
THE UNIVERSITY OF CHICAGO

# Swipe Left, or Right? Decoding Generative AI's Believable Social Behavior and Decision-Making in a Simulated Dating App

By

Aidi Li

June 2025

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Dr. Zhao Wang

Preceptor: Dr. Henry Dambanemuya

## Abstract

Dating apps shape romantic opportunities but remain inaccessible to researchers due to commercial secrecy and privacy concerns. We developed an alternative approach using authentic user profiles and Large Language Model agents,powered by GPT-4o to simulate dating decisions in a controlled environment. These agents engaged in 80,000+ simulated interactions, generating both choices and explanations for their preferences. Our supervised machine learning analysis (regression models, decision trees) and unsupervised semantic clustering revealed consistent patterns: education emerged as the dominant predictor of dating desirability across demographic groups. Gender differences were particularly pronounced in evaluation hierarchies—female agents prioritized income in male profiles, while male agents emphasized age when assessing females. Control experiments demonstrated high behavioral fidelity, with agents consistently acting according to their assigned demographic attributes and producing reasoning patterns that closely mirror documented human mate selection criteria. This methodology offers a transparent believable method for studying bias patterns in social decision making typically hidden behind proprietary algorithms.


**Keywords:** Large Language Models; Online Dating; Social Biases; AI Agents; Elo Ratings; Social Actor

# Contents

# 1 Introduction

Dating apps have transformed how people find romantic partners. These digital platforms now mediate millions of potential connections daily, replacing traditional face-to-face encounters with algorithm-driven matches. This shift represents a profound change from earlier forms of romantic communication where, as Barthes noted, exchanges required mutual emotional investment that shaped both connection and identity (Bandinelli, 2022).

As algorithmic systems increasingly mediate critical domains of social decision-making, researchers face mounting barriers to independent examination. Commercial secrecy, privacy regulations, and ethical constraints severely limit direct access to the underlying mechanisms shaping user interactions. Nowhere is this tension more pronounced than in romantic selection, where matching algorithms operate in private, protected spaces yet influence relationship formation in ways with profound social consequences. Despite growing concerns, the lack of transparent research tools has left key questions about algorithmic bias and social reproduction largely unanswered.

This study proposes a new simulation-based methodology to address these challenges: using Large Language Models (LLMs) as social proxies within a controlled dating environment. Rather than studying human participants directly, we deploy LLM agents to emulate decision-making processes based on authentic profile data. This design allows systematic measurement, analysis, and interpretation of preference patterns typically hidden within proprietary platforms. Our approach aligns with emerging work treating LLMs not merely as generative systems, but as tools for modeling socially situated behaviors under constrained conditions (Park et al., 2023).

Our research focus on two core questions:

1. **What decision-making mechanisms drive AI agents' swiping choices?**
2. **How are human biases introduced into and reflected in these AI-driven decisions?**

To investigate these questions, we hypothesize that AI agents will exhibit homophily bias, demonstrate systematic popularity patterns based on specific demographic attributes, and reflect intersectional biases in their decision rationales. The study quantifies social desirability through an adapted Elo rating system and analyzes linguistic patterns using cosine similarity metrics. This dual approach reveals both explicit demographic biases (through supervised machine learning approach) and implicit linguistic biases (through unsupervised analysis).

Through this framework, we aim to illuminate how LLM agents, trained on human-generated language data, internalize and operationalize social hierarchies. Our findings

offer a replicable, transparent proxy for studying opaque systems, a lens for understanding how algorithmic mediation might subtly reproduce patterns of social valuation.

# 2 Literature Review

## 2.1 Dating Apps: Algorithmic Mediation and Research Challenges

Dating applications have transformed romantic initiation from face-to-face encounters into algorithm-driven interactions, influencing not only who meets whom but how people conceptualize desirability itself (Bandinelli, 2022; Wu & Trottier, 2022). The "swipe" mechanics of these platforms reduce complex social evaluations to binary choices, prioritizing efficiency and entertainment over emotional investment (Bandinelli & Cossu, 2023; Bown, 2022). This mechanization reflects what Bandinelli terms a shift toward "de-romanticized social practice" where relationship formation increasingly resembles marketplace transactions governed by reputational metrics.

Research on these platforms faces substantial obstacles. Dating companies guard their matching algorithms as proprietary secrets, user data is protected by privacy regulations, and conducting field experiments raises ethical concerns (Castro & Barrada, 2020; Wu & Trottier, 2022). These constraints have created significant knowledge gaps about how algorithmic systems influence romantic decision-making and potentially reinforce existing social biases.

Studies examining dating outcomes consistently reveal selection patterns that echo broader social hierarchies. Online daters show strong preferences for partners sharing similar educational backgrounds and levels of physical attractiveness, producing matching structures that closely mirror offline social stratification (Hitsch et al., 2010). These patterns are further shaped by aspirational pursuits, where users disproportionately initiate contact with more socially desirable partners, intensifying competitive dynamics and reinforcing inequality within dating markets (Bruch & Newman, 2018). Moreover, large-scale behavioral analyses reveal that biases based on race, education, and age persist even when explicit preferences are not stated, suggesting that algorithmic mediation does not neutralize but rather subtly perpetuates existing social valuations (Rudder, 2014).

Similarly, research on both heterosexual and same-sex dating platforms shows preference patterns that privilege certain demographic characteristics while devaluing others (Chan, 2019; Wu & Ward, 2018). These preferences cannot be dismissed as merely individual tastes; they reflect and often reinforce cultural patterns of value and devaluation that advantage certain groups while disadvantaging others.

## 2.2 LLMs as Social Actors and Research Instruments

Recent advances in Large Language Models have created unprecedented opportunities for studying complex social phenomena. Modern LLMs function as what Park and colleagues term "generative agents"—computational entities capable of belief formation, preference development, and goal-directed social behavior (Park et al., 2023). Unlike earlier AI systems limited to narrowly defined tasks, these models can engage in sophisticated social reasoning that closely approximates human decision-making processes across diverse contexts.

The emergence of multi-agent LLM systems represents a significant methodological breakthrough. Li's CAMEL framework demonstrates how multiple LLM instances can engage in goal-oriented communication without human intervention (Li et al., 2023), while Wang's Voyager project shows how LLM-powered agents can explore environments, acquire skills, and make novel discoveries autonomously (Wang et al., 2023). These capabilities enable the creation of social simulations at scales and with control parameters previously impossible in human participant research.

Crucially, LLMs derive their capabilities from vast training datasets of human-generated text, inevitably capturing social biases embedded within human culture. As Vaswani explains in work on transformer architectures, these models create vector representations that encode semantic relationships between concepts, including socially constructed associations that reflect cultural biases (Vaswani et al., 2023). This characteristic makes LLMs valuable instruments for studying human social biases in contexts where direct experimentation would be challenging.

