Do Language Models Pass the Bechdel Test? Auditing Gender Biases in LLM-Generated Screenplays

MEGHA N. GOVINDU, University of Pennsylvania, USA
STEPHANIE T. WANG, University of Pennsylvania, USA
SORELLE A. FRIEDLER, Haverford College, USA
DANAÉ METAXA, University of Pennsylvania, USA

As large language models (LLMs) are increasingly used in media production from journalism to filmmaking, what impact do they have on the stories being told? Prior work has shown LLMs to perpetuate social biases, including those related to gender. We complement existing literature on gender bias in LLM outputs by auditing the network structure of LLM-generated movie screenplays through automating the Bechdel test, a popular measure of women's representation in literary and film works. We also introduce the use of social network analysis measures to further analyze representational bias in LLM-generated scripts. We evaluate screenplays generated by three state-of-the-art LLMs (GPT-5, Gemini 3 Pro, and Claude Sonnet 4.5) against 768 corresponding human-written screenplays, finding that human-written scripts are more likely to pass the Bechdel test. However, other network analyses, like centrality, homophily, and triadic relationships demonstrate that in some cases LLM-scripts have less bias, although all script types demonstrate some representational bias under most measures. We conclude by discussing the continued need for further quantitative assessments of media representations and AI-generated content.

CCS Concepts: • **Computing methodologies** → *Natural language processing*; • **Human-centered computing** → *Empirical studies in HCI*; *Social network analysis*; • **Applied computing** → *Media arts.*

Additional Key Words and Phrases: AI auditing, social network analysis, representational bias, text generation

ACM Reference Format:
Megha N. Govindu, Stephanie T. Wang, Sorelle A. Friedler, and Danaé Metaxa. 2026. Do Language Models Pass the Bechdel Test? Auditing Gender Biases in LLM-Generated Screenplays. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3805689.3812208

1 Introduction

In 2023, The Writers Guild of America (WGA) went on a 148 day strike, with AI usage in screenwriting a primary aspect in negotiations with the Alliance of Motion Picture and Television Producers [25]. Screenwriters on strike warned that AI could not provide a complex representation of human life, and especially of the experiences of marginalized groups: "They want to replace us with AI but I can tell you now definitively that AI does not have the trauma, the joy, of the lived experience [10]." Since the strike, the move towards replacing screenwriters has continued. In 2024, Lionsgate gave a startup permission to train their new model on the Lionsgate content

Authors' Contact Information: Megha N. Govindu, meggov@sas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Stephanie T. Wang, stephtw@seas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Sorelle A. Friedler, sorelle@cs.haverford.edu, Haverford College, Haverford, Pennsylvania, USA; Danaé Metaxa, metaxa@upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812208

library [31]. Most recently, OpenAI is backing an animated feature-length film made with an "AI-assisted workflow," set to debut at the Cannes Film Festival in 2026 [53].

We learn about ourselves and the world at the movies. Representation of marginalized identities in film shapes self-identity, experiences of discrimination, and individuals' capacities to imagine roles in society for themselves and others [30, 47]. Yet there is copious, well-documented evidence that large language models (LLMs) and other forms of generative AI exhibit biases on the basis of race [21], gender [33, 59], and a range of other attributes, as well as various intersections (see Section 2.3). Given this context, we ask: as LLMs are increasingly used in the film industry, will such biases affect the artistic content being produced?

This paper takes up this question by examining potential gender bias against women in LLM-generated movie scripts, and using these screenplays as a lens through which to examine LLMs for representational bias. One popular measure of the representation of women in film is the Bechdel test. A film passes the test if there are 1) two women who 2) talk to each other about 3) something besides a man [7]. We implement the Bechdel test and use it to measure the representation of women in LLM-generated scripts from three state-of-the-art LLMs (GPT-5, Gemini 3 Pro, and Claude Sonnet 4.5) and 768 corresponding real, human-written movie screenplays. Screenplays are generated using synopses of films whose real scripts are available and prompting the LLMs to first create scene lists and then scene-by-scene to generate scripts for those scenes. To analyze this data, our implementation of the Bechdel test creates a character network where each character is a node and each dialog interaction in the script is represented as an edge in the network. Using this network, we further analyze the representation of women characters on a range of social network analysis (SNA) measures, including homophily and centrality.

Our results show that human-written scripts are far more likely to pass the Bechdel test than LLM-generated ones. Examining other representational bias measures based on character networks yields more varied results; GPT-generated scripts outperform human ones on proportion of interactions involving women, and also have more heterophily and triadic relationships between female characters.

While these relative representational differences between generated and human scripts are important, overall we find that none of these script types have stronger representation of women than men; across all script types and centrality measures studied, men are more central within the network than women. The Bechdel test was introduced to point out the lack of representation of women in media; our representational analysis shows that these concerns remain across both human and LLM-generated screenplays.

2 Related Work

This study builds upon three main areas of work: media studies, analyses of social network representations of interpersonal relationships, and LLM bias audits.

2.1 Biases in Film and the Bechdel Test

Considerable research across disciplines from communication to media studies has examined gender and racial/ethnic representation in film and literature. Such work is motivated by the understanding that film, as a dominant form of media culture, both reflects and shapes social values and identities, offering insight into the contexts in which cultural narratives are produced [24].

Methodologically, prior work relied on highly manual studies in which researchers code the gender and ethnicity of characters to examine their presence and centrality within narratives [36, 52, 55]. More recently, advances in computational methods and data availability have enabled large-scale, data-driven approaches to quantifying stereotypes and biases in film [5, 27, 48, 66]. For instance, Bamman et al. [5] use computer vision to analyze gender and race representation across 2,307 films, revealing increasing diversity over time but persistent under-representation of Black actors in award-nominated films and of non-White actors and women in leading

roles. Other works analyze screenplays for gender bias using connotation frames [48] and by analyzing socio- and psycho-linguistic dimensions of dialogue [66].

In 1985, Alison Bechdel's comic strip "The Rule" introduced what became known as the Bechdel test. In order to pass the test, a given film or piece of fiction must satisfy three conditions: it must (1) have at least two named women in it, (2) who talk to each other, (3) about something other than a man [7]. Originally a humorous commentary and critique, the Bechdel test has since become a mainstream heuristic for benchmarking female representation in media [57].

Research has adapted the Bechdel test for computational analysis of gender representation at scale. Garcia et al. [17] quantify the Bechdel test by constructing a Bechdel score and applying this to movie scripts and MySpace and Twitter dialogue datasets to compare gender roles in both fictional and real-world online dialogues. Agarwal et al. [3] builds on this work to automate the Bechdel test, building a classifier and studying the effectiveness of various linguistic and SNA features, finding that SNA features are more effective at accurately labeling scripts with Bechdel scores. They also find that female characters exhibit lower centrality (in the movie's social network) in movies that fail the Bechdel test. Building on Agarwal et al. [3], we implement a computational Bechdel test and validate our implementation against theirs before extending to it to machine-generated screenplays. We apply this test alongside other network measures to to both human-written and LLM-generated screenplays to evaluate the degree to which gender biases in screenplays are amplified by language models.

2.2 Social Network Analysis

Social network analysis (SNA) uses network structures to model social structures, revealing patterns in group formation, influence, and connectivity. In SNA, networks are built of nodes (representing entities, like people) and edges (relationships between those entities). These methods have been applied to study human relationships in online social networks [56] and, more recently, to understand structural bias and demographic-driven disparities in social networks. For a survey, see Saxena et al. [49].

Social network analysis has also been used to analyze media such as film, producing empirical analyses of social ties among film characters [3, 23, 27, 34, 43, 63]. Agarwal et al. [2] develop methods for parsing movie screenplays into social networks to study character interactions. Weng et al. [63] introduce RoleNet, a method that constructs social networks by connecting characters appearing in the same scene, and develops an algorithm to classify characters into leading or supporting roles and identify various structures of social ties. Park et al. [43] build upon these ideas with Character-Net, which constructs the network from the accumulation of dialogue graphs between characters, treating dialogue as a relationship between two people. They use Character-Net to develop an algorithm that classifies characters into major, minor, and extra roles. StoryRoleNet is another project that integrates both video-based and subtitle-based networks by segmenting films into story units and computing relationship weights from both modalities within each unit before merging them into a social network for analysis [34]. Kagan et al. [23] also construct a social network representation of movies, and use SNA measures such as centrality and triangle analysis to analyze gender biases in women's portrayal in film. They also develop an automated classifier to predict whether a film passes the Bechdel test.

The methods above analyze social networks extracted from existing media. Recent work has begun applying these methods to LLM-generated networks as LLMs become capable of social simulation [42]. Recent work related to SNA explores LLM's abilities to *generate* social networks, and to simulate interactions over those networks. Papachristou and Yuan [41] study the network formation behaviors of multiple LLM agents, benchmarking them against human decisions. They find that LLMs exhibit differences in micro- and macro-level properties across different social contexts (e.g., homophily in friendship networks, heterophily in organizational settings) in ways similar to human social networks. Chang et al. [13] explore how realistic social networks generated by LLMs compare to real ones, and examine demographic biases in social ties. They find differences in realistic-ness of

generated networks among three prompting methods, and find that LLMs overestimate political homophily compared to real networks.

Instead of explicitly prompting LLMs to generate networks, we prompt them to generate film screenplays, then parse and extract social networks from the resulting output following Agarwal et al. [2, 3]. We validate the quality of the extracted networks by confirming the computational Bechdel test performs comparably on them to human-written ones, then apply it alongside character role classification [43] and SNA measures — including homophily, triangle analysis and centrality — to characterize the position of female characters within the implicit networks that emerge from LLM-generated narratives.

2.3 Audits of LLM Biases

Algorithm auditing is a technique for systematically probing black-box systems through repeated interaction to understand their behavior and investigate potential biases without direct access to their internals [39]. Such methods have been commonly used to study biases by attributes like gender or political leaning, in domains including search engines [38, 46], social media [61], and targeted advertising [28]. We apply this auditing framework to large language models, systematically probing their screenplay generation to uncover potential gender biases in narrative representation.

Prior work has identified two major types of harms resulting from bias in LLM models, termed *allocative* and *representational* [6, 8, 14]. Earlier work on word embeddings found that they embed human-like biases [12] and entrench gender stereotypes [9], and proposed methodologies for modifying embeddings to address these bias sources. An ongoing line of research documented numerous forms of bias in LLMs, including gender bias [33, 59, 64], differential expressions of respect across genders, races and sexual orientation [50], biases in occupational associations [26], racial bias [21, 32], biases associating Muslims with violence [1], and stereotyping [51].

As LLMs are increasingly integrated into human studies research, work has found they can misportray and flatten the representation of marginalized demographic groups when used as proxies for human participants [60]. In a context related to movies, an audit of GPT's content moderation system found that many real and generated TV scripts are flagged as content violations, with flagging strongly linked to age rating, genre, and violent content. The study brought to attention the censorship implications when LLM moderation systematically shapes the production of cultural narratives [35], a question we take up from another angle in this work.

Complementing existing bias studies of LLM systems, we conduct a social network–based audit of three state-of-the-art LLMs, prompting each to generate 768 film screenplays and examining whether and how these models exhibit gender bias in their narrative outputs.

