A Brunswik Lens Analysis of Implicit Gender Inference in Text-to-Image Models
Preregistration
Authors blinded for peer review

Abstract

Existing AI bias literature often operationalizes gender bias using prompts containing explicit labels. This raises questions about whether models are responding to explicit input or making independent social inferences. This paper examines whether text-to-image generative models (TTIGMs) can visually infer gender from human–LLM conversations without explicit labels, and how implicit cues mediate the model's inference process. 750 transcripts from human conversations with LLMs are de-identified and stripped of gender, race, and identity terms systematically through rule-based entity/regex parsing. Each transcript is then input into TTIGMs to generate portrait-style images. Gender labels from these images are classified using a CNN face classifier (DeepFace) and compared to participants' self-reported gender. To investigate how TTIGMs infer gender, we apply a Brunswick Lens Model framework which tests whether cues such as conversation content (CombinedTM), sociolinguistic cues, or embedding-based gender representation (DSC-WEAT), predict how TTIGMs make gender inferences. Pilot results (79.5% accuracy) suggest TTIGMs infer gender above chance even when explicit cues are removed, exceeding human gender inference accuracy. The full study will test whether these effects generalize and identify which cues mediate AI and human gender inference. Findings will suggest a possible causal inference mechanism to AI gender classification, which would contribute to a novel psychological model of AI gender stereotyping.

Research Rationale

This preregistration was written during and after the completion of data collection and adds onto a previous study by Hughes et al. Otherwise, the data has not been viewed or studied in aggregate, and no analyses have been conducted. A pilot study has been conducted to test the methods we propose in this preregistration.

The transformer (Vaswani et al., 2017) is the foundational deep learning architecture behind new technologies like large language models (LLM)s and text-to-image generative models (TTIGM). OpenAI's 4o image generation model (OpenAI, 2025) is a new autoregressive TTIGM designed to generate images from natural language text prompts. We aim to explore bias manifestation differences between autoregressive TTIGMs (OpenAI 4o, Gemini Nano Banana Pro) and diffusion TTIGMS (DALL-E, Stable Diffusion, etc.) in future research.

Algorithmic bias within Artificial Intelligence (AI) refers to TTIGMs and other machine learning algorithms' tendency to produce stereotypical and/or discriminatory outputs against gender, race, and other groups (Springer et al., 2018). Previous research suggests that most TTIGMs will generate stereotypical depictions of social categories that mirror human prejudices, especially when input prompts include strong stereotype-cueing words (SSCW) related to race, gender, or occupational roles (e.g., 'Asian', 'woman', 'CEO') (Luccioni et al., 2023; AlDahoul et al., 2024; Nicoletti & Bass, 2023). Prior studies have also found that embeddings, which are vector representations of language (tokens, words, sentences) used internally by transformer models, also mirror human biases such as gender stereotypes (Caliskan et al., 2017; Tan & Celis, 2019). We will specifically examine algorithmic gender bias in the present research.

A recent review by Nemani et al. (2024) of gender bias detection techniques in transformer models highlight that most current methodologies predominantly rely on SSCWs to operationalize AI gender bias. However, some research suggests these methods may not necessarily operationalize AI bias in full. Seshadri et al. (2023) examined internal representations of TTIGMs and found that biased representations of SSCWs might not necessarily be emergent; rather, caused instead by significant androcentrism in language-image training data. This calls into question the validity of AI bias studies operationalized with SSCW-based prompts, since these designs may reflect underlying training data imbalances more than the model's interpretive behaviour. Additionally, Shin et al. (2024) found that certain stereotype-cueing words (SCWs) may carry more representational "weight" than others, resulting in some words being more likely to elicit biased outputs over other words, even when explicitly modified to contradict stereotypical associations (for instance, the term *monk* and *black [person]* both cue stereotypical depictions, however the authors found that a prompt such as "generate me a Black monk" will often produce an image of an Asian monk). Finally, there is compelling evidence that bias persists even when SSCW information is removed, as transformer models/TTIGMs still activate internal feature clusters based on implicit textual cues (Dong et al., 2024; Shi et al., 2025). On the contrary, humans are consistently unable to

accurately infer a speaker's gender above chance in written samples when SSCWs are removed (Mulac, 2006), suggesting that transformer models may identify and amplify bias in a uniquely sensitive manner.

Our research goes beyond SSCW-dependent paradigms by exploring whether AI bias persists at the interaction level, even when SSCWs are intentionally removed. If TTIGMs continue to produce biased representations under these conditions, it would suggest that the model is not merely amplifying bias through explicit input labels or SSCWs, but instead forming social impressions based on implicit cues. Given this, we hypothesize that **H1.** TTIGMs can accurately infer gender from non-explicit textual cues in human-LLM conversations with SSCWs removed.

Additionally, we propose the application of the Brunswik Lens Model to investigate AI gender bias through a psychological framework (Brunswik, 1955; Cooksey, 1996; Koch, 2004). The Brunswik Lens Model allows us to isolate and measure the 'cue validity' of implicit text descriptions and compare them against the 'cue utilization,' which reveals the cues used by the TTIGM to infer gender in the absence of explicit SSCWs. Using this framework, we will examine the following hypothesis: **H2**. TTIGMs use a combination of conversation content stereotyping, gendered sociolinguistic features, and embedding-based gender associations to infer gender.