## 2.3 LLMs as Mirrors of Human Bias in Social Contexts

Multiple studies confirm that LLMs reliably reproduce measurable human biases across various domains. Hanna's research shows that LLMs generate different healthcare recommendations based on patient demographics, mirroring disparities documented in human healthcare provision (Hanna et al., 2023). Luo finds that LLM responses to identical queries vary significantly based on the language used, reflecting cultural biases embedded in different language communities (Luo et al., 2023). Suguri Motoki's work reveals systematic political biases in LLM outputs that parallel patterns in human political discourse (Suguri Motoki et al., 2023).

These biases manifest not only in explicit content but in the underlying vector representations themselves. Thongtan and Phienthrakul demonstrate how sentiment classification techniques can reveal subtle patterns of association within these models' semantic spaces (Thongtan & Phienthrakul, 2019). Such patterns provide a window into how social biases are encoded and reproduced through linguistic structures, offering insights that might not

be accessible through traditional research methods.

The convergence of these research streams creates a compelling methodological opportunity. By deploying LLM-powered agents in a simulated dating environment, we can bypass traditional research constraints while gaining insights into how socially informed decision patterns may be reproduced in computational systems. While our focus is not on auditing specific dating platforms, but rather on modeling bias re-emergence in agentic decision-making, we discuss this distinction and its implications more detailed in the limitations section.

This approach eliminates privacy concerns, allows precise control over parameters, and provides transparency into decision processes that remain opaque in commercial platforms (Gatter & Hodkinson, 2016; Park et al., 2023). While this methodology inherits certain limitations—the behavior of LLM agents inevitably reflects both architectural constraints and prompt design—it offers a novel window into the complex interplay between technology design and social bias. By positioning LLMs simultaneously as social actors and as instruments for studying bias, we create a framework for examining how algorithmic systems might shape dating possibilities in ways that either challenge or reinforce existing social hierarchies.

# 3 Data & Methods

## 3.1 Data

### 3.1.1 Dataset Selection and Preparation

This study employs the publicly available OkCupid dataset (https://www.kaggle.com/datasets/andrewmvd/ok profiles/data), which contains approximately 60,000 authentic user profiles from the popular dating application. Initial data cleaning procedures removed incomplete entries with missing values or invalid data points, resulting in a working dataset of 8,767 complete profiles. From this refined dataset, we extracted ten key features across three categories (Table 1).

| Category | Features |
|---|---|
| Numerical attributes | `age`, `height`, `income` |
| Categorical variables | `sex`, `orientation`, `ethnicity`, `education`, `religion` |
| Text fields | "My self summary", "The first thing people usually notice about me" |

Table 1: Key features extracted from the OkCupid dataset

### 3.1.2 Dimensional Reduction and Binarization

For analytical tractability, we selected six key dimensions (`sex`, `ethnicity`, `age`, `income`, `education`, and `height`) based on their significance in previous dating research (Hitsch et al., 2010; Wu & Trottier, 2022). Each dimension was converted into a binary attribute to facilitate intersectional analysis while maintaining statistical power.

| Features Dimension | Binary Categories | Binarization Method |
|---|---|---|
| `sex` | Male / Female | Original categorical |
| `ethnicity` | White / Minority | Grouped categories |
| `age` | Younger / Older | Split at median (32 years) |
| `income` | Low-Mid / High | Split at 80th percentile |
| `education` | Low / Mid-High | Grouped by degree level |
| `height` | Not-Tall / Tall | Split at gender-specific medians |

Table 2: Binary encoding of demographic dimensions and corresponding classification rules

Continuous variables were binarized using appropriate statistical thresholds, while categorical variables were grouped into logical clusters (Table 2). This binarization serves analytical purposes only and does not constrain the full attribute complexity available to AI agents during their decision-making processes.

### 3.1.3 Sampling Strategy

Each agent in the simulation was instantiated from a real user profile drawn from the OkCupid dataset. These profiles serve as the informational basis for agent identity, providing demographic attributes and narrative self-descriptions used in decision-making. To ensure diverse and balanced representation across key social dimensions, we employed an intersectional sampling strategy targeting variation in `sex`, `ethnicity`, `age`, `income`, `education`, and `height`.

We employed an intersectional sampling approach to ensure comprehensive representation across demographic combinations. The six binary dimensions theoretically yield $2^6 = 64$ possible intersectional categories. Our sampling protocol targeted 20 profiles from each valid category, for an intended sample of 1,280 profiles.

However, 26 of the 64 theoretical categories contained fewer than 20 representatives in the original dataset—a reflection of real-world demographic distributions rather than methodological limitations. Rather than artificially creating synthetic profiles, we excluded these underrepresented categories, resulting in 760 valid profiles across 38 intersectional categories. Each profile was used to instantiate one LLM-based agent, producing a total of

760 experimental agents. These agents form the core of our simulation and are tasked with evaluating potential matches based on profile content and generating decision rationales.

| Attribute | Category | Count | Percentage |
|---|---|---|---|
| `sex` | Male | 430 | 56.6% |
| | Female | 330 | 43.4% |
| `ethnicity` | White | 400 | 52.6% |
| | Minority | 340 | 44.7% |
| `age` | Older | 440 | 57.9% |
| | Younger | 300 | 39.5% |
| `income` | Low-Mid | 560 | 73.7% |
| | High | 180 | 23.7% |
| `education` | Mid-High | 500 | 65.8% |
| | Low | 240 | 31.6% |
| `height` | Not-Tall | 440 | 57.9% |
| | Tall | 300 | 39.5% |

Table 3: Distribution of binary-encoded demographic attributes in the final profile dataset

The final profile set exhibits distributional characteristics shown in Table 3. While not perfectly balanced, this distribution maintains sufficient representation across all key dimensions to support robust statistical comparisons while reflecting authentic demographic patterns observed in real dating platforms.

To establish behavioral benchmarks, we supplemented the simulation with 20 control agents programmed with fixed decision-making patterns, bringing the total agent population to 780. These controls serve as calibration points to assess the behavioral consistency of experimental agents under known conditions.

For demographic combinations with insufficient representation for inclusion in the experimental group, we note the most significant underrepresentation occurred among profiles combining high-`income`, younger, minority females with low `education`. This exclusion reflects actual demographic distributions rather than sampling bias, though it does constrain the generalizability of findings to these specific intersectional categories.

## 3.2 Methods

### 3.2.1 Agent Interaction Design

We implemented a controlled simulation environment in which LLM-based agents, instantiated from real dating profiles, engage in structured decision-making tasks. Each agent is as-

signed a complete profile—including demographic attributes and self-description text—and is asked to evaluate potential matches based on this identity. The simulation focuses on swiping decisions, operationalized as structured preference choices made under constrained informational settings.

Rather than using conventional one-to-one (binary) presentation common in commercial dating apps, we adopted a batched-sequential interaction framework that enables comparative evaluation (Figure 1). This design responds to documented limitations in LLM-based simulations, particularly their tendency to exhibit excessive agreeableness in isolated binary decision tasks—a behavioral artifact attributed to alignment training, where models default to positive or affirming responses when uncertain (Ganguli et al., 2022; Perez et al., 2022). In our task context, such behavior undermines the ability to detect the the preference structure, as agecan accept any match presentedcomplete" the interaction.