3 Dataset Design and Collection

To compare gender bias in LLM outputs relative to human outputs, we compile a dataset of human-written scripts and their anonymized synopses. Then, we use these synopses to generate a dataset of counterpart LLM-generated movie scripts from three models: GPT 5, Gemini 3 Pro, and Claude Sonnet 4.5.

3.1 Human-written Screenplays

We collected 1,116 human-written, English-language screenplays from The Internet Movie Script Database[1](IMSDb) and corresponding metadata — including short synopsis, cast information, genre, and rating — from The Movie Database[2] (TMDB), both publicly available sources used in prior computational and auditing work on screen media [3, 17, 19, 45]. Our selection was constrained to movie titles listed on IMSDb as of February 2025. Both sites are actively maintained; IMSDb requires administrator approval for script submissions, TMDB has a moderator

[1]https://imsdb.com/
[2]https://www.themoviedb.org/

system, and both include movies released as recently as 2025, the time of data collection. While IMSDb exclusively features screenplays that have been "made into movies," other site-level curation criteria are unknown.

From the initial 1,116 screenplays, we applied two filtering criteria for data completeness. First, we retained only screenplays with corresponding Bechdel test scores on the Bechdel Test Movie List[3], reducing the dataset to 825 movies. Second, we manually reviewed and excluded screenplays with synopses lacking meaningful plot details — synopses that did not describe story events, characters, setting or conflicts specific to the film, but instead described only its production context (e.g., director, cast, soundtrack), technical properties (e.g., lighting, film format), relationship to other works (e.g., sequel, remake, reboot), or reception and framing (e.g., critical reception, filmmaker's intention). For example, we excluded the synopses "A remake of the Wes Craven cannibal horror film" (*The Hills Have Eyes*, 2006) or "A short work by Si Fried, screened by the Ann Arbor Film Festival in 1972" (*Out of Site*, 1971). This yielded a final dataset of 783 screenplays with Bechdel metadata and meaningful plot synopses.

The finalized dataset spans more than a century of cinema (see Appendix Table 4 for the full list of films included in our dataset), from *Jane Eyre* (1910) to *Anora* (2024), with additional examples including *Pulp Fiction* (1994), *The Perks of Being a Wallflower* (2012), and *Deadpool* (2016). None of the included films are known to have been primarily generated by AI. Screenplays averaged 23,672 words in length (SD = 4,840). Movies were tagged with multiple genres, such as 'Thriller, Action' and 'Drama, Romance, Comedy,' and indicated as unrated, G, PG, PG-13, R, or NC-17.

3.2 Screenplay Generation

3.2.1 Model selection. We focus on three state-of-the-art LLMs, models underlying some of the most widely used chatbot systems at the time of writing [4]: GPT-5, Gemini 3 Pro, and Claude Sonnet 4.5. All experiments were conducted via the models' respective APIs in December 2025.

3.2.2 Synopsis anonymization. To generate full length screenplays, we prompted the LLMs using anonymized short synopses from TMDB for the 783 human-written screenplays in our dataset. Since we intended for the LLMs to generate the characters and their interactions, we cleaned all identifying information from the synopses – apart from including the synopsis and scene count in the prompt, we did not include any other information about the movie (number of characters, genre, year, title etc.). Synopses were anonymized by replacing character names with generic identifiers (e.g., PERSON 1), substituting gendered relational terms with neutral equivalents (e.g., "husband" or "wife" with "spouse"), and replacing gendered pronouns with gender-neutral ones. For example, *Confessions of a Dangerous Mind* (2002):

> **Original:** "Television producer by day, CIA assassin by night, *Chuck Barris* was recruited by the CIA at the height of *his* TV career and trained to become a covert operative…"
> **Anonymized:** "Television producer by day, CIA assassin by night, *PERSON 1* was recruited by the CIA at the height of *their* TV career and trained to become a covert operative…".

3.2.3 Prompt engineering. We conducted pilot experiments to develop an appropriate prompting approach for screenplay generation across GPT, Claude, and Gemini. Previous work that prompted GPT to generate screenplays used a prompt based on a single synopsis, but the resulting scripts were far shorter than human-written scripts [35]. We determined longer scripts necessary for the representational analysis we focus on in this paper, so length was the focus of our prompt engineering experiments.

We trialed two methods: (1) we first attempted to generate the entire movie screenplay with a single prompt that included the anonymized synopsis, similar to the work of [35], and (2) sequential prompting using consecutive paragraphs of the anonymized synopsis, with generated sub-scripts concatenated to form a full screenplay.

[3]https://bechdeltest.com/

Neither method generated scripts long enough to be comparable to human-written scripts, although the second was more successful. For both methods, we also experimented with providing a word count target for the output length. However, the LLMs did not respond accurately to this instruction, which aligns with their observed inability to recognize word counts [65].

Our final approach successfully generated long screenplays that were of similar or longer length than the human-written counterpart scripts. First, we prompted each model with the anonymized synopsis, specified the target number of scenes, and requested a numbered list of scene headings with brief descriptions. Second, for each scene heading, we issued a follow-up prompt to generate the full scene content. Finally, we concatenated the resulting scenes to form a complete screenplay. The exact prompts used in this two-step approach are shown in Appendices 5 and 6.

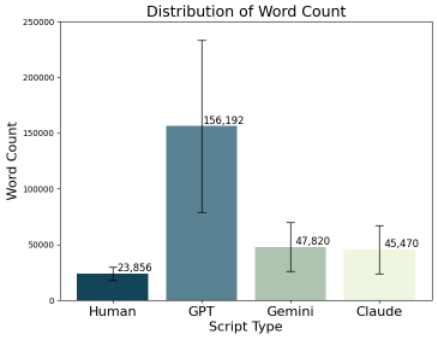

Fig. 1. Mean script length across the 4 script types; error bars indicate standard deviation.

3.2.4 Generated screenplay dataset. We generated 770 GPT scripts ($M = 156,192 words, SD = 77,030$), 771 Gemini scripts ($M = 47,820 words, SD = 21,725$), and 752 Claude scripts ($M = 45,470 words, SD = 21,605$).

During individual scene generation, models sometimes refused to generate scenes due to content moderation. Claude refused 36 scene generations, while GPT-5 refused twice, each with explicit content-filter errors. Gemini, by contrast, did not issue explicit refusals but instead generated 50 empty scenes. Models also refused or generated empty screenplays with Claude missing 31 screenplays, GPT missing 13 screenplays, and Gemini missing 12 screenplays. We proceed rather than regenerating to remain consistent with a one-time generation per script, and because prior work on refusal behaviors suggests regeneration would likely result in the same outcome.

All three models demonstrated high variation in output length; a given model generated a large range of script lengths. LLM-generated scripts were substantially longer than human-written scripts, with GPT producing scripts ~3.3 times longer than the other LLMs and ~6 times longer than human-written scripts (Figure 1).

Occasionally LLMs reproduced character names despite synopsis anonymization, which may be due to recognizing the source film from training data. Importantly, however, each LLM structures its generated narrative differently, as reflected in varying script lengths and social network structures. In a spot-check reading of 100 screenplays — 25 films and 4 script types — we did not find evidence that LLMs simply replicated human-written scripts. Our analysis focuses on these structural differences, rather than whether the original film was recognized.

4 Analytical Methods

We now describe our methods for (1) constructing character interaction networks from human- and LLM-generated screenplays, (2) implementing an automated Bechdel test, and (3) computing network measures to compare structural similarities and differences and to assess gender bias within these networks.

4.1 Character Network Construction

Characters were extracted and converted to nodes in our network representation. For both human-written and LLM-generated screenplays, we extracted character names based on formatting patterns: (a) names followed by a colon, (b) names written in all-uppercase, and (c) exclusion of scene directions such as "FADE IN" or "CUT TO."

Edges connect character nodes to indicate interactions, with edge weights representing the number of times a given pair of characters interacts. Following Agarwal et al. [3], we define a character interaction as two characters appearing consecutively within the same scene. For example:

```
JAMAL: Where did you get the gun?
SALIM: Bought it. Now, I'm going to have to throw this beauty in the sea.
LATIKA: You didn't need to kill him.
```

In this dialogue excerpt from *Slumdog Millionaire* (2008) [11], the interaction pairs are Jamal/Salim and Salim/Latika. Jamal and Latika are not considered to have interacted, as they did not speak consecutively. This consecutive-speaker approach provides a more accurate representation of character interaction than connecting all characters appearing in the same scene, in keeping with past work [3].

Using this graph construction method, 768 out of the 783 human-written scripts were able to be converted to character networks (the remaining 15 unconverted scripts lacked scene breaks and other necessary cues). We then generated social networks for the LLM counterparts of the 768 network-convertible human-written scripts. This was possible for all Gemini scripts; five Claude scripts did not contain dialogue, and one GPT script flagged that the prompt asked for too many scenes. This brought our totals to 768 human-written, 754 GPT, 756 Gemini, and 732 Claude character networks.

4.1.1 Determining Character Gender. In the network representation of human-written scripts, we ascribed gender to each character node by matching characters to cast members in TMDB and using the actors' listed genders (no available gender, female, male, or other). Manual review of the "other" category revealed inconsistencies: some nodes represented cisgender characters played by non-binary actors, while others were transgender characters played by cisgender actors. Given these inconsistencies and the binary gender framework of the Bechdel test, we reclassified all such nodes, along with 'no available gender' nodes, using a name-based approach, classifying character names rather than actor names because actors do not necessarily play characters that correspond to their personal gender identity.

We used the NomQuamGender [58] Python package to classify unmatched character names in the human-written screenplays and all character names in the LLM-generated ones. NomQuamGender uses a database of names with known gender distributions to calculate the probability of a name belonging to a man or woman based on historical usage patterns, with probability values near 0.5 indicate roughly equal usage across genders. The algorithm applies an uncertainty threshold in classifying names; we tuned NomQuamGender on all character names from TMDB, which determines the smallest uncertainty threshold that can gender-classify >85% of the character names at a starting threshold of 0.3. The algorithm was maximally able to classify 62% of the character names at this threshold, so the algorithm proceeded with an uncertainty threshold of 0.3 in our work. The remaining 38% of the dataset had unclassifiable character names that were role-based (e.g., Therapist, Barista) or gender-neutral (e.g., Ash, Casey).

We acknowledge that inferring gender from names is inherently limited, and risks reinforcing essentialist links between names and gender categories [18]. At best, this is only a proxy: names do not determine gender identity,

vary significantly across cultures and time periods, and can be gender-neutral or used across genders. However, for the purposes of Bechdel test analysis—which relies on perceived gender presentation in dialogue—a probabilistic name-based approach provides a consistent method applicable to both human-written and LLM-generated scripts.

Using this combined approach, we successfully assigned a gender to 70% of human-written characters. For LLM-generated scripts, gender classification rates were slightly higher: 75% for GPT, 78% for Gemini, and 69% for Claude. Character roles are determined in Section 4.1.2. The majority of unclassified characters are peripheral figures lacking explicit names or gender markers in the script — such as "Police Officer" or "Teacher" — and only 6.90% of major characters remained unclassified.

