Gender Bias in AI-Generated Texts: A Comparative Perspective between English and Chinese Language Models

Xiaojun Liang Shengzi Huang

xiaojun.liang@studio.unibo.it
shengzi.huang@studio.unibo.it

January, 2026

1 Abstract

Gender bias in AI-generated texts has become a growing concern with the widespread use of large language models (LLMs). Most existing studies focus on English and emphasize surface-level outputs, leaving cross-linguistic differences underexplored. This review provides a systematic comparison of gender bias in English and Chinese language models.

Based on a structured analysis of recent literature, we examine how gender bias is detected, how it manifests at surface, representation, and behavioral levels, and how linguistic and cultural characteristics shape its expression across languages. The findings show that English-language models tend to exhibit bias through explicit stereotypes and evaluative framing, while Chinese-language models more often encode bias implicitly through default gender inference, omission, and asymmetric safety behaviors.

We further discuss the implications of these patterns for fairness, inclusiveness, and responsible AI development, arguing that technical performance alone is insufficient to ensure socially responsible systems. Overall, this review highlights the need for language-aware and context-sensitive evaluation frameworks to support fair and inclusive multilingual AI systems.

Keywords: Artificial intelligence; Gender bias; Language models; Cross-lingual evaluation

2 Introduction

Artificial intelligence has become an integral component of modern society, with AI-generated text playing a central role across a wide range of application domains. The development of text generation systems has evolved from early rule-based approaches, such as ELIZA, to statistical models and neural architectures, including recurrent neural networks (RNNs). With continued advances in neural modeling and, more recently, the emergence of large language models (LLMs), contemporary systems are capable of producing fluent and contextually coherent text that increasingly resembles human writing [1]. Despite these advances, the rapid proliferation of AI-generated text has raised substantial concerns regarding algorithmic bias [2]. Among the various forms of social bias that may be propagated by language models, gender bias has received particular attention due to its subtle yet pervasive nature. It refers to systematic and unfair differences in how gender identities are represented, evaluated, or associated within model behaviors and outputs [3]. Such patterns may unintentionally reinforce stereotypical or unequal social roles, shaping perceptions of fairness and inclusiveness and posing challenges to the development of trustworthy AI systems.

Importantly, gender is expressed and encoded differently across languages. In English, gender distinctions are often marked explicitly through pronouns or occupational terms, whereas Chinese relies more heavily on pronouns and characters that carry gendered semantic meanings. These structural differences can influence how language models learn and reproduce gender-related patterns. Cross-lingual studies provide empirical support for this observation, showing that even under matched prompting or probing procedures, models generate systematically different gendered descriptions across languages [4]. Together, these findings indicate that linguistic structure and sociocultural context play a critical role in shaping how gender bias emerges, underscoring the importance of cross-language investigation.

Existing research has identified diverse manifestations of gender bias in large language models. Prior studies report imbalanced gender representations in text generation, systematic stereotyping linking gender with specific traits or professions, and interaction bias, whereby equivalent prompts yield different responses due to gender-related cues. Beyond single-axis analyses, intersectional work further shows that gender bias often co-occurs with race, culture, and other identity attributes. At a deeper level, gendered associations have also been observed within internal model representations, including embedding spaces and multilingual alignment, indicating that bias extends beyond surface outputs.

While these findings reveal multiple forms of gender bias, existing evaluation practices remain largely centered on explicit and directly observable behaviors. Increasing evidence suggests that gender bias in LLMs can instead be examined across multiple analytical layers, including explicit outputs, evaluative preference patterns, and implicit semantic associations. Importantly, these layers do not necessarily align, as seemingly neutral outputs may coexist with biased latent tendencies [5]. This multi-level perspective therefore motivates the need for integrated evaluation frameworks capable of assessing bias beyond surface-level performance alone [6].

Accordingly, this review aims to provide a systematic synthesis of existing research on gender bias in AI-generated texts from a cross-linguistic perspective, with a particular focus on comparisons between English and Chinese language models. Specifically, the study examines how gender bias is identified and evaluated across different linguistic settings, highlighting similarities as well as language-dependent divergences in bias manifestation and measurement practices. By organizing prior work across multiple analytical levels and evaluation approaches, this review seeks to clarify the methodological assumptions underlying commonly used bias detection metrics and to assess their applicability across languages. Through this analysis, the paper emphasizes the importance of language-aware and context-sensitive evaluation practices, and outlines key challenges and future directions for more reliable and inclusive assessments of fairness in multilingual AI systems.

3 Method

3.1 Premises

In everyday use, gender is usually understood as a social and cultural difference between men and women. Although traditional dictionaries do not define gender bias as a separate term, it has become an important concept in academic research, especially in studies of language, education, and artificial intelligence. From a sociolinguistic perspective, gender bias comes from social systems that assign different roles and expectations to different gender groups. These patterns are reflected in language and are gradually reinforced through daily communication.