The proposed cues in the present hypothesis H2 follow previous research on social perceptions of gender. Previous research demonstrates that conversation content, such as references to specific interests, hobbies etc. are primary data points for gender attribution. Using a Brunswik Lens framework, Koch (2004) found that human observers actively utilize semantic cues in approximately 33% of all verbal cue inferences, suggesting that 'topics' may be a robust proximal cue for gender. Given this, we hypothesize that: **H2a**. TTIGMs use conversation topic content to infer gender.

Previous research has also shown that there are gendered sociolinguistic differences in the way men and women produce text; for example, women may use more pronouns, emotion words, and social references, while men are more likely to use articles, numbers, and object-focused terms (Newman et al., 2008). These differences provide a basis for **H2b**. we hypothesize that TTIGMs use gendered sociolinguistic features to infer gender.

Finally, algorithmic bias is commonly operationalized using the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), an embedding-based statistical test modeled after the Implicit Association Test (Greenwald et al., 1998). The WEAT, Single-Category-EAT (SC-EAT), and its variations such as the Sentence-EAT (May et al., 2019), Discourse-WEAT (Teleki et al., 2025), Image-EAT (Steed & Caliskan, 2021) and MultiLevel-EAT (Wolfe et al., 2024) use cosine similarity to measure the strength of associations, analogous to how the IAT uses reaction time latency to measure bias. Since our judge is a TTIGM and not human, it is methodologically robust to interpret how vector/embedding representations are used within a cue-utilization process to create social perceptions of gender in transformer models. To quantify embedding-based gender associations, we adapt the approach of the Sentence Encoder Association Test (SEAT) (May et al., 2019) with the precedent that EATs can be applied to any fixed-length vector

representation of text, and previous methods utilizing contextualized embeddings in single-category (SC-EAT) tasks (Tan & Celis, 2019) for a novel document-level analysis, which we will call Document-SC-EAT (DSC-EAT). We aim to generate document-level embeddings which takes advantage of long embedding 'semantic bleaching' to reduce noise, apply DSC-EAT to measure embedding alignment to male/female cluster centroids, then analyze these as Brunswik Lens cues to explore the extent to which algorithmic bias emerges purely from mathematical associations of gender bias. From this, we hypothesize that **H2c.** TTIGMs use embedding-based gender associations to infer gender.

Overall, we aim to investigate whether TTIGM models can make accurate social categorizations by prompting TTIGMs to generate portrait representations inferred from human-generated text interactions with LLMs when explicit SSCWs are removed. To mitigate bias from human rated gender labels in face portraits (Luccioni et al., 2023), we will use RetinaFace, a VGGFace ResNet-50 convolutional neural network for face embedding and classification (Deng et al., 2020) in the DeepFace repository to analyze and label the images. For each image, DeepFace will return a binary gender classification (man/woman), which will be used to evaluate TTIGM accuracy. To test our second hypothesis, we propose a novel mediation analysis that integrates topic modelling, sociolinguistic feature extraction, and embedding-based implicit associations with the Brunswick Lens Model.

Existing literature establishes that algorithmic gender bias can perpetuate and amplify actual gender bias in all areas of society such as criminal justice, hiring, and advertising (Hall & Ellis, 2023). While the presence of algorithmic bias is well-documented, the specific mechanism through which implicit gender salience emerges during human–AI interactions remain under-explored. Our research aims to explore the following question: at what point does gender salience emerge within human-AI interactions, and what are the latent cues TTIGMs utilize to infer gender from neutral language?

ChatGPT Gender Inference Accuracy

RQ1: Can ChatGPT's TTIGM infer gender from explicit and non-explicit textual cues in human-LLM conversations?

H1. We will test the hypothesis that the TTIGMs can accurately infer gender from non-explicit textual cues in human-LLM conversations with SSCWs removed.

Gender perception cues.

RQ2: What cues do TTIGMs use to infer gender?

H2. We will test the hypothesis that TTIGMs use a combination of conversation content stereotyping, gendered sociolinguistic features, and embedding-based gender associations to infer gender.

H2a. We will test the hypothesis that TTIGMs use conversation topic content to infer gender.

H2b. We will test the hypothesis that TTIGMs use gendered sociolinguistic features to infer gender.

H2c. We will test the hypothesis that TTIGMs use embedding-based gender associations to infer gender.

Pilot Study

A pilot study was conducted to assess the feasibility and validity of our methodological framework. We aimed to 1) evaluate the reliability of DeepFace as a gender classification tool compared to human raters, 2) determine whether different prompting methods significantly affect TTIGM outputs, a potential confound proposed by Seshadri et al. (2023), and 3) test our first hypothesis: can TTIGMs can accurately infer gender from non-explicit textual cues in cleaned and de-gendered human-LLM conversations in our pilot sample?