4.1.2 Determining Character Role. We implement the algorithm developed by Park et al. [44] to classify characters into two categories: MAJOR and NON-MAJOR (includes MINOR and PERIPHERAL). This classification is relevant for analyzing film scripts because characters in different roles exert distinct narrative influence. By categorizing characters by role prominence, we can assess whether gender representation varies across these hierarchical levels.

The algorithm uses degree centrality as its primary measure. A character's degree centrality is calculated by dividing its degree (the number of other characters it is connected to) by the maximum possible degree in the network (N-1, where N is the total number of characters). The classification process proceeds in three steps: (1) All characters are sorted in descending order by degree centrality. (2) For each adjacent pair of characters in this sorted list, we calculate the difference in their degree centrality values. (3) We identify the pair with the largest centrality difference; this difference defines the boundary between MAJOR and MINOR characters. The higher-centrality character in this pair is classified as the last MAJOR character, while the lower-centrality character becomes the first MINOR character. The last minor character has the lowest degree centrality that is still above the mean degree centrality of the network. (4) All nodes with values below the mean degree centrality are classified as peripheral. However, since we are primarily interested in gendered characters that have influence in the network, we merged the MINOR and PERIPHERAL roles into the singular NON-MAJOR role.

4.2 Automating the Bechdel Test

The Bechdel test assesses women's presence and dimensionality in literary or film work through three sequential criteria, each conditional on passing the previous one. Following Agarwal et al. [3], we operationalized each criterion:

Test 1: Two named *female characters...* If a movie's character network contained at least two *named* female-classified nodes, the movie passed Test 1. Otherwise, if the character network had fewer than two named female characters, the movie failed Test 1, and was ineligible to continue the test.

Test 2: ...who talk to each other... A movie passed Test 2 if its character network contained at least one edge between any two female character nodes. If the network had no edges between female characters, the movie failed Test 2, and was ineligible to continue.

Test 3: ...about something other than a man. To automatically determine whether female-female character conversations were about men, we used a machine learning approach based on SNA features. Following Agarwal et al. [3], who demonstrated that using the set of SNA features yielded higher F1 scores than alternative feature sets — including word unigrams, distribution of conversations over topics, linguistic features capturing mentions of "men" in dialogue, and features proposed by Garcia et al. [17] — we extracted 43 SNA features characterizing the placement of female characters in the character network (see Appendix 1). We then trained an RBF Support Vector Machine (SVM) model to predict Test 3 scores. The model was trained on 576 randomly selected films (75% of the 768 human-written films), with hyperparameters optimized via grid search ($C = 0.5$, $\gamma = 0.01$), and subsequently applied to the full dataset of scripts in our testing pipeline. The SVM had a prediction error of 0.265 and the following accuracy scores: $P = 0.65, R = 0.80, F1 = 0.72$.

4.2.1 Validation. We validated our Bechdel test method on human-written scripts using the Bechdel Test Movie List[4], a crowdsourced dataset of human evaluations used in the work of Agarwal et al. [3]; our validation also includes films added to the list over the past decade since that publication. Admin approval is required to rate a movie and at the time of data collection, the site had contributions within the same month. In our dataset of 768 movies, 95% of movies pass the first test, 67% pass the second, and 54% pass the third. Note that movies that pass Test 2 are highly likely to also pass Test 3. Our method achieved comparable performance to Agarwal et al. [3], with comparable precision, recall and F1 scores for each test (see Appendix Table 2 for precise numbers).

Since Test 3 is trained and validated on human-written screenplays, we further assessed robustness under domain shift to LLM-generated screenplays. We hand-annotated 300 LLM screenplays — 100 each from Gemini, GPT, and Claude — for pass/fail on the Bechdel test and compared this to our model's assessment. Our model achieves comparable precision, recall, and F1 scores for LLM scripts relative to human-written screenplays, indicating our model generalizes well to LLM-generated screenplays (see Appendix Table 3 for precise numbers).

4.3 Assessing Network Structure

To supplement Bechdel performance, we analyze character centrality, homophily, and triangle relationships as additional measures of gender bias in character interaction and positioning. Below we define and motivate these measures.

Centrality. The centrality of a given character node expresses its influence in the character network; a character with higher involvement with other characters is more central [62]. There are three main notions of centrality that each offer distinct perspectives: (1) *Degree centrality* is based on the idea that a more central character must be connected to more characters in the network, and is defined as a ratio of the number of others to which a given character connects compared to the maximum possible such value [62]. (2) *Closeness centrality* offers an alternate view: a character is more central the faster that node can reach other characters' nodes through the network, calculated by measuring the inverse average distance between a given character n_i and all other characters [62]. (3) Finally, *betweenness centrality* suggests that a character is more central if their node mediates interactions between non-adjacent characters; that is, that a more central character must lie on more shortest paths between two non-adjacent characters, or the average probability that a given character lies on the shortest path between two other non-adjacent characters [62].

Homophily. The Bechdel test only accounts for female-to-female interactions. To examine cross-gender interactions, we analyze the network's homophily — the pattern in which members of a social group are more likely to share interaction links with each other than with members of other groups [37, 40]. We examine homophily at the dyadic level, measuring the global tendency for characters to form same-gender versus opposite-gender pairs. This offers a different perspective on the positioning of female characters: high homophily indicates that characters primarily engage in gender-segregated interactions rather than integrating across gender lines within the broader cast.

Triangles. Beyond pairwise connections, we examine how characters cluster into the smallest community structures: triangles (characters connected in a triad). Triangles are fundamental units in social networks, forming the minimal building blocks of community structure [16, 20, 29]. While homophily reveals the overall tendency for same-gender connections, triangle analysis reveals the gender composition of these minimal community triads. A network could exhibit moderate homophily but predominantly contain male-majority triangles, indicating that while individual connections may be mixed, actual community structures remain gender-segregated. We examine how the distribution of all-male, all-female, and mixed-gender triangles varies across script types.

[4]https://www.bechdeltest.com/

5 Findings

We now describe the findings from evaluating the four screenplay types: human-written, GPT-5, Gemini 3 Pro, and Claude Sonnet 4.5. First, we found that the number of interactions in a network, measured as the sum of edge weights, correlates strongly with script type and script length; LLM scripts were significantly longer and had significantly more interactions. Additionally, our analyses heavily rely on edge-defined relationships; 41 of the 43 SNA features tracked in Test 3 and all following social network analyses are dependent on the edge relationships present in the network. We experimented with additional controls alongside the number of interactions, including total characters, average utterance length, and average number of interactions per character, and found that our trends hold. We therefore control for the number of interactions in all analyses that follow.

Since we aim to identify differences between the four script types on various measures that capture representational bias against women, we primarily use an OLS regression model. This model computes a line of best fit for the four script types and control variable(s) as independent variables relative to a particular measure (dependent variable), to identify correlational relationships. The resulting coefficients and p-values show the correlation (if any) between each script type and a particular bias measure. We run all regressions with human-written scripts as the reference; reported coefficients are relative to a human-written script reference point. We first report performance on our automated Bechdel test model, then examine SNA measures outlined in 4.3 to measure biases in their representation of women characters.

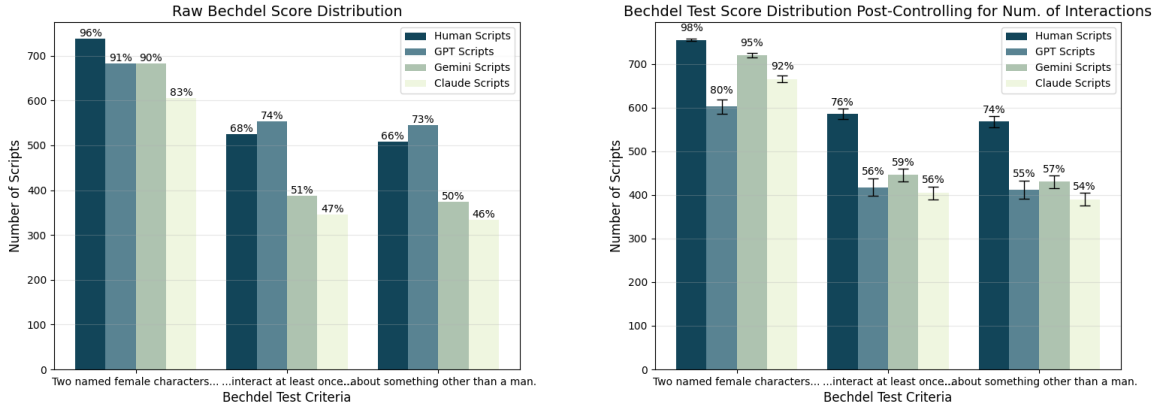

Fig. 2. Bechdel test performance before controlling for the number of interactions (left; raw pass rates) and after controlling for interactions (right; predicted pass rates); right error bars indicate predicted rates after controlling for interactions ± standard error.

5.1 Human-written screenplays score higher than LLMs on the Bechdel test

Based on the three tests described above, each script receives a Bechdel test score from 0 to 3. A script scores 0 if it fails Test 1 (lacks two female characters), 1 if it passes Test 1 but fails Test 2 (female characters do not interact), 2 if it passes Tests 1 and 2 but fails Test 3 (female characters only discuss men), and 3 if it passes all three tests. To pass the Bechdel test overall, a script must achieve a score of 3; scores of 0 to 2 indicate failure.

On a pass/fail basis. Figure 2 shows 74% of GPT scripts passed, 66% of human scripts passed, 50% of Gemini scripts passed, and 46% of Claude scripts passed. When we control for the number of interactions [5] in a logistic regression model that returns the odds of a specific script type passing the Bechdel test, we find that the number of interactions is correlated with pass/fail outcomes ($\beta = 0.001, p < 0.001$) and all LLM scripts are significantly less likely to pass the Bechdel test than human-written scripts.

Out of the LLM scripts, Gemini scripts are most likely to pass ($\beta = -0.77, p < 0.001$), followed by GPT scripts ($\beta = -0.88, p < 0.001$), and lastly, Claude scripts ($\beta = -0.91, p < 0.001$). We use the logistic regression model to predict the pass rate at the mean number of interactions, shown in Figure 2. This suggests that LLM-generated script performance on the Bechdel test may be inflated due to longer script length and thus more allowance to form interactions.

On an exact score basis. After controlling for the number of interactions [6], all LLM-generated scripts are associated with significantly lower Bechdel scores than human-written scripts. Relative to human-written scripts, Gemini-generated scripts exhibit the smallest decrease in Bechdel score ($\beta = -0.43, p < 0.001$), followed by GPT-generated scripts ($\beta = -0.52, p < 0.001$), while Claude-generated scripts show the largest decrease ($\beta = -0.57, p < 0.001$).

On a differential score basis. For each LLM-generated script, we compute the difference in Bechdel score relative to its human-written counterpart as $\text{Score}_{\text{human}} - \text{Score}_{\text{LLM}}$. In an OLS regression model with Claude-generated scripts as the reference category [7], the score differences for GPT-generated screenplays and Gemini-generated scripts are not statistically significantly different from those for Claude-generated screenplays.

Across evaluations, human-written screenplays consistently score higher than LLM-generated screenplays on the Bechdel test. In both the binary and continuous score evaluations, Gemini-generated screenplays exhibit the least reduction in performance relative to human-written screenplays, followed by GPT-generated screenplays, with Claude-generated screenplays showing the largest reduction. However, when modeling the Bechdel score gap between human-written and LLM-generated screenplays, differences among LLMs are not statistically significant. We next turn to network-based measures of gender bias.