When artificial intelligence systems are trained on large amounts of human-written text, they learn these language patterns as statistical regularities. As a result, natural language processing models may unintentionally absorb gender stereotypes from training data and reproduce them in tasks such as text generation or prediction. From a technical point of view, bias in the data can be encoded into internal model representations, such as word embeddings and contextual features, and later influence the model’s observable behavior. This explains why gender bias in AI-generated text can be seen as a continuation of existing social and linguistic inequalities.

However, gender bias does not appear in the same way across different languages. Most previous studies focus mainly on English data and evaluation methods, which limits how well the findings apply to other languages. English and Chinese differ greatly in both linguistic structure and cultural background. For example, the Chinese pronoun system (他 / 她 / ta) has historically used the male form as the default, which is different from the English system of he, she, and they. In addition, culture-specific gender stereotypes related to occupations or personal traits may also affect how bias is learned and expressed. Despite these differences, systematic comparisons between English and Chinese remain limited. Based on these premises, this review aims to summarize and compare existing studies in both languages in order to better understand similarities and differences in the origins, expressions, and evaluation of gender bias in AI-generated text.

3.1.1 Model Classification

In this review, English models are defined as large language models that are primarily trained and evaluated in English. These models are typically built on corpora dominated by English-language texts and are assessed using English prompts and outputs. In contrast, Chinese models refer to models mainly trained on Chinese corpora and optimized for Chinese linguistic and cultural contexts, with evaluation commonly conducted through Chinese-language prompts and generated texts. This distinction provides a basis for examining how different linguistic environments influence the manifestation of gender bias in AI-generated text.

For bilingual or multilingual models that incorporate both English and Chinese data during training, classification is determined by the language context of the task rather than by the model’ s institutional origin. When a model is evaluated using English prompts and produces English outputs, it is examined as an English model; when the same model is applied to Chinese-language tasks, it is treated as a Chinese model. This task-based classification strategy helps maintain analytical consistency and avoids conflating a model’ s technical background with the linguistic setting of the study, thereby supporting clearer cross-linguistic comparison of gender bias.

3.1.2 Keyword Source and Treatment

When author-defined keywords are available, they are adopted as the primary source of analysis and are used without further processing. For papers without explicit keywords, derived keywords are constructed exclusively from abstracts. For these documents, candidate key-phrases are automatically extracted using the YAKE algorithm, a statistical keyword extraction method based on term frequency, positional features, and contextual information.

In this implementation, multi-word key-phrases of up to three tokens are extracted, and the top fifteen candidates are retained for each document lacking author-defined keywords. The extracted keywords are subsequently refined through manual review following predefined criteria. Specifically, evaluative or descriptive expressions are removed, while conceptually informative terms that denote research topics, methods, datasets, or phenomena are retained. During this refinement stage, keyword selection is guided by their conceptual roles rather than surface lexical forms, with attention to maintaining a balanced representation of different analytical dimensions (e.g. technical focus, research problem, and evaluation resource) within each document. Lexical variants and acronym forms referring to the same underlying concept are consolidated to improve conceptual consistency. The same extraction configuration was applied uniformly across all applicable documents to ensure methodological consistency.

Although this procedure involves a degree of interpretative judgment, it enables documents lacking author-defined keywords to be incorporated into a unified analytical framework. The subsequent keyword-based analysis is exploratory in nature and aims to identify conceptual structures rather than to assess statistical significance or research impact.

3.2 Research questions

To organize this review and support its main goal of understanding gender bias in English and Chinese language models, this study proposes one general research question and four sub-questions. The general question defines the overall scope of the review. The four sub-questions are designed to follow a logical progression, guiding the analysis from bias detection methods to bias manifestations, cross-linguistic influences, and broader implications for responsible AI.

G: What similarities and differences exist in gender bias between English and Chinese language models?

Q1: What methodologies and metrics are used in detecting and measuring gender bias in AI-generated texts?

Q2: What forms of gender bias are present in AI-generated texts produced by language models?

Q3: How do linguistic and cultural characteristics of English and Chinese contribute to the differences in gender bias observed in AI-generated texts?

Q4: What are the implications of gender bias in English and Chinese AI-generated texts for fairness, inclusiveness, and responsible AI development?

3.3 Sources

In particular, the literature considered in this review is selected based on two main criteria: recency and disciplinary relevance. Most of the included studies were published after 2020, reflecting the rapid technological advancements in recent years. In terms of disciplinary scope, the review primarily draws on research from computer science, while also incorporating relevant contributions from linguistics and psychology.