To address this, we structured each decision round as a forced-choice task: agents are presented with a batch of potential partners and must select exactly one preferred candidate—or reject the entire batch if none are acceptable. This comparative setup encourages discriminative judgment and more closely mirrors how real users make selections across multiple options during dating app sessions.

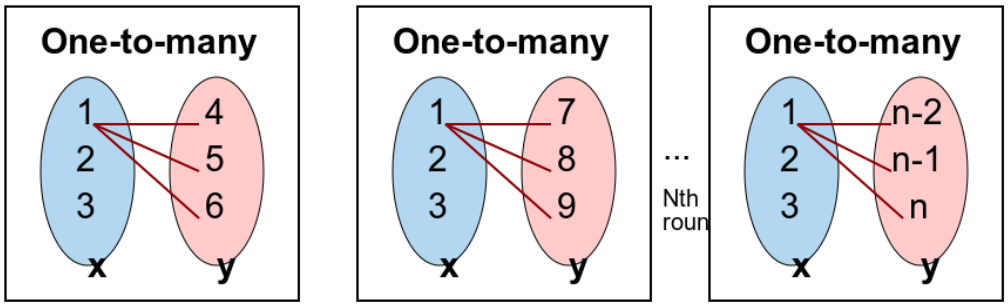


Figure 1: Batched-sequential interaction framework. In each round, a fixed set of agents (X) evaluates a new batch of opposite-group profiles (Y) in a one-to-many format. The batches iterate across rounds until all potential matches have been evaluated exactly once.

To ensure that each agent evaluated every potential opposite-gender match exactly once, we implemented a batched interaction design. In each decision round, a female agent was shown `batch_x` male profiles, while a male agent viewed `batch_y` female profiles. These batch sizes determined how many candidates were visible to each agent per round and were tuned to balance exposure across genders despite unequal group sizes (430 males, 330 females).

We calculated the total number of rounds $R$ needed to ensure full pairwise coverage as:

$$R = \max\left(\left\lceil \frac{F}{\text{batch}_y} \right\rceil, \left\lceil \frac{M}{\text{batch}_x} \right\rceil\right), \quad \text{where } M = \text{total male agents}, \ F = \text{total female agents} \tag{1}$$

This formulation guaranteed that each agent encountered every eligible partner of the opposite gender exactly once across the experiment. By structuring evaluation as a forced-choice among multiple candidates per round, rather than single binary decisions, we also mitigated common LLM artifacts such as default-acceptance bias (Perez et al., 2022). Compared to sequential one-to-one designs, this format introduced a comparative reasoning constraint more consistent with real-world user behavior on dating platforms.

The batched-sequential framework offers three key advantages:

1. It reduces over-selection artifacts stemming from LLMs' tendency to affirm in low-discrimination contexts.

2. It enforces within-round contrastive evaluation, increasing behavioral resolution and decreasing random agreement.

3. It enables full agent-to-agent pairing under controlled exposure conditions, supporting interpretable comparisons of social desirability.

### 3.2.2 LLM Prompt Design

The simulation's behavioral fidelity depended on prompt engineering (Figure 2). We developed a two-component architecture consisting of system and task prompts:

System prompts established agent identity by incorporating complete profile information from the original OkCupid dataset. Rather than fabricating artificial personas, this approach grounded agent behavior in authentic self-presentation patterns from real dating app users. Critically, the system prompt instructed agents to make decisions based entirely on their assigned profile characteristics without imposing predetermined value systems or decision criteria.

Task prompts delivered candidate profiles for evaluation, with presentation order randomized to mitigate position bias—a documented tendency of LLMs to favor items based on their order in a list, particularly under uncertainty (Pezeshkpour & Hruschka, 2023). Each agent returned two parameters: (1) the unique identifier of their selected match, and (2) a brief explanation justifying their choice. If an agent found no suitable matches within a batch, they could return a value of `''-1''`, indicating rejection of all presented candidates.

To ensure response consistency, we implemented OpenAI's JSON mode, which standardized output formatting while allowing natural variation in decision rationales.

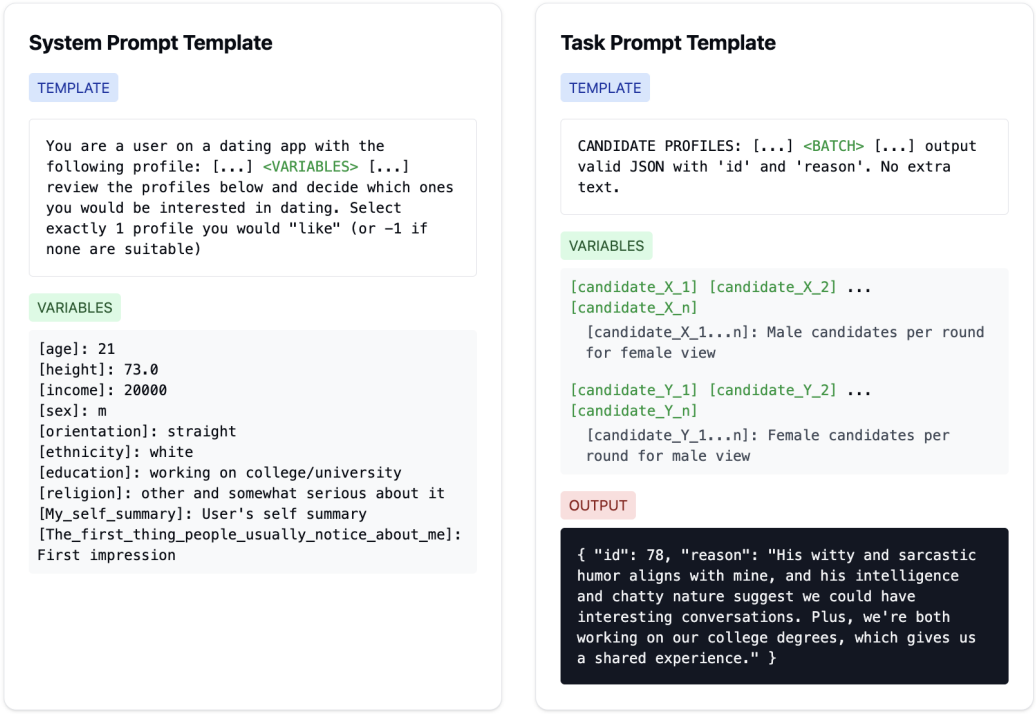


Figure 2: Prompt engineering interface that dynamically assembles system and task prompts for AI agents, using variable substitution to personalize dating profile evaluations while enforcing structured JSON responses for consistent decision-making.

