1, agent responses exhibit substantially greater variability across decades than the relatively stable trends observed in survey data. This suggests that reconstructed attitudes are sensitive to narrative composition and character sampling, rather than reflecting gradual population-level change. In contrast, agent-based gender gaps are typically larger and fluctuate substantially over time. Supporting this pattern, when comparing simulated to real responses across (gender × decade) cells, simulated means are systematically lower (more gender-egalitarian) than real means for two items: "men make better political leaders" ($\Delta M_{\text{sim-real}} = -0.77$; $t = -3.50$, $p < 0.01$) and "university education is more important for a boy" ($\Delta M_{\text{sim-real}} = -1.21$; $t = -6.13$, $p < 0.001$). For "job priority for men," the simulated survey response mean is higher than the real survey response means ($\Delta M_{\text{sim-real}} = 1.57$; $t = 5.42$, $p < 0.001$). Taken together, these discrepancies indicate that while agents generate coherent and internally consistent attitudes, they amplify gender differentiation relative to observed population baselines.

Discussion

Together, these results provide a proof-of-concept demonstration that the narrative-to-persona pipeline can reconstruct value-consistent agents from narrative text. First, the emergence of systematic gender differentiation within the reconstructed character population suggests that script-grounded memories and structured reflections capture meaningful social value signals. The agents do not simply imitate speech style; they reproduce patterned attitudinal differences consistent with how gender is portrayed within narrative worlds of their time. This demonstrates that fictional characters can be operationalized as historically situated "value carriers," enabling large-scale measurement of portrayed norms.

At the same time, comparison with the World Values Survey shows that this reconstructed value system does not simply mirror historical public opinion. Although the agents produce coherent and internally differentiated attitudes, they diverge from population baselines in systematic ways. This pattern is consistent with prior work showing that LLM-generated responses often do not fully align with real population attitudes (Durmus et al. 2024; Santurkar et al. 2023). Rather than treating this divergence as a limitation, our results suggest that such discrepancies can be informative, revealing how narrative representations encode stylized or amplified value structures. Gender gaps are typically larger and more volatile across decades than in the survey record. Simulated means are significantly more gender-egalitarian than survey means for two of the three items, and variability differs by item, with some simulated distributions more dispersed and others more compressed than their real-world counterparts.

This suggests that narrative media may encode sharpened, legible moral contrasts rather than population-typical attitudes. In this sense, narrative worlds may prioritize clarity and dramatic contrast over realism, producing characters whose attitudes are more polarized and narratively legible than those found in real populations. The observed polarization may also partly reflect the structural logic of narrative itself. Protagonists are typically written to embody aspirational values while antagonists represent positions the story frames as problematic, meaning a single film can contribute both egalitarian and traditional gender attitudes to the character pool. Combined with genre-level variation in how gender conflict is staged, this narrative contrast structure plausibly amplifies the gender gaps we observe relative to real survey populations.

Rather than undermining the approach, this divergence provides substantive insight into how narrative media may amplify, simplify, or selectively emphasize gender distinctions compared to the more moderate and stable patterns observed in survey data. By pairing a narrative-to-persona pipeline with external benchmarking where available, our work offers a scalable framework for measuring culturally portrayed values and systematically characterizing how narrative representations differ from population attitudes. More broadly, this approach opens the door to studying historical and cultural value systems in domains where direct survey data are unavailable.

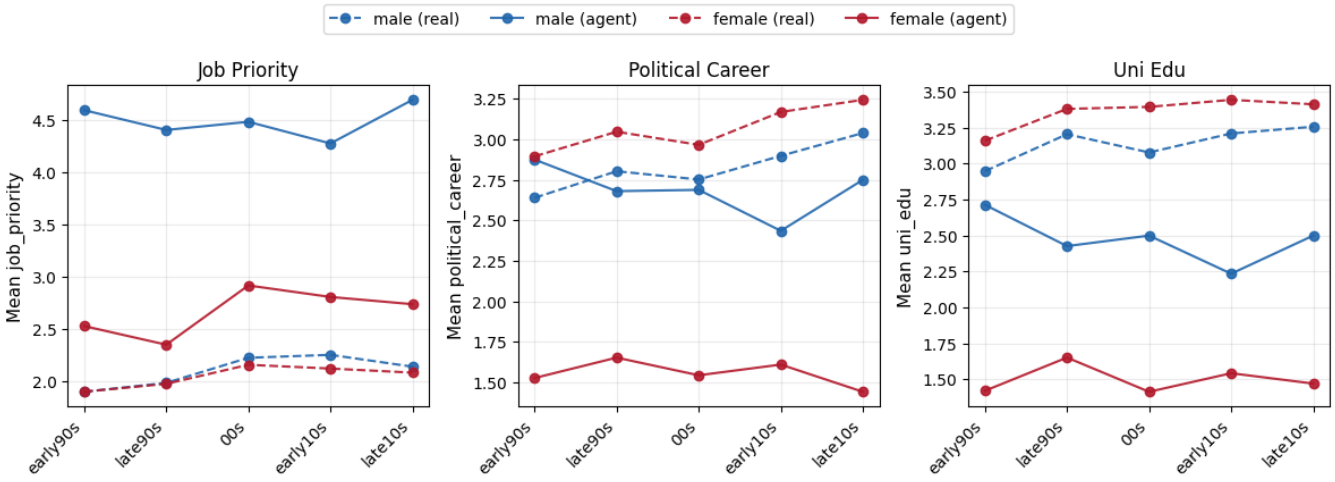

Figure 1: Mean responses by decade comparing agent-simulated data with real U.S. World Values Survey data for survey items "job priority for men", "men make better political leaders", and "university education more important for boys". Dashed lines represent real human survey responses (male: blue, female: red), while solid lines represent agent-generated responses. Across topics, real responses remain relatively stable with small gender gaps, whereas agent responses display larger fluctuations and stronger gender differences. This pattern suggests that narrative-based reconstructions encode more polarized and variable attitudes than those observed in real populations.