In detail, the sources used for intensive search (via keyword-based queries) are the following:

- Scopus—https://www.scopus.com
- ScienceDirect—http://www.sciencedirect.com
- IEEE Xplore—http://ieeexplore.ieee.org
- arXiv—https://arxiv.org
- Google Scholar—https://scholar.google.com/
- UNESCO—https://www.unesco.org/en

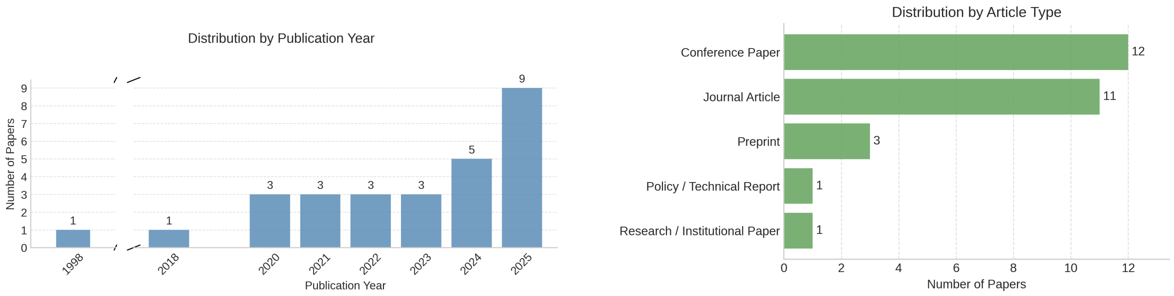
Each of these sources is queried with the following combinations of keywords:

- gender bias AND text generation
- gender bias AND English AND Chinese
- gender bias AND large language models
- gender bias AND AI-generated texts
- gender bias AND LLMs

To ensure comprehensive coverage, additional keyword combinations reflecting variations in terminology across model types, linguistic settings, and gender-related bias concepts were also included. The full list of keyword combinations is reported in Appendix A.

3.4 Papers classification

The papers selected are classified according to their publication year and article type. Two figures are used to illustrate the overall distribution of the reviewed literature.



(a) Distribution by publication year

(b) Distribution by article type

Figure 1: Overview of the selected papers in terms of publication year and article type.

Figure 1a shows the number of papers published in different years. Only a small number of studies were published before 2020. These early works mainly focused on theoretical or social aspects of gender bias. After 2021, the number of publications increased clearly. Most of the selected papers were published between 2023 and 2025. This trend shows that gender bias in large language models has become an important research topic in recent years.

Figure 1b presents the distribution of papers by article type. Most of the selected studies are conference papers and journal articles. Conference papers are mainly published in NLP venues such as ACL, EMNLP, and NAACL. Journal articles usually provide a more detailed analysis and discussion. A small number of papers are preprints or institutional reports. These works help reflect new research trends and social concerns.

4 Results of the Review

4.1 English Language Model

4.1.1 Methodologies and Metrics for Detecting Gender Bias in AI-Generated Texts

In studies of gender bias in AI-generated texts, metrics and methodologies play complementary but distinct roles. Specifically, existing studies can be broadly categorized into (i) **bias quantification metrics**, which formally define and measure gender bias, and (ii) **bias analysis methodologies**, which investigate gender bias at different analytical levels.

The latter category further includes:

- **Statistical-level analysis**, focusing on validating whether observed gender differences are statistically significant.
- **Corpus-level analysis**, aimed at identifying potential sources of gender bias in training data.
- **Prediction-level analysis**, examining how gender bias manifests in model outputs.

To maintain the readability of the main text, detailed mathematical formulations of the metrics and methodologies

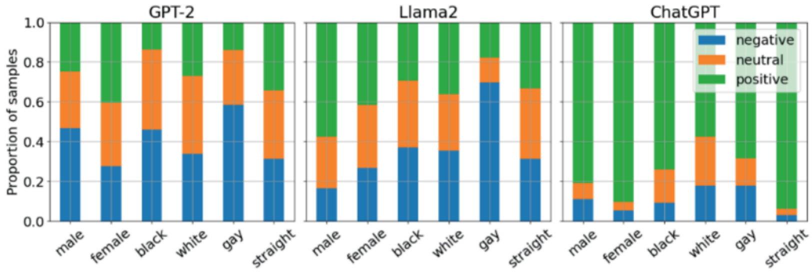


Figure 2: Sentiment distribution for bias evaluation in language models. Reproduced from [13].

are provided in Appendix B.

4.1.1.1 Metrics for Quantifying Gender Bias A variety of quantitative metrics have been proposed to measure gender bias in AI-generated texts. These metrics provide standardized means to compare how different models handle gender-related expressions and to identify systematic disparities between gender groups. Existing metrics can be broadly grouped into probability-based measures, fairness-oriented indicators, and embedding-based association tests.