### 3.2.3 Quantifying and Analyzing Social Desirability Patterns

To measure and analyze social attractiveness in our simulated dating environment, we integrated a modified Elo rating system with complementary statistical approaches. Originally developed for chess rankings and previously used by dating applications, the Elo system provides a transparent framework for quantifying relative desirability based on interaction outcomes.

$$\text{ELO}_{\text{new}} = \begin{cases} \text{ELO}_{\text{current}} + 10 & \text{if mutual selection} \\ \text{ELO}_{\text{current}} + 5 & \text{if selected only} \\ \text{ELO}_{\text{current}} - 5 & \text{if selecting only} \\ \text{ELO}_{\text{current}} & \text{if neither selects} \end{cases} \tag{2}$$

| Scenario | Male Action | Female Action | Male ELO | Female ELO |
|---|---|---|---|---|
| Mutual Match | Like | Like | +10 | +10 |
| One-way Interest | Like | Ignore | -5 | +5 |
| Reversed Interest | Ignore | Like | +5 | -5 |
| No Interest | Ignore | Ignore | 0 | 0 |

Table 4: Elo score adjustments based on interaction outcomes. Control agents with specific sexual orientation preferences serve as stability checks in the system.

Each agent begins with a 1400-point baseline score, with subsequent adjustments reflecting dating interaction outcomes (Equation 2 and Table 4). Unlike chess ratings that fluctuate based on win probabilities, our system uses fixed value increments: +10 points for mutual matches, +5 points for being selected, -5 points for unrequited selection, and no change when neither agent selects the other. We incorporated 20 gay-identified agents as control cases who systematically decline opposite-sex matches, with their Elo scores remaining unchanged at 1400 throughout the experiment to provide reference points validating system reliability.

To analyze the relationships between demographic attributes and dating success (measured by final Elo scores), we implemented two complementary machine learning approaches. We assess the magnitude and robustness of such effects in the Results section.:

1. **Linear regression** provides interpretability through feature coefficients, allowing us to quantify the marginal contribution of each demographic attribute to predicted desirability. For instance, a positive coefficient for `younger_group` would indicate that, all else equal, younger profiles are predicted to be more desirable.

2. **Decision trees** capture non-linear relationships and interaction effects by partitioning data based on the most discriminative features at each node. This approach reveals complex conditional dependencies between attributes that might be missed by linear models.

For both modeling approaches, we conducted parallel analyses using two feature representations: raw demographic values (`age`=23, `income`=75000) and binary categorical features (`younger_group` and `high_income`). This dual approach leverages complementary advantages—raw values preserve granular information while categorical features better capture threshold effects in dating preferences. By analyzing the patterns of Elo score distributions across demographic dimensions, we can identify systematic biases in AI decision-making that might reflect and potentially amplify human social preferences.

#### 3.2.4 Semantic Analysis of Decision Reasoning

Beyond quantitative outcomes, we examined the qualitative reasoning behind agent preferences. Each agent provided textual explanations for their "like" decisions, offering a window into how preferences are articulated and rationalized. We randomly sampled five stated reasons from each agent, yielding 2,150 samples from male agents and 1,650 from female agents.

These textual expressions were transformed into high-dimensional semantic vectors using the Nomic embedding model, which captures contextual relationships between words and phrases. The resulting vectors underwent hierarchical clustering analysis to identify natural groupings in the data without imposing predetermined categories. This approach reveals thematic structures within decision reasoning and enables visualization as a semantic dendrogram, with branch height representing conceptual distance between reasoning clusters. By comparing cluster formations between demographic groups, we identified systematic differences in how agents articulate attraction, revealing both explicit and implicit patterns of preference that might not appear in quantitative analysis alone.

Our experiment yielded clear patterns in social desirability scores across the simulated dating environment. The mean Elo score was 1403 (SD = 254.66), indicating modest movement from the baseline of 1400. However, this aggregate measure conceals significant gender differences: female agents averaged significantly higher scores (M = 1466.68) than male agents (M = 1353.34). The distribution showed pronounced right skewness, with some agents achieving substantially higher popularity than would be expected by chance.

# 4 Results

Our experiment revealed distinct patterns in social desirability across the simulated dating environment. While the overall mean Elo score (1403, SD = 254.66) showed modest movement from the baseline of 1400, this aggregate measure conceals significant gender differences: female agents averaged substantially higher scores (M = 1466.68) than male agents (M = 1353.34). The distribution showed pronounced right skewness, with some agents achieving popularity levels far exceeding random chance expectations.

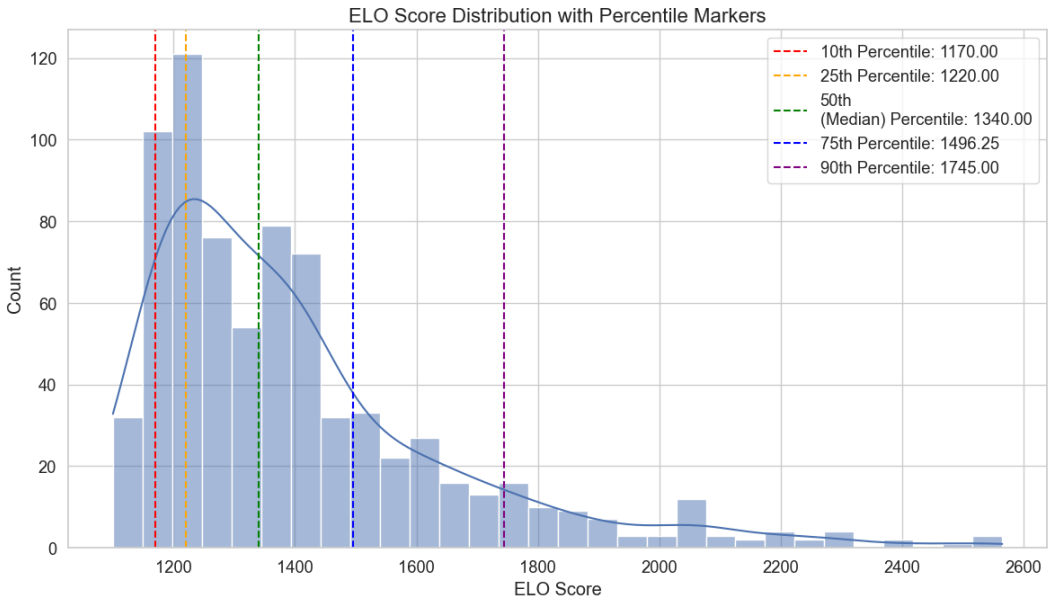


Figure 3: ELO Score Distribution with Percentile Markers. The histogram shows a right-skewed distribution, a small subset of agents achieved disproportionately high popularity scores.

The distribution analysis reveals substantial variation in dating outcomes despite the relatively modest shift in mean scores. As shown in Figure 5, the 10th and 90th percentiles (1170.00 and 1745.00, respectively) span a range of 575 points, indicating pronounced stratification in social desirability. This pattern aligns with observations in real-world dating markets, where attractiveness distributions typically follow power-law rather than normal distributions.