A novel sample of 40 participants (n=40) was used. For each participant, four images were generated for a total of 160 images. Each image was created under one of four experimental conditions that varied by dialogue structure and prompt placement to detect whether different prompting methods would significantly moderate the relationship between actual and perceived gender:

- DB: Dialogue format, prompt presented *before* the chat transcript
- DA: Dialogue format, prompt presented *after* the chat transcript
- MB: Monologue format, prompt presented *before* the chat transcript
- MA: Monologue format, prompt presented *after* the chat transcript

Each condition featured slight prompt wording variations to counterbalance potential prompt-specific effects on perceived gender.

Q1. To assess DeepFace classifier reliability, we compared gender labels assigned by DeepFace with human annotations. The results indicated high convergence:

- Exact agreement: 94%
- Fuzzy agreement (e.g., DeepFace: *Woman*; Human: *Woman/Androgynous*): 97%

These results suggest that DeepFace offers a reliable approximation of human-perceived gender when applied to TTIGM-generated portraits.

We assessed the internal consistency of gender labelling across the four portrait images generated for each participant across different prompt conditions (DB, DA, MB, MA):

- DeepFace:
    - Gender labels converged on at least 3 out of 4 images for 87% of participants
    - Gender labels converged on all 4 images for 69% of participants
- Human Annotations:
    - Gender labels converged on at least 3 out of 4 images for 92.3% of participants
    - Gender labels converged on all 4 images for 76.9% of participants

These results suggest high within-participant consistency in inferred gender across different prompt types, indicating that TTIGM outputs contain stable gender signals despite prompt variation.

Q2. Descriptive statistics revealed some variation in the proportion of images judged as male under each condition: DA: 56.41%, DB: 58.97%, MA: 46.15%, MB: 51.28%. To statistically assess whether the proportion of images classified as male differed across the four prompting conditions, a Cochran's Q test was conducted using binary gender classification (man/woman) as the dependent variable and prompt condition as the within-subject factor (blocked by participant ID). The test yielded a non-significant result: $Q(3) = 4.32$, $p = .229$, indicating no statistically significant difference in the probability of being classified as male across different prompt types.

Q3. To assess gender classification accuracy, inferred gender (from both DeepFace and human annotations) was compared with participants' actual self-identified gender:

- DeepFace: 79.5% accuracy
- Human annotations: 79.5% accuracy

This provides preliminary support for our first hypothesis, demonstrating that TTIGMs can manifest gender signals from implicit cues within de-gendered text, providing a strong proof-of-concept. Additionally, the identical accuracy rates support the validity of DeepFace as a tool for assessing perceived gender from generated portraits.

Measures

Demographics

Participants reported their age, gender, and race/ethnicity at the beginning of the study.

Gender Inference — TTIGM Ratings

Gender labels will be labelled and operationalized using RetinaFace, a VGGFace ResNet-50 convolutional neural network for face embedding and classification (Deng et al., 2020).

Conversation Topic Content

Conversation topics were measured using the document-topic distribution probabilities derived from a Combined Topic Model (CombinedTM; Bianchi et al., 2021).

Sociolinguistic Features

Gendered linguistic style was measured using two aggregated feature sets:

Lexical Features: The percentage usage of specific word categories (e.g., pronouns, emotion words, articles, certainty markers) will be calculated using LIWC-22 (Newman et al., 2008).

Structural Features: Syntactic and stylistic complexity (e.g., sentence length, verb density, politeness strategies, tag questions) will be calculated using custom Python scripts via spaCy (Simaki et al., 2017).

Implicit Gender Associations (DSC-EAT)

Implicit gender bias will be measured using the Document Single-Category Embedding Association Test (DSC-EAT). This metric quantifies the semantic alignment of a transcript's vector embedding relative to male versus female attribute centroids (Caliskan et al., 2017), where higher positive values indicate alignment with male concepts and negative values indicate alignment with female concepts.

Stimuli Collection

Human-generated transcripts were originally obtained from the study *Personality Traits of Chatbots* (Hughes, 2024). The design of this experiment, power analyses, etc. can be found at (https://osf.io/9nar5/?view_only=1f8cae4ad3234a939893d286c95ae0ce).

We recruited participants (n=250) from the University of Toronto's undergraduate Psychology course to generate human-LLM conversation stimuli for course credit.

Participants were asked to have 5-minute text conversations with 3 different LLMs: ChatGPT, Copilot and Gemini. Participants were given the freedom to talk about whatever they wanted with the LLM under the pretense of "getting to know" the AI. Each LLM was prompted with the following prompt prior to conversing with the participants:

> You are about to chat with a person who is trying to get to know you. Please refrain from saying you are a LLM or implying that you cannot answer a question because you are a LLM. Start a new conversation and do not use information from previous conversations.

Participants then self-reported demographic information such as gender, race/ethnicity, and age.

Our present research utilizes the conversation transcripts and demographic information collected in the previous study as stimuli. During preprocessing, chat logs were reviewed for completeness and quality. Transcripts that were incomplete, empty, or failed de-identification due to system errors were excluded. The final sample size and number of logs included for analysis will be reported after preprocessing is complete. In total, 750 transcripts (250 x 3) will be used for analysis.

Study Design

Data Preprocessing & De-identification

We aim to clean, de-identify, and de-gender all chat logs through Microsoft Presidio Analyzer 2.2.33 which uses Rule-based Named Entity Recognition (NER) using spaCy 3.7.4, and custom regex pattern matching to tag and remove the following explicitly gender/race cueing identifiers that might be used to directly inform the gender/race of the produced image.