5.2 Some LLM-generated scripts demonstrate better female character representation than human scripts on other network measures

The Bechdel test is one network evaluation measure of gender bias for film media. To investigate further, we examine several more network metrics, including the centrality imbalance between male and female characters, gender homophily, triadic relationships between male and female characters, and the proportion of edges involving female characters. In doing so, we find a more complex story: some LLM-generated scripts display better women's representation than human ones on other meaningful measures, described below.

5.2.1 All script types display gender bias in MAJOR character centrality. The Bechdel test treats all character interactions equally and does not account for differences in character role. Accordingly, we examined gender differences in centrality for MAJOR characters (see Section 4.1.2 for definition and construction). For each of the three notions of centrality previously defined, we conducted pairwise t-tests to assess differences in centrality between male and female characters.

All scripts display gender bias towards female characters, as seen in the ratio of female to male centrality falling under 1 in Figure 3. However, a statistical difference in centrality between male and female characters is only seen in GPT and human-written scripts, with GPT scripts demonstrating the most uneven representation of female

[5] `verdict ~ script type + number of interactions`
[6] `score ~ script type + number of interactions`
[7] `score difference ~ script type + number of interactions`

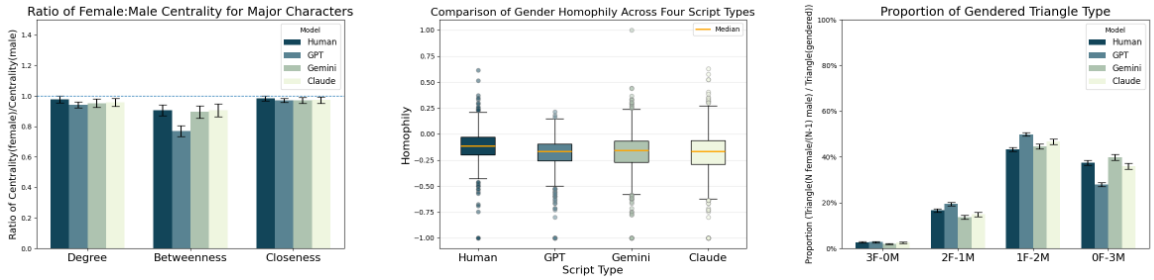

Fig. 3. Distribution of various network measures across 4 script types. Left: ratio of female centrality to male centrality for MAJOR characters, across all 4 script types and 3 centrality notions. Middle: box plot distribution of gender homophily. Right: proportion of triangles per gender composition type; error bars indicate mean ± standard error on bar graphs

MAJOR characters. GPT-generated screenplays exhibited a statistically significant difference between the mean degree centrality of male ($mean = 0.61$) and female ($mean = 0.58$) MAJOR characters, with male MAJOR characters more central ($t = 2.68, p = 0.007$). In contrast, human-written, Claude, and Gemini scripts show no significant difference in mean degree centrality between genders. Only GPT ($t = 5.72, p < 0.001$) and human-written ($t = 2.42, p = 0.016$) scripts demonstrated a higher betweenness centrality for male than female MAJOR characters. GPT scripts exhibited the larger betweenness centrality gap between male and female MAJOR characters, as shown by the smaller ratio of female-to-male betweenness centrality ($ratio_{GPT} = 0.78, ratio_{Human} = 0.91$). We did not find statistical differences between genders for closeness centrality across script types.

5.2.2 GPT screenplays exhibit the lowest homophily of female characters. All scripts displayed heterophily (the inverse of homophily) with mean scores from −0.11 to −0.19, as seen in Figure 3, reflecting more inter-gender interactions than expected. (Unlike movie scripts, which establish connection strictly through dialogue, real-world relationships are defined more robustly and tend to be more homophilous.) In an OLS regression model [8], GPT scripts displayed the highest heterophily ($\beta = -0.13, p < 0.001$), followed by Claude scripts ($\beta = -0.08, p < 0.001$), Gemini scripts ($\beta = -0.07, p < 0.001$), and lastly human-written scripts. This suggests that while human-written and Gemini scripts score higher than GPT and Claude scripts on the Bechdel test, this may be a result of better interaction between female characters than cross-gender interactions with male characters.

5.2.3 GPT screenplays have more female-inclusive triadic relationships than others. The second Bechdel test is a measure of female-female dyads in a network (two women that have a conversation). We examine a stronger bar: triads, network segments involving three characters. We examine four types of triadic relationships, classified by gender composition: three female and zero male (3F–0M), two female (2F–1M), two male (1F–2M), and three male characters (0F–3M) (Figure 3). Across all script types, the most common triangle is 1F–2M, followed by 0F–3M, 2F–1M, and finally 3F–0M. These patterns are consistent with previous observations in human-written film screenplays [23]. Aggregating triangles with at least one female character as *female-inclusive*, and triangles with only male characters as *female-exclusive*[9], GPT scripts exhibited a significantly higher proportion of female-inclusive triangles ($\beta = 0.09, p < 0.001$) compared to the other three types, which performed similarly.

5.2.4 GPT screenplays have the highest proportion of interactions involving female characters. We analyze representational bias in films by examining both the proportion of female characters and the proportion of interactions

[8] `homophily ~ script type + number of interactions`
[9] `proportion of female-inclusive triangles ~ script type + number of interactions`

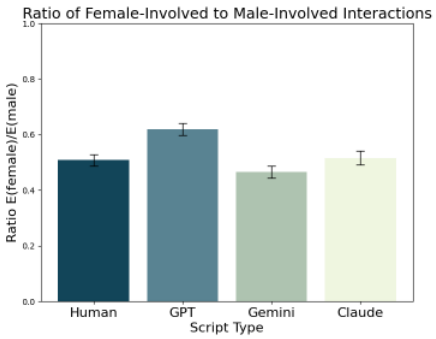

Fig. 4. Ratio of female-involved interactions (edges) to male-involved interactions; error bars indicate mean ± standard error.

involving them. While the Bechdel test incorporates the proportion of female characters, it does not distinguish characters by role type or account for the ratio of interactions in which female characters participate. We examine the proportion of female characters among MAJOR characters as defined in prior work; these characters play a central role in the narrative structure of films [43]. Using an OLS regression model controlling for number of character interactions [10], we find no statistically significant differences across script types: all contain roughly similar proportions of female MAJOR characters, ranging from 21–23% (n.s.). Examining the ratio of interactions involving female versus male characters with another regression [11], GPT scripts show the highest ratio ($\beta = 0.11, p < 0.001$), while human-written, Gemini ($\beta = -0.04, p = n.s.$), and Claude scripts ($\beta = 0.01, p = n.s.$) do not differ significantly, as shown in Figure 4.

6 Discussion

We find that human screenplays score higher than LLM-generated ones on the Bechdel test, a popular measure of gender bias in film. Our findings are more mixed on other aspects of their character networks; GPT-generated scripts display better representation of women than human ones on some measures, including all-female network triads and heterophily. We next discuss implications for future studies, and the broader risks of AI replacing human creative labor.

6.1 Expanding LLM Evaluation Strategies

The introduction of LLMs and their surge in popularity has led to a parallel explosion in studies auditing and evaluating LLM outputs for bias [1, 26, 32, 33, 35, 39, 50, 59]. Auditing AI systems is critical for raising awareness among users about potential harms, and seeking accountability from developers. In that spirit, in this work we sought to analyze representational fairness [6, 8, 14] in LLM depictions of women. Conducting such representational bias evaluations entails several challenges compared to studying questions of allocative fairness, first among them the unstructured nature of their outputs. The long-form text produced by LLMs is hard to quantitatively analyze, leading many bias auditors to instead force LLMs to give binary or numerical answers [22, 64], or to extract and analyze LLMs' internal vector representations of different concepts [15]. While such audits are useful, they necessarily remove some of the nuance of generative textual responses.

In this work, we take a complementary approach drawing from social network analysis (SNA), converting LLM outputs into network graphs that can be quantitatively analyzed using SNA measures. This has the added benefit

[10] `proportion of MAJOR female characters ~ script type + number of interactions`
[11] `ratio of female-involved interactions to male-involved interactions ~ script type + number of interactions`

of allowing us to compare LLM outputs with human texts, as such analyses are commonly done on human-written content in the domain of film [2, 3, 23, 27, 34, 43]. Our findings demonstrate the value of a network analysis approach to this topic, and support its broader use in LLM evaluations, including across other demographic groups and/or forms of media beyond screenplays. More generally, we encourage researchers in this domain to identify other ways to impose structure and leverage proven techniques to develop new forms of LLM evaluation.

6.2 Beyond Bechdel and Bias

A full and nuanced sense of self-identification and representation in media — the sense of seeing yourself in a character — cannot be captured by any quantitative measure. The same is true for positive or equitable representation of women (or any other group). In this work, we examine a variety of measures that capture different dimensions of visual representation — centrality within the narrative, tendency to form cross-gender relationships, proportion of characters of the same gender, etc. — to capture some aspect of these social concepts. But even if LLM scripts could match or outperform humans on our metrics, this would neither give people a sense of genuine visual representation, nor ensure fairness or equity in that representation.

Beyond representation, another crucial issue is at stake: the treatment of the human artists whose work is increasingly ingested, often without permission, to power systems that attempt to replace them. Alongside research on LLM biases in creative tasks, more attention is needed to the experiences of creatives on the front lines of this shift. Such questions complement, and reinforce, questions of representation in media; screenwriters should have a voice in determining both whether and how LLM screenwriting tools are used, and in assessing the quality of any resulting portrayals of the human experience. Participatory methods, such as those described by Tseng et al. [54] in their co-design of a journalist-controlled LLM are promising paths forward. By involving stakeholders whose livelihoods are most affected, like journalists and screenwriters, as co-designers, these approaches help preserve their decision-making power.

6.3 Limitations

As mentioned in previous sections, this work entails several limitations worth discussion. The first is that our work relies on name-to-gender inference to enable our large-scale analysis of gender bias in screenplays, which are dialogue-heavy and lack sufficient pronoun references. Automated name annotation itself introduces error, names are imperfect proxies for identity, and this method limits our work to a binary framing. Further work is sorely needed to address these issues, especially as characters beyond the binary are increasingly being represented in film.

We also collected scripts and Bechdel scores from crowdsourced datasets. Though these sites are moderated, their data can be highly variable. For example, 45% of the human-written scripts were not tagged with maturity ratings and each film is tagged with multiple genres. This limitation extends to LLM-generated scripts, which similarly lack genre/rating classifications, so we were unable to segment our analyses along these axes. Bechdel scores were also imperfect; when manually reviewing score disagreement between our model and the human-annotated baseline (from The Bechdel Movie List), we found several instances where the latter had inconsistently applied the test and/or users disagreed on the score. Additionally, we acknowledge that LLMs are non-deterministic. We believe our full dataset of over 700 screenplays is robust, but further work could be done on model consistency.

Finally, since we collected scripts from publicly available sources, there is a high likelihood that the scripts appear in the LLMs' training data despite our attempts at anonymization. However, given that the resulting scripts were substantially longer than the human-written versions, had structurally different social networks, and did not replicate human scripts during a spot check reading of 100 screenplays, we do not believe the generated scripts directly reproduced such training data. Future work could further investigate the extent to which the generated scripts represented new plot ideas and different character development.