The gender disparity in Elo scores represents a key finding that informs subsequent analyses. This systematic difference suggests that AI agents have internalized gender-specific evaluation standards from their training data, despite receiving no explicit instructions about gender preferences. These baseline differences provide essential context for understanding the subsequent models of demographic determinants of dating success.

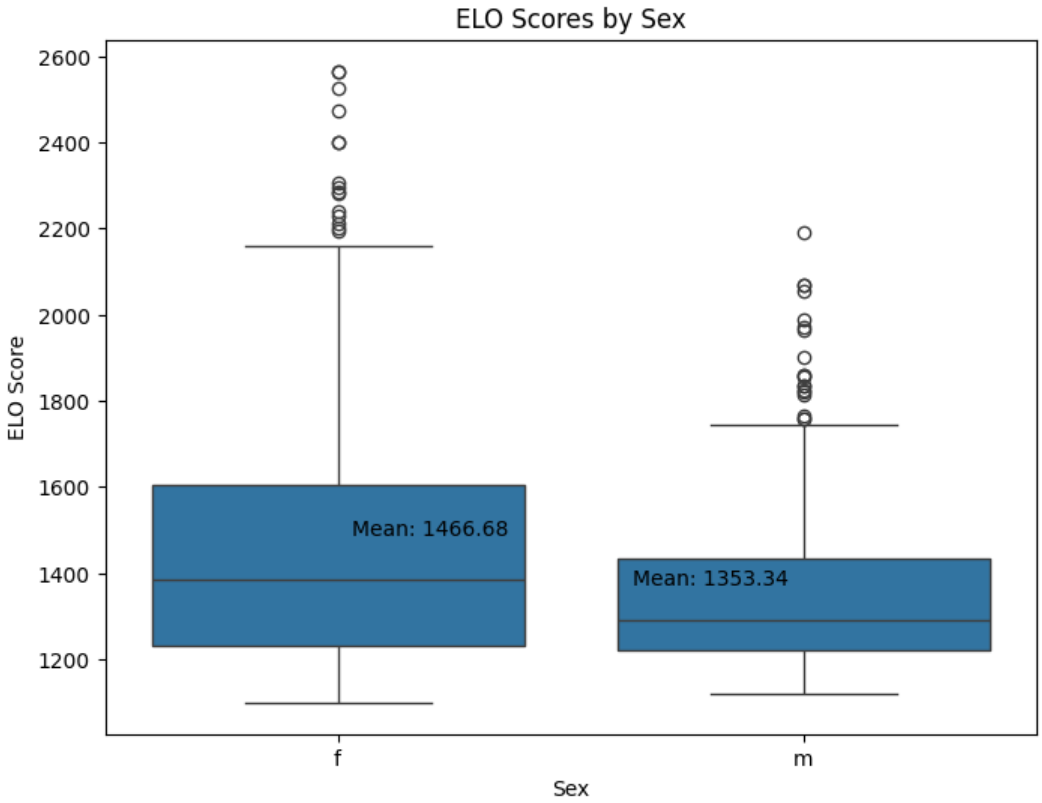


Figure 4: ELO Scores by Sex, with female agents also showing greater variance and more extreme outliers at the upper end of the distribution.

### 4.1 Control Group Validation

Agents designated with gay sexual orientation served as methodological controls since they were expected to maintain baseline scores in an opposite-sex matching environment. These agents achieved a mean Elo score of 1422.85 (SD = 64.14), with no statistically significant difference from the experiment group ($p = 0.0670$). The relatively low standard deviation among these control agents compared to the general population confirms that agents maintained fidelity to their demographic identities rather than selecting randomly. The slight deviation from the perfect baseline of 1400 likely reflects occasional interactions with agents identifying as bisexual rather than systematic selection errors.

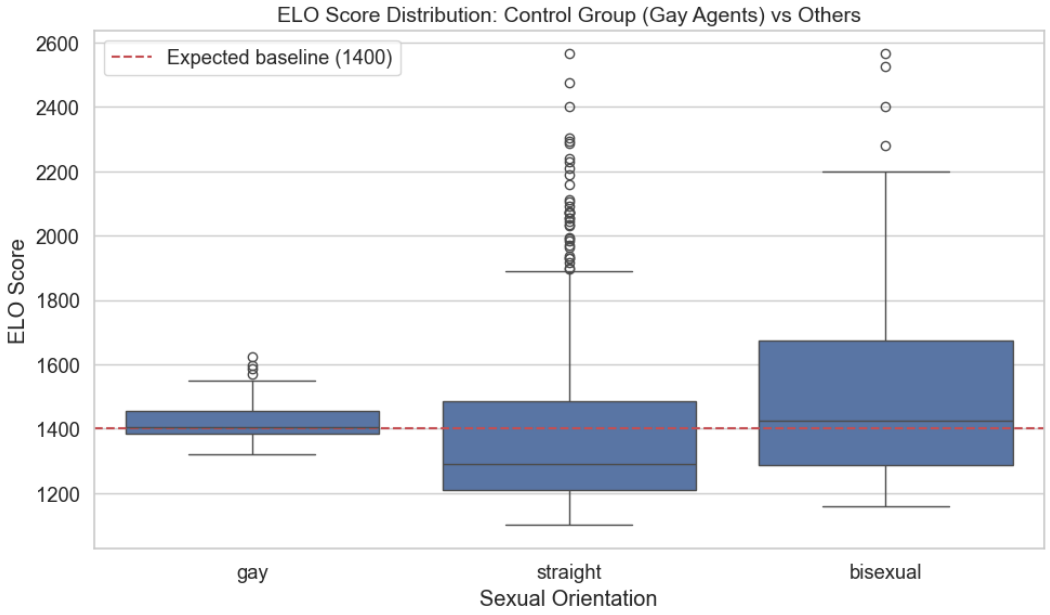


Figure 5: Elo Score Distribution: Control Group (Gay Agents) vs straight and bisexual agents. The control group shows a markedly stable distribution compared to the experimental groups, confirming that agent behavior follows expected patterns based on their assigned demographic characteristics.

Figure 5 illustrates the stark contrast between the stable Elo distribution of gay agents and the more dispersed scores of straight and bisexual agents. This pattern provides strong validation that our experimental framework accurately captures preference-driven social dynamics rather than random selection behavior. The narrow distribution of control group scores around the baseline demonstrates that agents reliably implement decisions consistent with their assigned sexual orientation.