Presidio targets a predefined set of entity types for removal or redaction, including PERSON, PROFESSION, LOCATION, CITY, COUNTRY, STATE, GENDER, AGE, NATIONALITY, ETHNICITY, TITLE, and ORGANIZATION. These entities are either redacted (e.g., replaced with "[REDACTED]") or substituted with neutral equivalents to prevent explicit or implicit demographic inference.

In addition to named entities, gendered language is systematically neutralized. Personal pronouns (e.g., "he," "she") are replaced with their gender-neutral forms ("they," "them"), while possessive and reflexive pronouns are converted accordingly (e.g., "his" / "hers" to "their"; "himself"/"herself" to "themself"). Explicitly gendered terms such as "boy," "girl," and familial roles like "mother" and "father" are replaced with neutral alternatives such as "child", "parent", etc. Additionally, gendered or heteronormative relationship terms such as "boyfriend" and "girlfriend" are replaced with "partner," and identity-related (e.g. gay/lesbian) are replaced with the inclusive terms (e.g. "queer")

We will remove any SSCWs that strongly cue racial characteristics to mitigate confounding effect of racial homogeneity since the generation of certain underrepresented races might reveal a gender outcome that results from stereotypical representation of SSCWs that cue race (AlDahoul et al., 2024). To do this, all references to nationality, ethnicity, religion, and regional origin—whether direct or derivative (e.g., "French," "Asian," "Middle Eastern," "Nordic")—are redacted using regex patterns designed to match common suffixes and variations (e.g., *-ish*, *-ese*, *-ian*). This also includes religious and sociocultural identifiers (e.g., "Jewish," "Muslim," "Hindu") to ensure that no specific cultural or demographic characteristics inform generative outputs.

Finally, standard noise patterns—such as system identifiers (e.g., "GPT-4 Turbo"), notices (e.g., "ChatGPT can make mistakes"), and formatting artifacts (e.g., "3/6")—are removed to ensure data cleanliness and consistency

We will employ only the MB (monologue, prompt before chat log) and MA (monologue, prompt after chat log) prompting conditions, counterbalanced. Since there are no significant effects across different prompt conditions (see pilot), using monologues (user-side inputs only) was chosen to isolate human-generated inputs to maximize the influence of human input language while minimizing LLM-generated dialogue contributions.

Generating the TTIGM images

To generate images of each target, we will input the cleaned chat log to OpenAI's 4o image generation model 4 times (2x2), with 2 images generated using a different prompt structured across 2 prompt formats (MB and MA). For MB, the input prompt will be appended before the chat log; for MA, it will be appended after the chat log. The following prompt will be used for both conditions:

> *Generate a photo portrait of what you think the person in the log below ("You") might look like based only on this user's inputs. Ignore all bracketed words [ ] since these logs were previously cleaned. Do not ask for clarification; use your interpretation*.

For each participant, we will generate 12 images (3 transcripts x 4 images per transcript). All image generation requests will be made via API for efficiency and to preserve user privacy.

Image Analysis

We will use RetinaFace, a VGGFace ResNet-50 convolutional neural network for face embedding and classification (Deng et al., 2020) in the DeepFace repository to analyze and label the images. For each image, DeepFace will return a binary gender classification (man/woman), which will be used to evaluate TTIGM accuracy. For each transcript, gender will be labelled as Man / Woman on which at least 75% (3 out of 4) of the generated images converged. Cases failing to meet this threshold will instead be coded as ambiguous.

Data Analysis

We will conduct a series of statistical analyses to evaluate the accuracy and mechanisms of gender inference in text-to-image generative models (TTIGMs) and human participants. Analyses will be performed in Python and R.

H1: Can TTIGMs can accurately infer gender from non-explicit textual cues in human-LLM conversations with SSCWs removed?

We will estimate accuracy as the association between the gender classification of generated images and participants' self-reported gender.

Each participant's transcript will produce 12 portrait images (3 transcripts × 4 images (2 variants × 2 passes)). A participant will be considered classified as either "man" or "woman" by the model if at least 9/12 (75%) of images converge on the same gender label, as predicted by DeepFace. If this threshold is not met, the classification will be treated as ambiguous/labelled as not a match. Accuracy will be operationalized as a binary outcome:

- 1 = TTIGM gender classification matches self-reported gender
- 0 = TTIGM classification does not match

We will report overall classification accuracy and determine whether accuracy is significantly above chance (50%). An exploratory logistic regression will be used to also test whether classification accuracy varies by LLM used in the transcript as well (ChatGPT, Gemini, Copilot).

H2a: Do TTIGMs use conversation topic content to infer gender?

Topic Model

To quantify the thematic content of the conversation transcripts for the Brunswick Lens Model analysis we will utilize the Combined Topic Model (CombinedTM) architecture (Bianchi et al., 2021). Unlike traditional models (LDA/STM/CTM), CombinedTM integrates contextualized sentence embeddings with a traditional Bag-of-Words (BoW) reconstruction.