7 Conclusion

LLMs have well-documented biases against various marginalized groups, including women, but their long-form outputs make these challenging to study quantitatively. Existing analyses have largely been restricted to studying internal vector representations or constraining LLM outputs to single-word or otherwise structured answers. In this work, we complement these approaches in a timely creative domain, applying social network analysis to analyze LLM-generated screenplays. We curate a dataset of 768 human-written scripts and create matching LLM-generated scripts by GPT, Gemini, and Claude based on the same film synopses. We create character networks for each script and analyze the network structures in two parts: (1) using the Bechdel test, a popular measure of gender bias in film, and (2) analyzing the representation of female characters using network measures informed by the literature, including centrality and homophily. We find that human scripts outperform generated ones on the Bechdel test, but that LLM-generated scripts display less bias against women in some of the subsequent tests. We find that across tests and script type, representational bias against women remains.

Based on these findings, we discuss implications for creative fields like filmmaking, and for researchers studying fairness and accountability. These findings highlight the value of LLM audits for evaluating potential biases, and the importance of developing methodological tools – such as the social network analyses performed here – for conducting them. Representational bias is particularly concerning as AI-generated text, images, and video increasingly shape our media environment. While human-produced media has long reflected structural biases, frameworks such as the Bechdel test have helped make these patterns visible and have contributed to gradual improvements in representation. Just as human writers' rooms can bring in consultants or screenwriters from diverse backgrounds, affected stakeholders should likewise be involved as co-designers in determining whether and how LLMs are used in creative production.

Statement on Generative AI Usage

The authors disclose that generative AI, beyond being used as an object of study, aided in coding figures in Python, translating pseudocode into executable Python code, and rephrasing already-written text for clarity.

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 298–306. doi:10.1145/3461702.3462624

[2] Apoorv Agarwal, Sriramkumar Balasubramanian, Jiehan Zheng, and Sarthak Dash. 2014. Parsing Screenplays for Extracting Social Networks from Movies. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, Anna Feldman, Anna Kazantseva, and Stan Szpakowicz (Eds.). Association for Computational Linguistics, Gothenburg, Sweden, 50–58. doi:10.3115/v1/W14-0907

[3] Apoorv Agarwal, Jiehan Zheng, Shruti Kamath, Sriramkumar Balasubramanian, and Shirin Ann Dey. 2015. Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Rada Mihalcea, Joyce Chai, and Anoop Sarkar (Eds.). Association for Computational Linguistics, Denver, Colorado, 830–840. doi:10.3115/v1/N15-1084

[4] Evan Bailyn. 2025. Top Generative AI Chatbots by Market Share – December 2025. https://firstpagesage.com/reports/top-generative-ai-chatbots/ Section: SEO Blog.

[5] David Bamman, Rachael Samberg, Richard Jean So, and Naitian Zhou. 2024. Measuring diversity in Hollywood through the large-scale computational analysis of film. *Proceedings of the National Academy of Sciences* 121, 46 (Nov. 2024), e2409770121. doi:10.1073/pnas.2409770121 Publisher: Proceedings of the National Academy of Sciences.

[6] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: From allocative to representational harms in machine learning. In *SIGCIS conference paper*.

[7] Alison Bechdel. 1985. *Dykes to Watch Out For*. Firebrand Books. https://dykestowatchoutfor.com/

[8] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. doi:10.18653/v1/2020.acl-main.485

[9] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[10] Joel Kim Booster. 2023. GLAAD Media Awards 2023: Fire Island's Joel Kim Stands Strong With WGA In Acceptance Speech: "Labor Issues Are Queer Issues". https://glaad.org/glaad-media-awards-2023-fire-islands-joel-kim-stands-strong-wga-acceptance-speech-labor-issues/.

[11] D. Boyle and L. Tandan. 2008. Slumdog Millionaire.

[12] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. doi:10.1126/science.aal4230

[13] Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. 2025. LLMs Generate Structurally Realistic Social Networks but Overestimate Political Homophily. *Proceedings of the International AAAI Conference on Web and Social Media* 19 (June 2025), 341–371. doi:10.1609/icwsm.v19i1.35820

[14] Kate Crawford. 2017. The Trouble with Bias. In *Keynote at NeurIPS*.

[15] Hannah Cyberey, Yangfeng Ji, and David Evans. 2025. Unsupervised Concept Vector Extraction for Bias Control in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 28333–28355. doi:10.18653/v1/2025.emnlp-main.1439

[16] Nurcan Durak, Ali Pinar, Tamara G. Kolda, and C. Seshadhri. 2012. Degree relations of triangles in real-world networks and graph models. In *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12)*. Association for Computing Machinery, New York, NY, USA, 1712–1716. doi:10.1145/2396761.2398503

[17] David Garcia, Ingmar Weber, and Venkata Garimella. 2014. Gender Asymmetries in Reality and Fiction: The Bechdel Test of Social Media. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 131–140. doi:10.1609/icwsm.v8i1.14522

[18] Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. 2024. Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic Attributes in NLP. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 323–337. doi:10.18653/v1/2024.gebnlp-1.20

[19] Philip John Gorinski and Mirella Lapata. 2015. Movie Script Summarization as Graph-based Scene Extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Rada Mihalcea, Joyce Chai, and Anoop Sarkar (Eds.). Association for Computational Linguistics, Denver, Colorado, 1066–1076. doi:10.3115/v1/N15-1113

[20] Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology* 78, 6 (1973), 1360–1380.

[21] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633, 8028 (Sept. 2024), 147–154. doi:10.1038/s41586-024-07856-5 Publisher: Nature Publishing Group.

[22] Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2025. Evaluating Bias in LLMs for Job-Resume Matching: Gender, Race, and Education. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, Weizhu Chen, Yi Yang, Mohammad Kachuee, and Xue-Yong Fu (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 672–683. doi:10.18653/v1/2025.naacl-industry.55

[23] Dima Kagan, Thomas Chesney, and Michael Fire. 2020. Using data science to understand the film industry's gender gap. *Palgrave Communications* 6, 1 (May 2020), 92. doi:10.1057/s41599-020-0436-1 Publisher: Palgrave.

[24] D. Kellner. 1995. *Media Culture: Cultural Studies, Identity and Politics Between the Modern and the Postmodern*. Routledge. https://books.google.com/books?id=GjbdsiZ0q10C

[25] Molly Kinder. 2024. Hollywood writers went on strike to protect their livelihoods from generative AI. Their remarkable victory matters for all workers. https://www.brookings.edu/articles/hollywood-writers-went-on-strike-to-protect-their-livelihoods-from-generative-ai-their-remarkable-victory-matters-for-all-workers/

[26] Hannah Rose Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021. Bias out-of-the-box: an empirical analysis of intersectional occupational biases in popular generative language models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, 2611–2624.

[27] Arjun M. Kumar, Jasmine Y. Q. Goh, Tiffany H. H. Tan, and Cynthia S. Q. Siew. 2022. Gender Stereotypes in Hollywood Movies and Their Evolution over Time: Insights from Network Analysis. *Big Data and Cognitive Computing* 6, 2 (June 2022), 50. doi:10.3390/bdcc6020050 Publisher: Multidisciplinary Digital Publishing Institute.

[28] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management science* 65, 7 (2019), 2966–2981.

[29] David Laniado, Yana Volkovich, Karolin Kappler, and Andreas Kaltenbrunner. 2016. Gender homophily in online dyadic and triadic relationships. *EPJ Data Science* 5, 1 (May 2016), 19. doi:10.1140/epjds/s13688-016-0080-6

[30] Peter A. Leavitt, Rebecca Covarrubias, Yvonne A. Perez, and Stephanie A. Fryberg. 2015. "Frozen in Time": The Impact of Native American Media Representations on Identity and Self-Understanding. *Journal of Social Issues* 71, 1 (2015), 39–53. doi:10.1111/josi.12095 _eprint: https://spssi.onlinelibrary.wiley.com/doi/pdf/10.1111/josi.12095.

[31] Benjamin Lee. 2024. Lionsgate partners with AI firm to train generative model on film and TV library. *The Guardian* (Sept. 2024). https://www.theguardian.com/film/2024/sep/18/lionsgate-ai

[32] Eric Justin Liu, Wonyoung So, Peko Hosoi, and Catherine D'Ignazio. 2024. Racial Steering by Large Language Models: A Prospective Audit of GPT-4 on Housing Recommendations. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3689904.3694709

[33] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J. Martin (Eds.). Association for Computational Linguistics, Virtual, 48–55. doi:10.18653/v1/2021.nuse-1.5

[34] Jinna Lv, Bin Wu, Lili Zhou, and Han Wang. 2018. StoryRoleNet: Social Network Construction of Role Relationship in Video. *IEEE Access* 6 (2018), 25958–25969. doi:10.1109/ACCESS.2018.2832087

[35] Yaaseen Mahomed, Charlie M. Crawford, Sanjana Gautam, Sorelle A. Friedler, and Danaë Metaxa. 2024. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show?. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 660–686. doi:10.1145/3630106.3658932

[36] Janice McCabe, Emily Fairchild, Liz Grauerholz, Bernice A. Pescosolido, and Daniel Tope. 2011. Gender in the Twentieth-Century Children's Books: Patterns of Disparity in Titles and Central Characters. *Gender and Society* 25, 2 (2011), 197–226. http://www.jstor.org/stable/23044136

[37] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27 (2001), 415–444. http://www.jstor.org/stable/2678628

[38] Danaë Metaxa, Joon Sung Park, James A. Landay, and Jeff Hancock. 2019. Search Media and Elections: A Longitudinal Investigation of Political Search Results. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 129:1–129:17. doi:10.1145/3359231

[39] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344. doi:10.1561/1100000083

[40] M. E. J. Newman. 2003. Mixing patterns in networks. *Physical Review E* 67, 2 (Feb. 2003), 026126. doi:10.1103/PhysRevE.67.026126

[41] Marios Papachristou and Yuan Yuan. 2025. Network formation and dynamics among multi-LLMs. *PNAS Nexus* 4, 12 (Dec. 2025), pgaf317. doi:10.1093/pnasnexus/pgaf317

[42] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3586183.3606763

[43] Seung-Bo Park, Yoo-Won Kim, Mohammed Nazim Uddin, and Geun-Sik Jo. 2009. Character-Net: Character Network Analysis from Video. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1. 305–308. doi:10.1109/WI-IAT.2009.54

[44] Seung-Bo Park, Kyeong-Jin Oh, and Geun-Sik Jo. 2012. Social network analysis in a movie using character-net. *Multimedia Tools Appl.* 59, 2 (2012), 601–627. doi:10.1007/s11042-011-0725-1

[45] Grace Proebsting, Oghenefejiro Isaacs Anigboro, Charlie M. Crawford, Danaé Metaxa, and Sorelle A. Friedler. 2025. Identity-related Speech Suppression in Generative AI Content Moderation. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '25)*. Association for Computing Machinery, New York, NY, USA, 185–217. doi:10.1145/3757887.3763010

[46] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–22. doi:10.1145/3274417

[47] Muniba Saleem and Srividya Ramasubramanian. 2019. Muslim Americans' Responses to Social Identity Threats: Effects of Media Representations and Experiences of Discrimination. *Media Psychology* 22, 3 (2019), 373–393. doi:10.1080/15213269.2017.1302345 _eprint: https://doi.org/10.1080/15213269.2017.1302345.