### 4.2 Predictive Regression Models of Social Desirability

We developed two parallel regression approaches—one using original continuous features and another using binary categorical features—to identify demographic predictors of social desirability.

| Performance Metric | Original Features Model | Binary Features Model |
|---|---|---|
| Test $R^2$ | 0.0779 | 0.0842 |
| Cross-validation $R^2$ | 0.0405 | 0.0244 |
| Test MSE | 61502.80 | 61084.39 |
| Regularization | Ridge ( = 50.0) | Ridge ( = 50.0) |

Table 5: Performance comparison of predictive models for social desirability

The regularized linear regression model using original features achieved modest but significant predictive power (Table 5). `education` emerged as the strongest predictor, with high school education or less receiving severe penalties. `gender` and `sexual_orientation` also showed substantial effects, with male and straight agents receiving lower scores. Religious affiliations demonstrated moderate effects, with atheism and Christianity both reducing attractiveness. `ethnicity` showed milder effects, with mixed and Indian backgrounds receiving modest bonuses.

| Feature | Coefficient | Feature | Coefficient |
|---|---|---|---|
| Education (High school or less) | -68.30 | Education (Mid-to-high) | +120.84 |
| Sexual orientation (Straight) | -47.60 | Sexual orientation (Straight) | -102.32 |
| Gender (Male) | -45.59 | Sexual orientation (Gay) | -54.81 |
| Religion (Atheism) | -35.22 | Gender (Male) | -41.50 |
| Religion (Christianity) | -23.15 | Income (Low-to-mid) | +35.83 |
| Ethnicity (Mixed) | +19.95 | Ethnicity (Minority) | +18.75 |
| Ethnicity (Indian) | +17.37 | Age (Younger) | +12.46 |

Table 6: Top coefficients for original features model (left) and binary features model (right)

The binary feature model yielded similar predictive performance (Table 5). This model confirmed the `education` effect, showing an even stronger coefficient for mid-to-high education levels. `sexual_orientation` again emerged as a major factor, with straight and gay orientations both reducing scores compared to bisexual orientation. `gender` and `income` also significantly influenced outcomes (Table 6).

Both models converged on several key findings:

1. `education` level constitutes the strongest predictor of social desirability, with higher education conferring significant advantages

2. `gender` and `sexual_orientation` effects consistently show penalties for male and exclusively heterosexual agents

3. `ethnicity` predictors show relatively modest effects compared to socioeconomic factors like `education`

The binary model revealed additional nuance by highlighting `income` effects that were less apparent in the original feature model, while the original feature model provided finer-grained insights into religious preferences.

These converging results from distinct modeling approaches strengthen confidence in our findings, revealing systematic demographic biases in AI dating decisions that likely reflect patterns in the training data derived from human behavior. The models' modest $R^2$ values indicate that while demographic attributes significantly predict desirability, substantial variance remains unexplained by these factors alone, pointing to the complex, multi-faceted nature of attraction even in simulated environments.

### 4.3 Decision Tree Analysis Reveals Gendered Evaluation Hierarchies

Our decision tree models uncovered not just predictive factors but distinctly gendered evaluation pathways in dating preferences. Despite their shallow optimal depth (max_depth = 3), these models revealed nuanced mechanisms behind attractiveness assessments (Table 7).

| Model | Best Parameters | MSE | $R^2$ | Mean CV $R^2$ |
|---|---|---|---|---|
| Original Features | {max_depth: 3, min_samples_leaf: 2} | 61218.35 | 0.0822 | -0.0143 |
| Binary Features | {max_depth: 3, min_samples_leaf: 1} | 61669.64 | 0.0754 | 0.0258 |

Table 7: Decision tree model performance metrics

The tree structure revealed gender as the foundational split, creating distinct evaluation frameworks that mirror documented sociological patterns (Figure 6). For female profiles, `age` emerged as the critical secondary factor (importance 0.334), reflecting cultural emphasis on youth in feminine attractiveness standards. For male profiles, `income` became the decisive secondary determinant (importance 0.184), reinforcing traditional provider role expectations. This gender-specific bifurcation shows how AI agents recapitulate longstanding social biases where women are evaluated primarily on physical attributes while men face scrutiny of economic status.

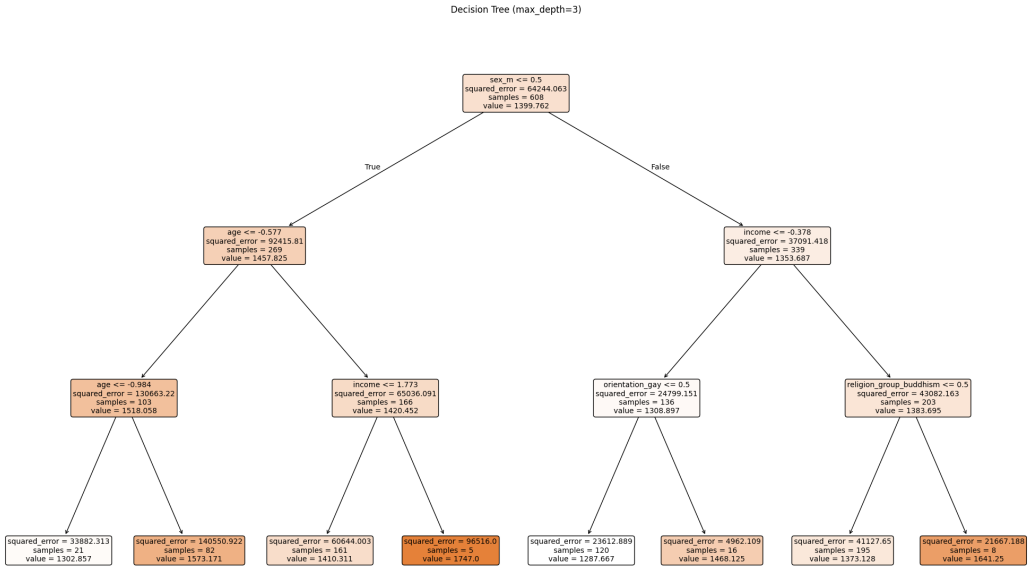


Figure 6: Original feature decision tree structure showing gender as the primary split, followed by age-based evaluation for female profiles and income-based evaluation for male profiles. This reveals how AI agents apply different evaluation criteria based on gender, mirroring documented human dating preferences.

| Feature | Importance |
|---|---|
| `age` | 0.334 |
| `sex_m` | 0.297 |
| `income` | 0.184 |
| `religion_group_buddhism` | 0.101 |
| `orientation_gay` | 0.084 |