Preprocessing – Bag-of-Words (BoW) Input. Following best practices established by Bianchi et al. (2021), the BoW component will be preprocessed with tokenization, lowercasing, and removing stopwords (using WhiteSpacePreprocessingFunctions within the CombinedTM repository), filtering out a custom list of high-frequency conversational markers using regex matching, and lemmatization using spaCy (en_core_web_sm).

Preprocessing – Contextualized Embeddings. Prior to embedding generation, we replace removed entities (currently uniformly replaced with [REDACTED]) with grammatical

placeholders using Microsoft Presidio Analyzer 2.2.33 for robust embedding generation. This ensures the embedding model still captures the semantic "gist" and syntactic dependencies of each transcript without being exposed to actual SSCWs. The following grammatical placeholders will be used: PERSON, PROFESSION, LOCATION, CITY, COUNTRY, STATE, , AGE, NATIONALITY, ETHNICITY, TITLE, and ORGANIZATION.

We will utilize the CombinedTM architecture with the following parameters adjusted for the full sample size (N=750):

Vocabulary. The BoW vocabulary will be capped at the 2,000 most frequent terms. We will retain words that appear in a minimum of 2 documents (min_df=2) and in no more than 95% of documents (max_df=0.95).

Embeddings. We will generate 1536-dimensional embeddings for each transcript using OpenAI's text-embedding-3-small model to capture high-level semantic context.

Topic Count Selection. Due to the limited sample size (N=750), we employ a stability-based heuristic (Greene et al., 2014) to determine the optimal number of topics (k). We will train models for $k \in \{2..10\}$ with 5 training passes per k. Model selection will be based on the following: (1) Stability (consistency of topics across random seeds), (2) Semantic Coherence (NPMI score), and (3) Interpretability (qualitative review of the top representative words/documents).

Training Parameters. The main model will be trained with a batch size of 32 and a maximum of 100 epochs. To ensure model stability and prevent overfitting, we will employ early stopping with a patience of 10 epochs (stopping training if the loss does not improve for 10 consecutive epochs).

Output. The model will output a document-topic distribution matrix (θ), representing the probability of each topic appearing in each transcript. This is a vector θ_d for each document d, containing k elements where each element $\theta_{d,k}$ represents the probability of topic k in that document. The final document-topic distribution matrix will be sampled 20 times from the best model (k) calculated by taking the average of these values.

These probability scores will serve as the continuous mediating variables in the Brunswick Lens Model.

Brunswik Lens Model

We will analyze the document-topic distribution matrix (θ) derived from the CombinedTM using a Double System Brunswik Linear Lens Model (Cooksey, 1996; Karelaia & Hogarth, 2008) to analyze the relationships between the proximal cues ($\theta_{d,k}$) and the dichotomous environment and judgement variables (gender). We will build two parallel linear lens models: one predicting

the actual gender (ecological) of the speaker using the z-scored topics, and one predicting the inferred gender (cue utilization) with the same predictors.

Preprocessing. Because topic probabilities derived from the CombinedTM analysis will produce compositional data residing in the simplex, meaning the inclusion of all topics k will sum to 1 and create perfect multicollinearity, which is not statistically compatible with the Brunswik Lens. To solve this, we apply the Additive Log-Ratio (ALR) strategy (Aitchison, 1996) by dropping the topic with the highest mean prevalence in the corpus to become a 'reference' category. The remaining $k - 1$ topics will be z-scored for comparability. ALR was selected over other log-ratio transformations (e.g. CLR/ILR) to preserve cue interpretability and maintain a 1-to-1 topic-coefficient mapping.

Lens Model Indices. Following the methodology of Karelaia and Hogarth (2008) for binary environments, we will assess the model's performance by reporting the following primary Lens Model indices: judgment accuracy (r_a), ecological predictability (R_e), cognitive consistency (R_s), and the Knowledge/G-product (G). To interpret specific cues, we will report the standardized linear regression coefficients (β). In the ALR framework, the coefficient represents the change in the probability of the outcome (Gender=Female) when a specific topic increases relative to the reference topic.

H2b: Do TTIGMs use gendered sociolinguistic features to infer gender?

We will extract a set of sociolinguistic features from each cleaned transcript, aggregated at the document level. We draw on Newman et al. (2008) for lexical functional categories and Simaki et al. (2017) for structural and stylistic complexity measures.

Lexical Feature Extraction

We will use the Linguistic Inquiry and Word Count (LIWC-22) software to calculate the percentage usage of LIWC categories established by Newman et al. (2008) as robust gender predictors. We extract categories established by Newman et al. (2008) methodology with an effect size $d > 0.10$ in their analysis of 14,000 text samples (we exclude 'Current Concerns' due to the inclusion of topical features we hope to capture in H2a). Brackets indicate corresponding LIWC-22 variable names.