[48] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation Frames of Power and Agency in Modern Films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 2329–2334. doi:10.18653/v1/D17-1247

[49] Akrati Saxena, George Fletcher, and Mykola Pechenizkiy. 2024. FairSNA: Algorithmic Fairness in Social Network Analysis. *Comput. Surveys* (April 2024). doi:10.1145/3653711 Publisher: ACMPUB27New York, NY.

[50] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3407–3412. doi:10.18653/v1/D19-1339

[51] Zara Siddique, Liam Turner, and Luis Espinosa-Anke. 2024. Who is better at math, Jenny or Jingzhen? Uncovering Stereotypes in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 18601–18619. doi:10.18653/v1/2024.emnlp-main.1035

[52] Dr Stacy L Smith, Dr Katherine Pieper, and Sam Wheeler. 2023. Inequality in 1,600 popular films: Examining Portrayals of Gender, Race/Ethnicity, LGBTQ+ & Disability from 2007 to 2022. (Aug. 2023). https://assets.uscannenberg.org/docs/aii-inequality-in-1600-popular-films-20230811.pdf

[53] Jessica Toonkel. 2025. Exclusive | OpenAI Backs AI-Made Animated Feature Film. https://www.wsj.com/tech/ai/openai-backs-ai-made-animated-feature-film-389f70b0

[54] Emily Tseng, Meg Young, Marianne Aubin Le Quéré, Aimee Rinehart, and Harini Suresh. 2025. "Ownership, Not Just Happy Talk": Co-Designing a Participatory Large Language Model for Journalism. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 3119–3130. doi:10.1145/3715275.3732198

[55] Riva Tukachinsky, Dana Mastro, and Moran Yarchi. 2015. Documenting Portrayals of Race/Ethnicity on Primetime Television over a 20-Year Span and Their Association with National-Level Racial/Ethnic Attitudes. *Communication Faculty Articles and Research* (Jan. 2015). doi:10.1111/josi.12094

[56] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron A. Marlow. 2011. The Anatomy of the Facebook Social Graph. (2011).

[57] Johann Valentowitsch. 2023. Hollywood caught in two worlds? The impact of the Bechdel test on the international box office performance of cinematic films. *Marketing Letters* 34, 2 (2023), 293–308. doi:10.1007/s11002-022-09652-5

[58] Ian Van Buskirk, Aaron Clauset, and Daniel B Larremore. 2023. An Open-Source Cultural Consensus Approach to Name-Based Gender Classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 866–877. https://github.com/ianvanbuskirk/nbgc.

[59] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3730–3748. doi:10.18653/v1/2023.findings-emnlp.243

[60] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence* 7, 3 (March 2025), 400–411. doi:10.1038/s42256-025-00986-z Publisher: Nature Publishing Group.

[61] Stephanie Wang, Shengchun Huang, Alvin Zhou, and Danaë Metaxa. 2024. Lower Quantity, Higher Quality: Auditing News Content and User Perceptions on Twitter/X Algorithmic versus Chronological Timelines. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (Nov. 2024), 507:1–507:25. doi:10.1145/3687046

[62] Stanley Wasserman and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

[63] Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. 2007. RoleNet: treat a movie as a small society. In *Proceedings of the international workshop on Workshop on multimedia information retrieval (MIR '07)*. Association for Computing Machinery, New York, NY, USA, 51–60. doi:10.1145/1290082.1290092

[64] Kyra Wilson and Aylin Caliskan. 2024. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 1578–1590. doi:10.1609/aies.v7i1.31748

[65] Nan Xu and Xuezhe Ma. 2025. LLM The Genius Paradox: A Linguistic and Math Expert's Struggle with Simple Word-based Counting Problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 3344–3370. doi:10.18653/v1/2025.naacl-long.172

[66] Yulin Yu, Yucong Hao, and Paramveer Dhillon. 2022. Unpacking Gender Stereotypes in Film Dialogue. In *Social Informatics*, Frank Hopfgartner, Kokil Jaidka, Philipp Mayr, Joemon Jose, and Jan Breitsohl (Eds.). Springer International Publishing, Cham, 398–405. doi:10.1007/978-3-031-19097-1_26

A Screenplay generation prompts

Two prompts used to generate screenplays using LLMs in a two-step process.

You are a helpful AI assistant. I will provide a movie's synopsis where all character names are redacted as PERSON IDs (e.g., PERSON 1, PERSON 33) and the number of scenes in the movie.
Your tasks:

(1) Assign each PERSON ID a unique, consistent character name based on the roles and relationships implied in the synopsis.
(2) Generate a numbered list of scenes for the movie based on the synopsis. Each scene must have:
   - "number": integer
   - "description": one concise, single-sentence summary of events, including the names of all characters present in that scene
   - "characters": list of character names present in the scene

Movie Synopsis: {synopsis}
Number of Scenes: {num_scenes}

Fig. 5. Prompt used to generate structured movie scenes from anonymized synopses.

You are a helpful AI assistant. I will provide a movie's synopsis and a summary of a scene from the movie. The movie's synopsis have character names redacted as PERSON IDs (e.g., PERSON 1, PERSON 33). The scene description provides the real character names. Use the real character names in the screenplay. Based on the movie synopsis and scene description, please write a detailed screenplay for the scene.
Requirements:

- Only character names should be in ALL CAPS before their dialogue.
- Do NOT use ALL CAPS for scene descriptions, stage directions, or any other text.
- Use professional screenplay formatting.

Movie Synopsis: {synopsis}
Scene Description: SCENE {scene.number}: {scene.description}

Fig. 6. Prompt used to generate structured screenplays from a given movie scene and anonymized synopsis.

B SNA features used in Bechdel test model

We record the list of 43 SNA features used to predict whether a script passed the third part of the Bechdel test, in keeping with the method used by Agarwal et al. [3].

Table 1. List of 43 SNA features used to determine whether movie script passed Test 3 of the Bechdel test

#	Feature	Type
1–4	Degree centrality (Min, Max, Mean, SD)	Normalized
5–8	Closeness centrality (Min, Max, Mean, SD)	Normalized
9–12	Betweenness centrality (Min, Max, Mean, SD)	Normalized
13–16	Number of male characters a female character is connected to (Min, Max, Mean, SD)	Absolute
17–20	Number of female characters a female character is connected to (Min, Max, Mean, SD)	Absolute
21–24	Number of male characters in common between 2 female characters (Min, Max, Mean, SD)	Absolute
25–28	Number of female characters in common between 2 female characters (Min, Max, Mean, SD)	Absolute
29	Ratio of number of female characters to male characters	Normalized
30	Ratio of number of female characters to total characters	Normalized
31	Percentage of female characters that form a 3-clique with another male character and female character	Normalized
32–34	Percentage of female characters in the in top 5 main characters (based on each of the 3 centrality measures)	Normalized
35–37	Boolean recording whether the main character is female (based on each of the 3 centrality measures)	Normalized
38–40	Boolean recording whether any female character connects another female character to the main male character (based on each of the 3 centrality measures)	Normalized
41–43	Percentage of female characters that connect the main male character to another female character (based on each of the 3 centrality measures	Normalized

C Bechdel test model validation

Table 2. Precision, recall, and F1 scores for our Bechdel test model and Agarwal et al. [3] on human-written screenplays, evaluated against human annotations from the Bechdel Test Movie List as ground truth. Our model performs comparably to Agarwal et al. [3].

Bechdel Test Criteria	**Our Model**			**Agarwal et al.**		
	P	R	F1	P	R	F1
Pass Test 1	0.96	0.97	**0.97**	0.97	0.96	0.96
Pass Test 2	0.77	0.78	**0.78**	0.67	0.90	0.77
Pass Test 3	0.65	0.80	0.72	0.83	0.66	**0.73**

Table 3. Precision, recall, and F1 scores for Test 3 (i.e., overall pass/fail) of our Bechdel test model on human-written and LLM-generated screenplays, using human annotations from the Bechdel Test Movie List for human-written screenplays and our own hand-annotations of 100 screenplays per model for LLM-generated screenplays as ground truth. Our model performs comparably across both screenplay types.

Script Type	**Our Model**		
	P	R	F1
Human	0.65	0.80	0.72
GPT	0.68	0.96	**0.80**
Gemini	0.56	**0.97**	0.71
Claude	**0.69**	0.92	0.79

D List of films

Table 4. All 783 films in the dataset before network conversion, listed alphabetically.