Table 8: Top five features by importance in the original decision tree model

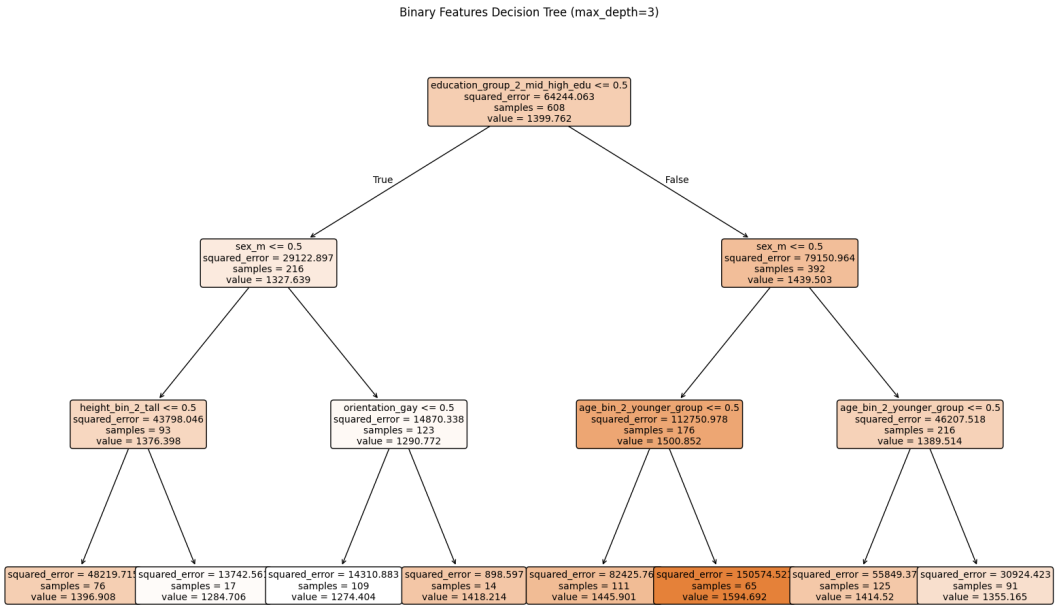


Figure 7: Binary feature decision tree revealing education level as the dominant overall predictor, even superseding gender in certain branches. This structure demonstrates how educational attainment functions as a class signifier that transcends gender boundaries in dating preferences.

The binary feature model revealed `education_group_2_mid_high_edu` (importance 0.359) as the strongest overall predictor, even outranking gender (Table 9, Figure 7). This suggests education functions as a class signifier that transcends gender boundaries, potentially serving as a proxy for cultural compatibility, intellectual engagement, and socioeconomic potential simultaneously. The emergence of education as the dominant binary predictor suggests dating decisions may increasingly reflect assortative mating patterns centered on cognitive and cultural capital rather than purely demographic characteristics.

| Feature | Importance |
|---|---|
| `education_group_2_mid_high_edu` | 0.359 |
| `sex_m` | 0.327 |
| `age_bin_2_younger_group` | 0.225 |
| `orientation_gay` | 0.053 |
| `height_bin_2_tall` | 0.036 |

Table 9: Top five features by importance in the binary decision tree model

Perhaps most striking is what these models revealed about evaluation sequence. The decision trees suggest human-like dating assessments follow a hierarchical filtering process rather than simultaneous attribute weighing. The primary gender split followed by gender-specific secondary criteria mirrors cognitive shortcuts documented in speed-dating research, where rapid initial categorization precedes more nuanced evaluation. This sequential filtering—particularly the role of gender as gatekeeper for how other attributes are interpreted—suggests AI agents have internalized not just attribute preferences but the very cognitive architecture humans employ in mate selection.

Notably absent from both trees' important features were several attributes that showed significance in linear models, including most `ethnicity` and `religion` categories. This absence suggests that while these attributes may subtly influence desirability scores, they lack the discriminative power to serve as primary sorting mechanisms in the decision hierarchy. Instead, they likely function as tertiary considerations applied only after the dominant socioeconomic and demographic filters have been satisfied—a finding consistent with research on how superficial preference statements often diverge from actual selection behavior in dating contexts.

### 4.4 Linguistic Analysis Reveals Deeper Selection Criteria

The text embedding analysis of agents' stated selection reasons revealed patterns both confirming our statistical findings and uncovering dimensions absent from profile attributes alone. Clustering of selection explanations produced distinct thematic categories (Tables

10, 11) that provide insights into the qualitative reasoning behind quantitative preferences.

| Index | Topic | Key Subtopics | Count | Percentage |
|---|---|---|---|---|
| 1 | Aspiring Academics | Personal Finance, Travel, Social Justice | 542 | 29.1% |
| 2 | Hispanic Culture | Personality Traits, Outdoor Activities | 319 | 17.2% |
| 3 | Navigating Complexity | Gaming, Spiritual Philosophy | 315 | 16.9% |
| 4 | Entertainment | Intellectual Interests, Education | 296 | 15.9% |
| 5 | Voting | Interpersonal Relationships, Sexuality | 290 | 15.6% |

Table 10: Male agent selection reasoning clusters. Interactive visualization available at: Atlas Link

| Index | Topic | Key Subtopics | Count | Percentage |
|---|---|---|---|---|
| 1 | Education | Science and Cinema, Christianity | 247 | 29.2% |
| 2 | Education (2) | Religion, Interpersonal Relationships | 226 | 26.7% |
| 3 | Culinary Lifestyle | Emotional Wellbeing, Communication | 195 | 23.1% |
| 4 | Interpersonal Relationships | Sports, Arts and Culture | 143 | 16.9% |
| 5 | Entertainment | Career, Outdoor Lifestyle | 20 | 2.4% |

Table 11: Female agent selection reasoning clusters. Interactive visualization available at: Atlas Link

Three significant findings emerge from this analysis:

First, the dominance of education-themed clusters in female agent reasoning (56

Second, the emergence of categories like "Interpersonal Communication," "Entertainment," and "Emotional Wellbeing" reveals that agents incorporate information beyond the structured demographic attributes. These themes correspond to content from the free-text fields ("My self summary" and "The first thing people usually notice about me"), demonstrating that LLMs leverage narrative self-descriptions in their decision-making. This finding suggests that dating preferences rely heavily on these implicit personality signals rather than just demographic matching—a phenomenon difficult to capture in traditional dating research using demographic variables alone.

Third, the gender-differentiated clustering patterns—with female agents showing more concentrated educational focus and male agents displaying more thematically diverse reasoning—explains the different decision pathways observed in our tree models. This suggests LLMs have internalized gendered patterns of mate selection criteria where women place greater emphasis on educational compatibility while men distribute attention across broader attribute categories.

## 5 Discussion & Conclusion

This study offers insights into both the mechanisms of LLM-based social decision-making and the methodological potential of using these models to study otherwise inaccessible social dynamics. The emergence of gendered evaluation pathways—where female agents prioritize education while male agents employ more diversified criteria—suggests that LLMs capture not just who is preferred in dating contexts but how these preferences are structured and applied. This finding extends beyond our initial research questions to illuminate the potential cognitive architecture behind AI social evaluations.

The substantial role of free-text descriptions in agent reasoning represents a critical finding with methodological implications. While traditional dating research typically focuses on demographic attributes, our linguistic analysis revealed that LLMs integrate narrative self-presentations into their decision-making in ways that statistical models alone cannot capture. This integration capability addresses a fundamental limitation in current computational social science approaches: the disconnect between quantifiable attributes and the rich textual data that often drives human social decisions (Park et al., 2023). However, this finding also raises questions about the black-box nature of embedding-based reasoning that merits further investigation.

The demographic imbalances in our dataset—particularly the gender disparity that necessitated differential batch sizing—represent a significant limitation that may have influenced our results. This technical constraint mirrors real-world dating platform demographics but complicates causal interpretation of gender-based preferences. The focus on a single LLM architecture also limits generalizability, as different models with varying training methodologies might produce distinct social bias patterns (Salinas et al., 2023). Additionally, the absence of visual inputs represents a substantial ecological validity concern, as physical appearance plays a central role in human dating decisions (Todorov et al., 2015).