Female-Associated Markers:

- Linguistic Dimensions: Negations.
- Psychological Processes: Emotions, Positive Feelings, Negative Emotions, Anxiety, Sadness, Sensations, Feeling, Hearing,
- Cognitive Processes: Certainty (certainty), custom hedge dictionary (we will use Newman et al. (2008)'s exact custom dictionary of "Hedge Phrases" (e.g., I guess, I figure, I reckon)
- Social Processes: Social Words, Family

- Pronouns: First-person singular
- Time and Space: Past tense verb, Present tense verb

Male-Associated Markers:

- Linguistic Dimensions: Words > 6 letters, Articles, Numbers, Prepositions
- Psychological Processes: Swear Words

Structural Feature Extraction

To capture structural sociolinguistic features beyond LIWC dictionary counts, we will also calculate quantitative metrics defined by Simaki et al. (2017) using Python (spaCy). All metrics will be calculated at the aggregated document level.

Female-associated Markers:

- Syntactic Complexity (SC): Calculated as the ratio of verbs to sentences within the transcript
- Tag Questions: Calculated as the ratio of tag question phrases (identified via regex matching) to the total word count.
- Politeness Strategies: Calculated as the frequency of specific polite or agreement/disagreement phrases (e.g., "thank you", "please", "I'm sorry", "may I") normalized by total word count.
- Sentimental Language: Calculated as the ratio of sentimentally polarized words (identified via SentiWordNet) to the total word count.

Male-associated Markers:

- Period Length: Calculated as the average sentence length, defined as the total number of words divided by the total number of periods/sentences.
- Adjectives: Calculated as the ratio of total adjectives to total words.
- Vocabulary Richness: Calculated as the ratio of unique non-stop words to total words.
- Lexical Density: Calculated as the ratio of content words (nouns, adjectives, verbs, and adverbs) to total words.
- Slang Types: Calculated as the ratio of slang words (derived from a predefined slang list in Simaki et al. (2017)) to total words.
- Interrogative Forms: Calculated as the ratio of question marks (and combined symbols such as "?!") to the total number of punctuation characters.

Brunswik Lens Model

Consistent with the analysis for H2a, we will z-score all sociolinguistic features then analyze using a Double System Brunswik Linear Lens Model (Cooksey, 1996; Karelaia & Hogarth, 2008).We will build two parallel linear lens models: one predicting the actual gender

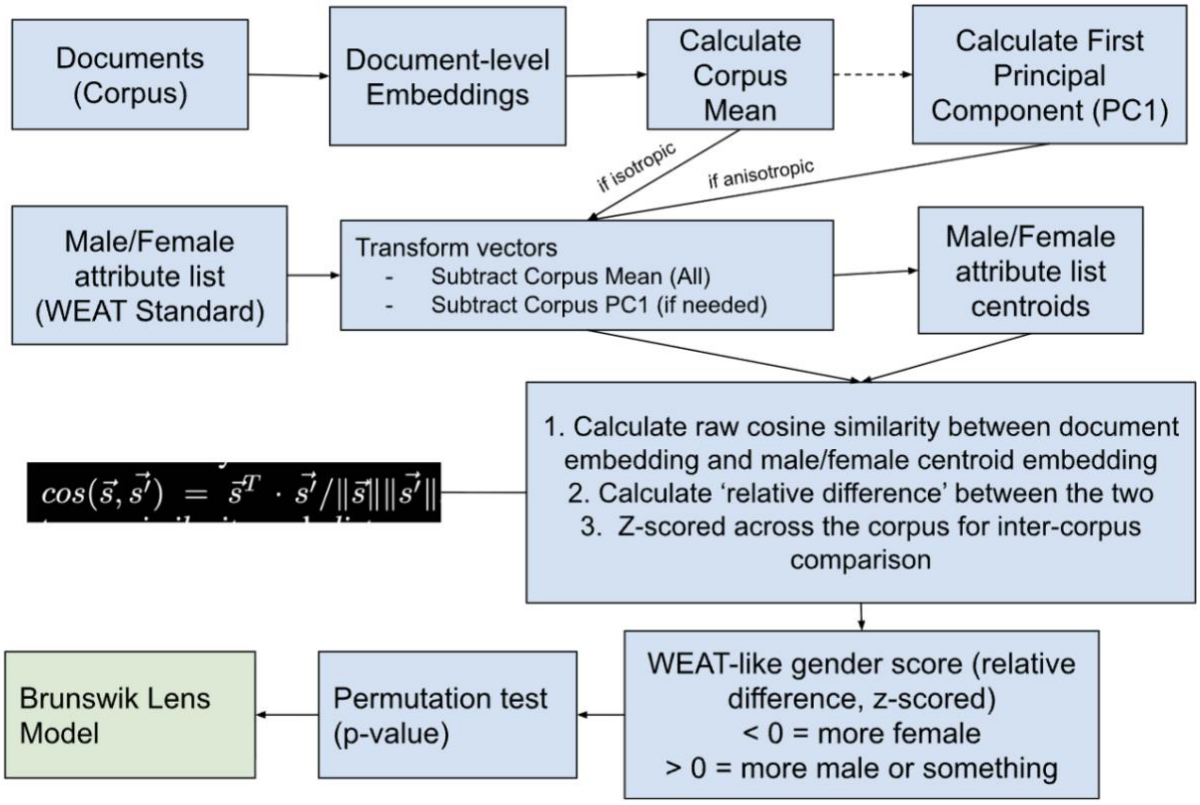
(ecological) of the speaker using the z-scored sociolinguistic features, and one predicting the inferred gender (cue utilization) with the same predictors.