Film (year)			
9 (2009)	*12* (2003)	*42* (2019)	*2012* (2009)
12 Years a Slave (2013)	*127 Hours* (2010)	*17 Again* (2009)	*20th Century* (2016)
28 Days Later (2002)	*30 Minutes or Less* (2001)	*44 Inch Chest* (2009)	*48 Hrs.* (1982)
8MM (1999)	*A Few Good Men* (1992)	*A Most Violent Year* (2014)	*A Nightmare on Elm Street* (2010)
A Perfect World (1993)	*A Quiet Place* (2016)	*A Real Pain* (2024)	*A Scanner Darkly* (2006)
A Serious Man (2009)	*A Walk to Remember* (2002)	*Ad Astra* (2016)	*Adaptation* (2022)
After.Life (2009)	*Agnes of God* (1985)	*Air* (2004)	*Air Force One* (2003)
Airplane (1980)	*Aladdin* (1992)	*Ali* (2001)	*Alien* (2025)
Alien 3 (1992)	*Alien Nation* (1988)	*Alien v. Predator* (2004)	*Aliens* (2014)
All About Eve (1950)	*All About Steve* (2009)	*All of Us Strangers* (2023)	*Almost Famous* (2000)
Amadeus (1984)	*Amelia* (2003)	*American Beauty* (1945)	*American Gangster* (2007)
American Graffiti (1973)	*American History X* (1998)	*American Hustle* (2013)	*American Outlaws* (2023)
American Pie (1999)	*American Sniper* (2014)	*American Werewolf in London* (1981)	*Amour* (2023)
An Education (2009)	*Analyze This* (1999)	*Anastasia* (1997)	*Annie Hall* (1977)
Anora (2024)	*Antitrust* (2001)	*Antz* (1998)	*Apocalypse Now* (1979)
Apt Pupil (1998)	*Arbitrage* (2012)	*Arctic Blue* (1993)	*Argo* (2017)
Armageddon (1998)	*Army of Darkness* (1992)	*Arsenic and Old Lace* (1969)	*As Good As It Gets* (1997)
Assassins (1995)	*Asteroid City* (2023)	*Autumn in New York* (2000)	*Avatar* (2009)
Babel (2019)	*Bachelor Party* (2009)	*Backdraft* (1991)	*Bad Boys* (2014)
Bad Day at Black Rock (1955)	*Bad Lieutenant* (1992)	*Bad Santa* (2003)	*Bad Teacher* (2011)
Barbie (2023)	*Barry Lyndon* (1975)	*Barton Fink* (1991)	*Basic* (2020)
Basic Instinct (1992)	*Basquiat* (1996)	*Batman* (1943)	*Bean* (2017)
Beasts of No Nation (2015)	*Beasts of the Southern Wild* (2012)	*Beauty and the Beast* (1997)	*Beginners* (2018)
Being John Malkovich (1999)	*Being the Ricardos* (2021)	*Being There* (2017)	*Belle* (2019)
Beloved (1976)	*Big* (1988)	*Big Eyes* (2014)	*Big Fish* (2004)
Black Rain (2021)	*Blackberry* (2023)	*BlacKkKlansman* (2018)	*Blade* (1973)
Blade II (2002)	*Blade Runner* (1982)	*Blood Simple* (1985)	*Blue Valentine* (2010)
Blue Velvet (1986)	*Body Heat* (1981)	*Body of Evidence* (1988)	*Bones* (2016)
Bonfire of the Vanities (1990)	*Bonnie and Clyde* (1967)	*Boogie Nights* (1997)	*Bookworm* (2024)
Bottle Rocket (2025)	*Bound* (2017)	*Braveheart* (1925)	*Break* (1986)
Brick (2006)	*Bridesmaids* (1989)	*Broadcast News* (1987)	*Broken Arrow* (2022)
Broken Embraces (2009)	*Bruce Almighty* (2003)	*Buffy the Vampire Slayer* (1992)	*Bull Durham* (1988)
Buried (2024)	*Burn After Reading* (2008)	*Cable Guy* (1996)	*Capote* (2005)
Carrie (1952)	*Cars 2* (2011)	*Case 39* (2009)	*Casino* (2011)
Cast Away (2017)	*Catch Me If You Can* (1998)	*Cecil B. Demented* (2000)	*Cedar Rapids* (2011)
Cellular (2020)	*Changeling* (2019)	*Chaos* (2021)	*Charade* (1963)
Chasing Amy (1997)	*Cherry Falls* (2000)	*Chinatown* (1974)	*Boogie Nights* (1997)
Cinema Paradiso (1998)	*Citizen Kane* (1941)	*Clash of the Titans* (1981)	*Clerks* (1994)
Cliffhanger (2026)	*Clueless* (1995)	*Cobb* (1994)	*Coco* (2000)
Cold Mountain (2003)	*Colombiana* (2011)	*Color of Night* (1994)	*Confessions of a Dangerous Mind* (2002)
Constantine (2005)	*Copycat* (2024)	*Coraline* (2009)	*Coriolanus* (1997)
Corpse Bride (2005)	*Cradle 2 the Grave* (2003)	*Crank* (2006)	*Crazy, Stupid, Love* (2011)
Creation (2020)	*Crouching Tiger, Hidden Dragon* (2000)	*Cruel Intentions* (2016)	*Crying Game* (1992)
Cube (2019)	*Custody* (2005)	*Dallas Buyers Club* (2013)	*Dances with Wolves* (1990)
Dark City (1990)	*Dark Star* (1974)	*Darkman* (1992)	*Date Night* (2024)
Dawn of the Dead (2004)	*Day of the Dead* (1985)	*Days of Heaven* (1978)	*Dead Poets Society* (1989)
Deadpool (2016)	*Dear White People* (2014)	*Death at a Funeral* (2007)	*Death to Smoochy* (2002)
Deception (2011)	*Defiance* (1993)	*Despicable Me* (2013)	*Devil in a Blue Dress* (1995)
Die Hard (1988)	*Die Hard 2* (1990)	*Diner* (1983)	*Disturbia* (2007)
Django Unchained (2012)	*Do the Right Thing* (1989)	*Dog Day Afternoon* (1975)	*Dogma* (1999)
Donnie Brasco (1997)	*Double Indemnity* (1973)	*Drag Me to Hell* (2009)	*Dragonslayer* (2011)
Drive (2022)	*Drop Dead Gorgeous* (2010)	*Duck Soup* (1942)	*Dumb and Dumber* (1994)
Dune (2021)	*E.T.* (1982)	*Eagle Eye* (2008)	*Eastern Promises* (2007)
Easy A (2010)	*Ed Wood* (1994)	*Edward Scissorhands* (1990)	*Election* (1999)
Elemental (2023)	*Evlis* (2005)	*Enemy of the State* (1998)	*Enough* (2021)
Erin Brockovich (2000)	*Escape From L.A.* (1996)	*Escape From New York* (1981)	*Eternal Sunshine of the Spotless Mind* (2004)
Even Cowgirls Get the Blues (1994)	*Event Horizon* (2017)	*Evil Dead* (2013)	*Ex Machina* (2015)
Excalibur (1981)	*Extract* (2009)	*Fair Game* (2017)	*Fantastic Beasts and Where to Find Them* (2016)
Fantastic Four (2015)	*Fargo* (1952)	*Fast Times at Ridgemont High* (1982)	*Fatal Instinct* (2014)

continued on next page

Table 4 – *continued*

Film (year)			
Fear and Loathing in Las Vegas (1998)	*Feast* (2023)	*Ferrari* (2023)	*Field of Dreams* (1989)
Fight Club (1999)	*Final Destination* (2000)	*Final Destination 2* (2003)	*Finding Nemo* (2003)
Five Easy Pieces (1995)	*Flash Gordon* (1980)	*Fletch* (1985)	*Flight* (1997)
Forrest Gump (1994)	*Four Rooms* (1995)	*Foxcatcher* (2014)	*Fracture* (2020)
Fred Claus (2007)	*Freddy vs. Jason* (2003)	*Friday the 13th* (2016)	*Fright Night* (2011)
From Dusk Till Dawn (1996)	*From Here to Eternity* (2014)	*Frozen River* (1929)	*Fruitvale Station* (2013)
Funny People (1981)	*G.I. Jane* (1951)	*Gamer* (2024)	*Gandhi* (1982)
Gangs of New York (1938)	*Garden State* (2024)	*Gattaca* (1997)	*Get Carter* (1971)
Get Low (2010)	*Get on Up* (2014)	*Get Out* (2017)	*Ghost Ship* (2002)
Ghost World (2001)	*Ghostbusters* (2016)	*Gladiator* (1992)	*Godzilla* (1998)
Good Will Hunting (1997)	*Gothika* (2003)	*Grabbers* (2012)	*Gran Torino* (2008)
Grand Hotel (2018)	*Gravity* (2014)	*Gremlins* (1984)	*Gremlins 2* (1990)
Groundhog Day (1993)	*Hackers* (1995)	*Hall Pass* (2011)	*Hancock* (1991)
Hanna (2011)	*Hannah and Her Sisters* (1986)	*Hannibal* (1972)	*Happy Feet* (1991)
Hard Rain (1998)	*Harold and Kumar Go to White Castle* (2004)	*Heathers* (1988)	*Heavenly Creatures* (1994)
Heavy Metal (1979)	*Hellboy* (2019)	*Hellraiser* (2022)	*Her* (Not Avail.)
Hesher (2010)	*High Fidelity* (2000)	*Highlander* (1986)	*His Girl Friday* (1940)
Hitchcock (2012)	*Hollow Man* (2000)	*Horrible Bosses* (2011)	*Hot Tub Time Machine* (2010)
Hotel Rwanda (2004)	*House of 1000 Corpses* (2003)	*How to Train Your Dragon* (2010)	*Hudson Hawk* (1991)
Human Nature (2025)	*I Am Number Four* (2011)	*I am Sam* (2001)	*I Love You Phillip Morris* (2010)
I Spit on Your Grave (2010)	*I Still Know What You Did Last Summer* (1998)	*I, Robot* (2004)	*In the Bedroom* (2024)
In the Loop (2009)	*Inception* (2010)	*Boogie Nights* (1997)	*Indiana Jones and the Last Crusade* (1989)
Indiana Jones and the Temple of Doom (1984)	*Inglourious Bastards* (2009)	*Initiation* (2011)	*Insidious* (2008)
Insomnia (2002)	*Interstellar* (2014)	*Interview with the Vampire* (1994)	*Into the Wild* (2007)
Into the Woods (2014)	*Intolerable Cruelty* (2003)	*Invictus* (2009)	*It* (2017)
It Happened One Night (1934)	*Jackie Brown* (1997)	*Jane Eyre* (1910)	*Jason X* (2001)
Jaws 2 (1978)	*Jay and Silent Bob Strike Back* (2001)	*Jerry Maguire* (1996)	*JFK* (2013)
John Wick (2014)	*Jojo Rabbit* (2019)	*Judge Dredd* (1995)	*Juno* (2007)
Jurassic Park (1993)	*Jurassic Park III* (2001)	*Kids* (2019)	*Kill Your Darlings* (2024)
Killers of the Flower Moon (2023)	*King Kong* (2012)	*Klute* (1971)	*Kill Bill* (2003)
King of Comedy (1982)	*Klute* (1971)	*Knocked Up* (2007)	*Kung Fu Panda* (2008)
La La Land (2016)	*Lake Placid* (1999)	*Last Chance Harvey* (2008)	*Last Tango in Paris* (1972)
Law Abiding Citizens (2021)	*Leaving Las Vegas* (1995)	*Legally Blonde* (2003)	*Legend* (2015)
Legion (2024)	*Les Miserables* (2012)	*Liar Liar* (1997)	*Life* (1999)
Life As A House (2001)	*Life of Pi* (2012)	*Limitless* (2021)	*Lincoln* (2006)
Little Nicky (2000)	*Lock, Stock and Two Smoking Barrels* (1998)	*Logan* (2013)	*Looper* (2012)
Lord of War (2005)	*Lost Highway* (1997)	*Lost Horizon* (1937)	*Lost in Space* (1998)
Lost in Translation (2003)	*Machete* (2010)	*Made* (2010)	*Magnolia* (1999)
Major League (1989)	*Malcolm X* (1992)	*Malignant* (2021)	*Man in the Iron Mask* (1998)
Man on the Moon (1999)	*Manhattan Murder Mystery* (1993)	*Manhunter* (1986)	*Margaret* (2011)
Margin Call (2011)	*Margot at the Wedding* (2007)	*Martha Marcy May Marlene* (2011)	*Marty* (2018)
Mary Poppins (1964)	*Max Payne* (2008)	*May December* (2023)	*Mean Streets* (1973)
Meet Joe Black (1998)	*Meet John Doe* (1941)	*Megamind* (2010)	*Memento* (2000)
Memory (2023)	*Men in Black* (1997)	*Metro* (2019)	*Miami Vice* (2006)
Midnight Cowboy (1969)	*Midnight Express* (1978)	*Midnight in Paris* (2011)	*Mighty Joe Young* (1998)
Milk (2009)	*Minority Report* (2002)	*Mirrors* (2008)	*Misery* (1990)
Mission Impossible (1996)	*Mission to Mars* (2000)	*Moneyball* (2011)	*Monkeybone* (2001)
Monte Carlo (2011)	*Moon* (2009)	*Moonrise Kingdom* (2012)	*Moonstruck* (1987)
Mrs. Brown (1997)	*Mud* (2013)	*Mulan* (1998)	*Music of the Heart* (1999)
My Girl (1991)	*My Week with Marilyn* (2011)	*Mystery Men* (1999)	*Nashville* (1975)
Natural Born Killers (1994)	*Never Been Kissed* (1999)	*New York Minute* (2003)	*Newsies* (1992)
Next (2007)	*Next Friday* (2000)	*Nick of Time* (1995)	*Nightbreed* (1990)
Nine (2009)	*Ninja Assassin* (2009)	*No Country for Old Men* (2007)	*No Hard Feelings* (2023)
Notting Hill (1999)	*Oblivion* (2013)	*Ocean's Twelve* (2004)	*Office Space* (1999)
Only God Forgives (2013)	*Onward* (2020)	*Oppenheimer* (2023)	*Ordinary People* (1980)
Origin (2023)	*Orphan* (2009)	*Pandorum* (2009)	*Panic Room* (2002)
ParaNorman (2012)	*Pariah* (2011)	*Passengers* (2016)	*Paul* (2011)
Pearl Harbor (2001)	*Peeping Tom* (1960)	*Peggy Sue Got Married* (1986)	*Pet Sematary* (1989)
Philadelphia (1993)	*Phone Booth* (2003)	*Pi* (1998)	*Pineapple Express* (2008)
Pirates of the Caribbean (2003)	*Pitch Black* (2000)	*Platoon* (1986)	*Point Break* (1991)
Poor Things (2023)	*Precious* (2009)	*Predator* (1987)	*Pride and Prejudice* (2005)
Priest (2011)	*Priscilla* (2023)	*Prom Night* (1980)	*Prometheus* (2002)
Psycho (1960)	*Public Enemies* (2009)	*Pulp Fiction* (1994)	*Purple Rain* (1984)
Queen of the Damned (2002)	*Quiz Lady* (2023)	*Rachel Getting Married* (2008)	*Raging Bull* (1980)