These limitations highlight a fundamental tension in LLM-based social simulation: these models simultaneously reflect human biases while introducing their own algorithmic distortions. Distinguishing between faithfully reproduced human biases and artificially amplified ones remains a significant challenge. Future research should employ counterfactual testing approaches—systematically manipulating profile attributes while holding others constant—to better isolate causal mechanisms behind dating preferences. Multi-model comparisons could also help differentiate universal social biases from model-specific artifacts.

Beyond dating applications, the gendered decision pathways observed here likely extend to other domains where social evaluation occurs, from hiring decisions to customer service interactions. Understanding these embedded evaluation frameworks represents a crucial step toward developing more transparent AI systems that can be aligned with human values

while mitigating harmful biases.

## Data and Code Availability Statement

The agent framework and analysis code used in this study are available in the GitHub repository: https://github.com/MACSS-Projects/LLM_datingAPP_AidiL/tree/main. This repository contains all necessary files to reproduce the experimental environment, including agent simulation code, prompt engineering templates, and data analysis scripts.

The OkCupid dataset used for agent profile generation is publicly available through Kaggle (https://www.kaggle.com/datasets/andrewmvd/okcupid-profiles/data).

The interactive text embedding visualizations for male and female agent reasoning clusters are accessible via the following Nomic Atlas links:

- Male "Like" Reasons: https://atlas.nomic.ai/data/aidi/male-straight-like-reasons/map/88eb953a-eb3b-49f2-aba4-33fae8839c47#uGs2
- Female "Like" Reasons: https://atlas.nomic.ai/data/aidi/female-straight-like-reasons/map/5c6d2541-72bb-4f41-b72e-e0e6f442ffdd#Yw5U

All agent decisions logs, evaluation metrics, and statistical analysis outputs are provided in the repository's `data` and `code` directory.

## References


Bandinelli, C. (2022). Dating apps: Towards post-romantic love in digital societies [Publisher: Routledge]. *International Journal of Cultural Policy*, *28*(7), 905–919. https://doi.org/10.1080/10286632.2022.2137157

Bandinelli, C., & Cossu, A. (2023). Bye bye romance, welcome reputation: An analysis of the digital enclosure of dating [Publisher: SAGE Publications Ltd]. *Sexualities*, 13634607231152427. https://doi.org/10.1177/13634607231152427

Bown, A. (2022). *Dream lovers: The gamification of relationships*. Pluto Press. https://doi.org/10.2307/j.ctv2k4fwzw

Bruch, E. E., & Newman, M. E. J. (2018). Aspirational pursuit of mates in online dating markets [Publisher: American Association for the Advancement of Science]. *Science Advances*, *4*(8), eaap9815. https://doi.org/10.1126/sciadv.aap9815

Castro, Á., & Barrada, J. R. (2020). Dating apps and their sociodemographic and psychosocial correlates: A systematic review. *International Journal of Environmental Research and Public Health*, *17*. https://doi.org/10.3390/ijerph17186500

Chan, L. S. (2019). Paradoxical associations of masculine ideology and casual sex among heterosexual male geosocial networking app users in china. *Sex Roles*, *81*(7), 456–466. https://doi.org/10.1007/s11199-019-1002-4

Ganguli, D., Hernandez, D., Lovitt, L., DasSarma, N., Henighan, T., Jones, A., Joseph, N., Kernion, J., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Elhage, N., Showk, S. E., Fort, S., Hatfield-Dodds, Z., Johnston, S., … Clark, J. (2022, October 3). Predictability and surprise in large generative models. https://doi.org/10.1145/3531146.3533229

Gatter, K., & Hodkinson, K. (2016). On the differences between tinder™ versus online dating agencies: Questioning a myth. an exploratory study (M. Kolle, Ed.) [Publisher: Cogent OA _eprint: https://doi.org/10.1080/23311908.2016.1162414]. *Cogent Psychology*, *3*(1), 1162414. https://doi.org/10.1080/23311908.2016.1162414

Hanna, J. J., Wakene, A. D., Lehmann, C. U., & Medford, R. J. (2023). Assessing racial and ethnic bias in text generation for healthcare-related tasks by ChatGPT1. *medRxiv*, 2023.08.28.23294730. https://doi.org/10.1101/2023.08.28.23294730

Hitsch, G. J., Hortaçsu, A., & Ariely, D. (2010). Matching and sorting in online dating. *American Economic Review*, *100*(1), 130–163. https://doi.org/10.1257/aer.100.1.130

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023). CAMEL: Communicative agents for "mind" exploration of large language model society [Publisher: arXiv Version Number: 2]. https://doi.org/10.48550/ARXIV.2303.17760

Luo, Q., Puett, M. J., & Smith, M. D. (2023, May 23). A perspectival mirror of the elephant: Investigating language bias on google, ChatGPT, wikipedia, and YouTube. https://doi.org/10.48550/arXiv.2303.16281

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. https://doi.org/10.1145/3586183.3606763

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., ... Kaplan, J. (2022, December 19). Discovering language model behaviors with model-written evaluations. https://doi.org/10.48550/arXiv.2212.09251

Pezeshkpour, P., & Hruschka, E. (2023, August 22). Large language models sensitivity to the order of options in multiple-choice questions. https://doi.org/10.48550/arXiv.2308.11483

Rudder, C. (2014). *Dataclysm : Who we are when we think no one's looking*. New York : Crown Publishers. Retrieved April 29, 2025, from http://archive.org/details/dataclysmwhowear0000rudd

Salinas, A., Shah, P., Huang, Y., McCormack, R., & Morstatter, F. (2023). The unequal opportunities of large language models: Examining demographic biases in job recommendations by ChatGPT and LLaMA. *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–15. https://doi.org/10.1145/3617694.3623257

Suguri Motoki, F. Y., Pinho Neto, V., & Rodrigues, V. (2023, July 18). More human than human: Measuring ChatGPT political bias. https://doi.org/10.2139/ssrn.4372349

Thongtan, T., & Phienthrakul, T. (2019, January 1). *Sentiment classification using document embeddings trained with cosine similarity* [Pages: 414]. https://doi.org/10.18653/v1/P19-2057

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August 1). Attention is all you need. https://doi.org/10.48550/arXiv.1706.03762

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023, October 19). Voyager: An open-ended embodied agent with large language models. Retrieved November 27, 2023, from http://arxiv.org/abs/2305.16291

Wu, S., & Trottier, D. (2022). Dating apps: A literature review [Publisher: Routledge _eprint: https://doi.org/10.1080/23808985.2022.2069046]. *Annals of the International Communication Association*, *46*(2), 91–115. https://doi.org/10.1080/23808985.2022.2069046

Wu, S., & Ward, J. (2018). The mediation of gay men's lives: A review on gay dating app studies [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/soc4.12560]. *Sociology Compass*, *12*(2), e12560. https://doi.org/10.1111/soc4.12560