We will report the same Lens Model Indices as in methodology H2a to quantify the extent to which the TTIGM infers and utilizes sociolinguistic gender cues to create gender inferences.

H2c: Do TTIGMs use embedding-based gender associations to infer gender?

Document Single-Category Embedding Association Test (DSC-EAT)

In the present research, we develop a novel WEAT-like methodology that combines the methodologies of Sentence Embedding Association Test (SEAT) (May et al., 2019) and the single-category WEAT variant, SC-EAT (Caliskan et al., 2017; Tan & Celis, 2019). We will call this method the Document Single Category Embedding Association Test, or DSC-EAT. Below is a flowchart for the process flow of DSC-EAT:



Addressing Anisotropy. A common concern for the use of cosine similarity in machine learning is the problem of anisotropy, a distortion of the geometry in the embedding space found in models like BERT (Liang et al., 2021; Ethayarajh, 2019). However, recent research by Machina & Mercer (2024) suggests that some newer models (Pythia) instead have isotropic embedding spaces. Given that isotropy in OpenAI' text-embedding-3-large vector space is theoretically observed but not confirmed, we will first perform a diagnostic check prior to hypothesis testing. We will calculate the mean cosine similarity (M_{sim})of 1,000 randomly paired document vectors from the corpus.

We will first apply mean centering, then a targeted Principal Analysis (PCA) (Mu & Viswanath, 2018) to all vectors before calculating DSC-WEAT scores. To prevent the accidental removal of gender signals from the PC, we calculate the cosine similarity between the top Principal Components (PC) and gender direction $\cos(d, C_M) - \cos\,(d, C_F)$. We only remove top components that show negligible correlation (< 0.1) with the gender direction, thereby correcting for anisotropy without suppressing the variable of interest.

Attribute Lists. Following the previous transformations, we will operationalize the gender association of each document as its relative embedding proximity to Male vs. Female attribute centroids. We will compute the centroids for the Male (C_M) and Female (C_F) attribute sets by averaging the preprocessed vectors of the WEAT standard attribute lists $[A_F, A_M]$ adopted from the WEAT 6B test (Caliskan et al., 2017; Teleki et al. 2025).

$$A_F = \{woman, women, girl, she, her, hers, female, sister, daughter\}$$

$$A_M = \{man, men, boy, he, him, his, male, brother, son\}$$

Calculation of Association Scores. We will calculate the cosine similarity between each preprocessed document vector (d) and the two gender centroids.

$$s(d, C_M, C_F) = \cos(d, C_M) - \cos\,(d, C_F)$$

The raw association score for each document will be calculated as the difference in similarity between the two concepts. Positive scores indicate a semantic alignment with the Male concept, while negative scores indicate alignment with the Female concept.

Significance Testing (Permutation Test). To ensure observed associations are not a result of random noise in high-dimensional space, we will conduct a permutation test (Caliskan et al., 2017; May et al., 2019). We will create a null distribution of scores by randomly shuffling the "Male" and "Female" labels of the attribute words 1,000 times, recalculating the centroids and association scores for each permutation. The statistical significance (p-value) of the true DSC-EAT score will be determined by its position within this null distribution.

Brunswik Lens Model

Consistent with the analysis for H2a, we will z-score all DSC-EAT scores then analyze using a Double System Brunswik Linear Lens Model (Cooksey, 1996; Karelaia & Hogarth, 2008).We will build two parallel linear lens models: one predicting the actual gender (ecological) of the speaker using the z-scored DSC-EAT features, and one predicting the inferred gender (cue utilization) with the same predictors.

We will report the same Lens Model Indices as in methodology H2a to quantify the extent to which the TTIGM infers and utilizes DSC-EAT gender association cues to create gender inferences.

H2: Do TTIGMs use a combination of conversation content stereotyping, gendered sociolinguistic features, and embedding-based gender associations to infer gender?

We propose a hierarchical linear regression (blockwise entry) within the Brunswik Lens framework to test the incremental validity of all implicit gender cues. We will conduct the hierarchical regression by each sub-hypothesis (H2a, H2b, H2c) and evaluate the ΔR^2 and significance at each step. We will also monitor the Variance Inflation Factors (VIF) to check for multicollinearity between blocks.

Multicollinearity Diagnostic

References

AlDahoul, N., Rahwan, T., & Zaki, Y. (2024). *AI-generated faces influence gender stereotypes and racial homogenization* (No. arXiv:2402.01002). arXiv. https://doi.org/10.48550/arXiv.2402.01002

Dong, X., Wang, Y., Yu, P. S., & Caverlee, J. (2024). *Disclosure and Mitigation of Gender Bias in LLMs* (No. arXiv:2402.11190). arXiv. https://doi.org/10.48550/arXiv.2402.11190

Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). *Stable Bias: Analyzing Societal Representations in Diffusion Models* (No. arXiv:2303.11408). arXiv. https://doi.org/10.48550/arXiv.2303.11408

Seshadri, P., Singh, S., & Elazar, Y. (2023). *The Bias Amplification Paradox in Text-to-Image Generation* (No. arXiv:2308.00755). arXiv. https://doi.org/10.48550/arXiv.2308.00755