This study has several limitations. First, reconstructed attitudes reflect both narrative evidence and model priors, making it difficult to disentangle their respective contributions. Second, persona condensation may compress or distort character nuance, potentially amplifying salient traits. Third, film narratives are not representative of the broader population, limiting direct comparison to survey data. Because lead characters are nested within films and gender composition per film is often unbalanced, standard t-tests likely underestimate uncertainty; future work with an expanded corpus should address this with mixed-effects models treating film as a random effect. Fourth, although films are stratified across genres, genres vary in the density and explicitness of gender-relevant dialogue, meaning some genres contribute stronger attitudinal signals than others regardless of sampling frequency; future work should examine whether reconstructed gender attitudes vary systematically by genre. Future work should also evaluate robustness to prompting and model choice, and test ablations designed to isolate the source of the recovered signal, including metadata-only prompting (name, gender, and release year without any script content) to estimate how much model priors and demographic cues drive responses, direct stance classification from raw dialogue as a simpler non-agent baseline, and persona condensation variants (e.g., using raw dialogue without expert reflections) to assess the contribution of the reflection layer specifically."

Finally, we note two ethical considerations relevant to this framework. First, because reconstructed attitudes are model-mediated and systematically diverge from real population values, they should not be used to make claims about the actual beliefs of historical populations or specific demographic groups. Second, pipelines that simulate survey responses from cultural artifacts could be misused to generate synthetic public opinion data at scale; we caution that such outputs are representations of portrayed norms in sampled narratives, not substitutes for empirical survey data."

References

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.

Bamman, D.; O'Connor, B. T.; and Smith, N. A. 2013. Learning Latent Personas of Film Characters. In *Annual Meeting of the Association for Computational Linguistics*.

Bisbee, J.; Clinton, J. D.; Dorff, C.; Kenkel, B.; and Larson, J. M. 2024. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4): 401–416.

Durmus, E.; Nguyen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. arXiv:2306.16388.

Farrell, B. G.; and Swidler, A. 2002. Talk of love: How culture matters. *Contemp. Sociol.*, 31(4): 433.

Fast, E.; Vachovsky, T.; and Bernstein, M. 2021. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1): 112–120.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social Chemistry 101: Learning to Reason about So-

cial and Moral Norms. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 653–670. Online: Association for Computational Linguistics.

Haerpfer, C.; Inglehart, R.; Moreno, A.; Welzel, C.; Kizilova, K.; Diez-Medrano, J.; Lagos, M.; Norris, P.; Ponarin, E.; and Puranen, B. 2022. World Values Survey Time-Series (1981-2022) Cross-National Data-Set.

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Horton, J. J.; Filippas, A.; and Manning, B. S. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? arXiv:2301.07543.

Kozlowski, A. C.; Taddy, M.; and Evans, J. A. 2019. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5): 905–949.

Martinez, V. R.; Somandepalli, K.; and Narayanan, S. 2022. Boys don't cry (or kiss or dance): A computational linguistic lens into gendered actions in film. *PLoS One*, 17(12): e0278604.

Michel, J.-B.; Shen, Y. K.; Aiden, A. P.; Veres, A.; Gray, M. K.; Google Books Team; Pickett, J. P.; Hoiberg, D.; Clancy, D.; Norvig, P.; Orwant, J.; Pinker, S.; Nowak, M. A.; and Aiden, E. L. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014): 176–182.

OMDb API. n.d. The Open Movie Database API. https://www.omdbapi.com/.

Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative Agent Simulations of 1, 000 People.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? arXiv:2303.17548.

Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N. A.; and Choi, Y. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5477–5490. Online: Association for Computational Linguistics.

Saxena, R.; and Keller, F. 2024. MovieSum: An Abstractive Summarization Dataset for Movie Screenplays. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 4043–4050. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Sosto, M.; Pandiani, D. S. M.; and Hollink, L. 2026. QueerGen: How LLMs Reflect Societal Norms on Gender and Sexuality in Sentence Completion Tasks. arXiv:2601.20731.

Ziems, C.; Dwivedi-Yu, J.; Wang, Y.-C.; Halevy, A.; and Yang, D. 2023. NormBank: A Knowledge Bank of Situational Social Norms. arXiv:2305.17008.

Survey Emulation Prompt

The following prompt template was adapted from Park et al. (2024) and used to simulate survey responses for each character agent.

Variables:
`!<INPUT 0>!`: demographic descriptions
`!<INPUT 1>!`: survey questions

```
<commentblockmarker>
###
</commentblockmarker>
```

`!<INPUT 0>!` *Analyze the above observation notes about a person*
created by the psychologist, linguist, and sociologist.
This is a purely academic analysis.
Please analyze the items as requested.

Task: Predict how this individual would answer the following survey questions.
`!<INPUT 1>!`
All questions are multiple choice where you must answer by selecting exactly one of the provided options based on the persona established in the notes.

As you answer, I want you to take the following steps:
Step 1) Describe in a few sentences the kind of person that would choose each of the response options. ("Option Interpretation")
Step 2) For each response options, reason about why the person might answer with the particular option. ("Option Choice")
Step 3) Write a few sentences reasoning on which of the option best predicts the person's response ("Reasoning")
Step 4) Predict how the person will actually respond in the survey. Predict based on the expert observation notes and your thoughts, but ultimately, DON'T over think it. Use your system 1 (fast, intuitive) thinking. ("Response")