continued on next page

Table 4 – *continued*

Film (year)			
Raising Arizona (1987)	*Real Genius* (1985)	*Rear Window* (1954)	*Rebel Without A Cause* (1955)
Red Planet (2000)	*Red Riding Hood* (2011)	*Remember Me* (1981)	*Repo Man* (1984)
Resident Evil (2002)	*Revolutionary Road* (2008)	*Ringu* (1998)	*Rise of the Guardians* (2012)
Rise of the Planet of the Apes (2011)	*RocknRolla* (2008)	*Rocky* (1976)	*Ronin* (1998)
Room (2016)	*Runaway Bride* (1999)	*Rush* (2013)	*Rush Hour* (1998)
Rush Hour 2 (2001)	*Saturday Night* (2024)	*Save the Last Dance* (2001)	*Saving Mr. Banks* (2013)
Saving Private Ryan (1998)	*Saw* (2003)	*Saw IV* (2007)	*Scream 2* (1997)
Scream 3 (2000)	*Scream 4* (2011)	*Se7en* (1995)	*Sense and Sensibility* (1995)
Serenity (2005)	*Serial Mom* (1994)	*Sex and the City* (2008)	*Shakespeare in Love* (1998)
Shame (2011)	*Sherlock Holmes* (2010)	*Shrek* (2001)	*Shrek the Third* (2007)
Sicario (2015)	*Sideways* (2004)	*Signs* (2002)	*Silence of the Lambs* (1991)
Silver Linings Playbook (2012)	*Sing Sing* (2024)	*Single White Female* (2000)	*Sister Act* (1992)
Six Degrees of Separation (1993)	*Sideways* (2004)	*Skyfall* (2012)	*Sleepless in Seattle* (1993)
Sleepy Hollow (1999)	*Slither* (2006)	*Slumdog Millionaire* (2008)	*Smashed* (2012)
Snatch (2001)	*Snow White and the Hunstman* (2012)	*So I Married an Axe Murderer* (1993)	*Solaris* (2002)
Source Code (2011)	*South Park* (1999)	*Spanglish* (2004)	*Spartan* (2004)
Speed Racer (2008)	*Sphere* (1998)	*Star Trek* (2009)	*Stepmom* (1998)
Straight Outta Compton (2015)	*Strange Days* (1995)	*Strangers on a Train* (1951)	*Sugar* (2009)
Sunset Blvd. (1950)	*Sunshine Cleaning* (2009)	*Superbad* (2007)	*Supergirl* (1984)
Surrogates (2009)	*Sweet Smell of Success* (1957)	*Swingers* (1996)	*Swordfish* (2001)
Synecdoche, New York (2008)	*Take Shelter* (2011)	*Taking Lives* (2004)	*Tamara Drewe* (2010)
Ted (2012)	*Tenet* (2020)	*Terminator* (1984)	*Terminator Salvation* (2009)
The Addams Family (1991)	*The Adjustment Bureau* (2011)	*The American* (2010)	*The American President* (1995)
The Apartment (1991)	*The Assignment* (1997)	*The Addams Family* (1991)	*The Avengers* (1998)
The Beekeeper (2024)	*The Best Exotic Marigold Hotel* (2012)	*The Big Lebowski* (1998)	*The Big Sick* (2017)
The Birds (1963)	*The Blind Side* (2009)	*The Bling Ring* (2013)	*The Boondock Saints* (2000)
The Bourne Identity (2002)	*The Bourne Ultimatum* (2007)	*The Box* (2009)	*The Boxtrolls* (2014)
The Breakfast Club (1985)	*The Brothers Bloom* (2008)	*The Butterfly Effect* (2004)	*The Cell* (2000)
The Cider House Rules (1999)	*The Color Purple* (2023)	*The Croods* (2013)	*The Crow* (1994)
The Curious Case of Benjamin Button (2008)	*The Danish Girl* (2016)	*The Dark Knight Rises* (2012)	*The Day the Earth Stood Still* (1951)
The Debt (2011)	*The Deer Hunter* (1978)	*The Departed* (2006)	*The Descendants* (2011)
The Devil Wears Prada (2006)	*The Doors* (1991)	*The English Patient* (1997)	*The Family Man* (2000)
The Fault in Our Stars (2014)	*The Fifth Element* (1997)	*The Fighter* (2010)	*The Flintstones* (1994)
The French Connection (1971)	*The Fugitive* (1993)	*The Game* (Not Avail.)	*The Girl with the Dragon Tattoo* (2011)
The Good Girl (2002)	*The Graduate* (1967)	*The Grapes of Wrath* (1940)	*The Great Gatsby* (2013)
The Green Mile (1999)	*The Grifters* (1990)	*The Grudge* (2004)	*The Hangover* (2009)
The Haunting (1999)	*The Hebrew Hammer* (2003)	*The Help* (2011)	*The Holdovers* (2023)
The Hudsucker Proxy (1994)	*The Hunt for Red October* (1990)	*The Hurt Locker* (2008)	*The Imaginarium of Doctor Parnassus* (2009)
The Invention of Lying (2009)	*The Iron Lady* (2011)	*The Island* (2005)	*The Italian Job* (2003)
The Jacket (1991)	*The Kids Are All Right* (2010)	*The Killer* (2023)	*The King of Comedy* (1982)
The Kingdom (2007)	*The Ladykillers* (2004)	*The Last Boy Scout* (1991)	*The Last Flight* (1931)
The Last of the Mohicans (1992)	*The Last Samurai* (2003)	*The Last Station* (2009)	*The LEGO Movie* (2014)
The Lincoln Lawyer (2011)	*The Little Mermaid* (1989)	*The Long Kiss Goodnight* (1996)	*The Losers* (2010)
The Man Who Knew Too Much (1956)	*The Manchurian Candidate* (2004)	*The Martian* (2015)	*The Master* (2012)
The Matrix (1999)	*The Matrix Reloaded* (2003)	*The Men Who Stare at Goats* (2009)	*The Menu* (2022)
The Miracle Worker (1962)	*The Mummy* (1999)	*The Next Three Days* (2010)	*The Nightmare Before Christmas* (1993)
The Ninth Gate (1999)	*The Other Boleyn Girl* (2008)	*The Pacifier* (2005)	*The Patriot* (2000)
The Perks of Being a Wallflower (2012)	*The Pianist* (2002)	*The Piano* (1993)	*The Postman* (1997)
The Prestige (2006)	*The Princess Bride* (1987)	*The Proposal* (2009)	*The Queen* (2006)
The Reader (2009)	*The Replacements* (2000)	*The Rescuers Down Under* (1990)	*The Revenant* (2016)
The Rock (1996)	*The Rocky Horror Picture Show* (1975)	*The Roommate* (2011)	*The Searchers* (1956)
The Secret Life of Walter Mitty (2013)	*The Sessions* (2012)	*The Seventh Seal* (1957)	*The Shawshank Redemption* (1994)
The Shining (1980)	*The Social Network* (2010)	*The Sting* (2003)	*The Substance* (2024)
The Talented Mr. Ripley (1999)	*The Thing* (1982)	*The Three Musketeers* (1993)	*The Time Machine* (2002)
The Tourist (2010)	*The Town* (2010)	*The Truman Show* (1998)	*The Ugly Truth* (2009)
The Usual Suspects (1995)	*The Verdict* (1982)	*The Visitor* (2008)	*The Way Back* (2011)
The Whistleblower (2011)	*The Wizard of Oz* (1939)	*The Wolf of Wall Street* (2013)	*The World is not Enough* (1999)
The Wrestler (2009)	*The Zone of Interest* (2023)	*They* (2002)	*Thirteen Days* (2000)
Thor (2011)	*Three Kings* (1999)	*Thunderbirds* (2004)	*Tin Cup* (1996)
Tinker Tailor Soldier Spy (2011)	*Titanic* (1997)	*TMNT* (2007)	*Tombstone* (1993)
Tomorrow Never Dies (1997)	*Top Gun* (1986)	*Total Recall* (1990)	*Toy Story* (1995)
Traffic (2000)	*Training Day* (2001)	*Trainspotting* (1996)	*TRON* (1982)
Tropic Thunder (2008)	*True Grit* (2010)	*True Lies* (1994)	*True Romance* (1993)
Tucker and Dale vs Evil (2010)	*Twilight* (2008)	*Twin Peaks* (1992)	*Twins* (1988)

continued on next page

Table 4 – *continued*

Film (year)			
Twister (1996)	*Unbreakable* (2000)	*Under Fire* (1983)	*Unknown* (2011)
Up (Not Avail.)	*Up in the Air* (2009)	*V for Vendetta* (2006)	*Valkyrie* (2008)
Vanilla Sky (2001)	*Walking Tall* (2004)	*Wall Street* (1987)	*Wanted* (2008)
War for the Planet of the Apes (2017)	*War Horse* (2011)	*Warrior* (2011)	*Watchmen* (2009)
Water for Elephants (2011)	*We Own the Night* (2007)	*What Lies Beneath* (2000)	*When a Stranger Calls* (1979)
While She Was Out (2008)	*Whiplash* (2014)	*White Christmas* (1954)	*Whiteout* (2009)
Wicked (2024)	*Wild At Heart* (1990)	*Wild Hogs* (2007)	*Wild Things* (1998)
Wild Wild West (1999)	*Willow* (1988)	*Win Win* (2011)	*Witness* (1985)
Wonder Boys (2000)	*Wonka* (2023)	*xXx* (2002)	*Year One* (2009)
Yes Man (2008)	*Youth in Revolt* (2010)	*Zero Dark Thirty* (2013)	*Zootopia* (2016)

Received 13 January 2026; revised 25 March 2026; accepted 16 April 2026