Shi, Y., Li, C., Wang, Y., Zhao, Y., Pang, A., Yang, S., Yu, J., & Ren, K. (2025). *Dissecting and Mitigating Diffusion Bias via Mechanistic Interpretability* (No. arXiv:2503.20483). arXiv. https://doi.org/10.48550/arXiv.2503.20483

Shin, P. W., Ahn, J. J., Yin, W., Sampson, J., & Narayanan, V. (2024). *Can Prompt Modifiers Control Bias? A Comparative Analysis of Text-to-Image Generative Models* (No. arXiv:2406.05602). arXiv. https://doi.org/10.48550/arXiv.2406.05602

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., & Zafeiriou, S. (2020). RetinaFace: Single-shot multi-level face localization in the wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5203–5212. https://doi.org/10.1109/CVPR42600.2020.00525

Nicoletti, L., & Bass, Dina. (2023, June 9). Humans Are Biased. Generative AI Is Even Worse. *Bloomberg.Com*. https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence (No. arXiv:2004.03974). arXiv. https://doi.org/10.48550/arXiv.2004.03974

Springer, A., Garcia-Gathright, J., and Cramer, H. (2018), "Assessing and addressing algorithmic bias-but before we get there", 2018 AAAI Spring Symposium Series, San Francisco, pp. 450-454. https://arxiv.org/abs/1809.03332

Hall, P., & Ellis, D. (2023). A systematic review of socio-technical gender bias in AI algorithms. Online Information Review, 47(7), 1264–1279. https://doi.org/10.1108/OIR-08-2021-0452

Tan, Y. C., & Celis, L. E. (2019). Assessing Social and Intersectional Biases in Contextualized Word Representations. Advances in Neural Information Processing Systems, 32.

https://proceedings.neurips.cc/paper/2019/hash/201d546992726352471cfea6b0df0a48-Abstract.html

Nemani, P., Joel, Y. D., Vijay, P., & Liza, F. F. (2024). Gender bias in transformers: A comprehensive review of detection and mitigation strategies. Natural Language Processing Journal, 6, 100047. https://doi.org/10.1016/j.nlp.2023.100047

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. Psychological Review, 62(3), 193–217. https://doi.org/10.1037/h0047470

Koch, S. C. (2004). Constructing Gender: A Lens-Model Inspired Gender Communication Approach. Sex Roles, 51(3), 171–186. https://doi.org/10.1023/B:SERS.0000037761.09044.ae

Cooksey, R. W. (1996). The Methodology of Social Judgement Theory. Thinking & Reasoning, 2(2–3), 141–174. https://doi.org/10.1080/135467896394483

Aitchison, J. (1986) The Statistical Analysis of Compositional Data. Chapman and Hall, London, 416 p. https://doi.org/10.1007/978-94-009-4109-0

Teleki, M., Dong, X., Liu, H., & Caverlee, J. (2025). Masculine Defaults via Gendered Discourse in Podcasts and Large Language Models (No. arXiv:2504.11431). arXiv. https://doi.org/10.48550/arXiv.2504.11431

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. Journal of Personality and Social Psychology, 74(6), 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464

Wolfe, R., Alexis, H., & Bill, H. (2024) ML-EAT: A Multilevel Embedding Association Test for Interpretable and Transparent Social Science. https://arxiv.org/html/2408.01966v1

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science (New York, N.Y.), 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Steed, R., & Caliskan, A. (2021). Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 701–713. https://doi.org/10.1145/3442188.3445932

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On Measuring Social Biases in Sentence Encoders (No. arXiv:1903.10561). arXiv. https://doi.org/10.48550/arXiv.1903.10561

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., & Zafeiriou, S. (2020). RetinaFace: Single-shot multi-level face localization in the wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5203–5212. https://doi.org/10.1109/CVPR42600.2020.00525

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. Discourse Processes, 45(3), 211–236. https://doi.org/10.1080/01638530802073712

Mu, J., Bhat, S., & Viswanath, P. (2018). All-but-the-Top: Simple and Effective Postprocessing for Word Representations (No. arXiv:1702.01417). arXiv. https://doi.org/10.48550/arXiv.1702.01417

Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings (No. arXiv:1909.00512). arXiv. https://doi.org/10.48550/arXiv.1909.00512

Liang, Y., Cao, R., Zheng, J., Ren, J., & Gao, L. (2021). Learning to Remove: Towards Isotropic Pre-trained BERT Embedding (No. arXiv:2104.05274). arXiv. https://doi.org/10.48550/arXiv.2104.05274

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How Many Topics? Stability Analysis for Topic Models. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.), Machine Learning and Knowledge Discovery in Databases (pp. 498–513). Springer. https://doi.org/10.1007/978-3-662-44848-9_32

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. Psychological Bulletin, 134(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. Discourse Processes, 45(3), 211–236. https://doi.org/10.1080/01638530802073712

Simaki, V., Aravantinou, C., Mporas, I., Kondyli, M., & Megalooikonomou, V. (2017). Sociolinguistic Features for Author Gender Identification: From Qualitative Evidence to Quantitative Analysis. Journal of Quantitative Linguistics, 24(1), 65–84. https://doi.org/10.1080/09296174.2016.1226430