- **GenBiT Score** [7] is a probability-based metric designed to quantify gender bias through lexical association asymmetry. It measures the strength of association between gender-related terms and other words in a corpus by comparing conditional probabilities across gender contexts. In the original GenBiT work, a logarithmic transformation is applied to the ratio-based formulation to improve numerical stability and interpretability.
  In the study [8], GenBiT was adopted to evaluate gender bias across multiple large language models, revealing consistent variations in bias magnitude among models of different architectures and parameter scales.
- **Average Difference in Prediction Intensity (avg. Δ)** [9] measures gender bias by comparing model-predicted emotion intensity scores for gender-swapped sentence pairs.
- **Selection Rate Ratio (SRR)**, a ratio-based criterion commonly used to operationalize demographic parity [10], evaluates whether predicted outcomes are distributed proportionally across gender groups.
- **Average Confidence Score (ACS)**, a heuristic confidence-based metric proposed in prior work [11] to characterize relative differences in prediction intensity across social groups.
- **Word Embedding Association Test (WEAT)** [12] quantifies gender bias by measuring differential associations between gender target word sets and stereotypical attribute sets using cosine similarity.
- **Sentence Embedding Association Test (SEAT)** [12] is an extension of WEAT to contextualized embeddings, measuring gender bias at the sentence level using templated sentences.
- **Sentiment Distribution**, evaluates gender bias by comparing the proportion of positive, neutral, and negative continuations generated for male- and female- referenced prompts. A higher share of negative outputs for women indicates a systematic sentiment imbalance. An illustrative example as shown in 2

4.1.1.2 Statistical-level Methods Beyond descriptive metrics, statistical-level methods are employed to assess whether observed gender bias differences across models are statistically reliable. Such methods focus on validating whether measured disparities reflect systematic effects rather than random variation, thereby strengthening the credibility of comparative evaluations. Common approaches include variance analysis and statistical

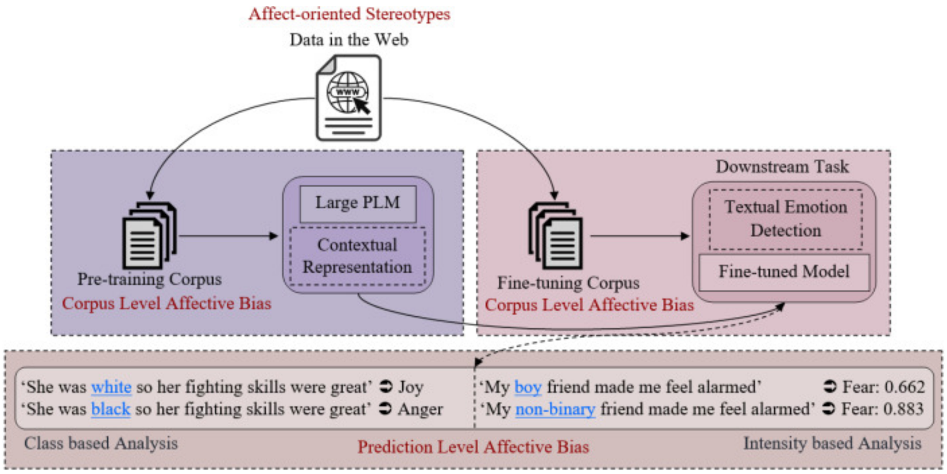

Affect-oriented Stereotypes
Data in the Web
Pre-training Corpus
Corpus Level Affective Bias
Large PLM
Contextual Representation
Downstream Task
Textual Emotion Detection
Fine-tuned Model
Fine-tuning Corpus
Corpus Level Affective Bias
'She was white so her fighting skills were great' ➲ Joy
'She was black so her fighting skills were great' ➲ Anger
'My boy friend made me feel alarmed' ➲ Fear: 0.662
'My non-binary friend made me feel alarmed' ➲ Fear: 0.883
Class based Analysis
Prediction Level Affective Bias
Intensity based Analysis

Figure 3: Corpus-level and prediction-level affective bias in large language models [11].

significance testing.

- **One-way ANOVA** [8] is commonly used to compare GenBiT or other bias scores across multiple models, testing whether observed differences represent statistically significant variation rather than random noise.
- **Two-way ANOVA** may further be applied to examine both main effects and interaction effects among experimental factors (e.g. model type, narrative structure, or protagonist gender), enabling analysis of whether the magnitude or direction of gender bias varies across conditions.
- Statistical significance is typically assessed using p**-values** [14]. Following standard statistical practice, a p-value is computed to evaluate whether observed differences in prediction scores between gender groups are statistically meaningful. A smaller value indicates stronger evidence against the null hypothesis of no difference between the groups.

4.1.1.3 Corpus-level Methodologies Corpus-level methodologies focus on identifying gender bias within the datasets used to train or fine-tune language models. These analyses aim to determine whether uneven gender associations are already embedded in the data before model training. Typical approaches examine the occurrence and co-occurrence patterns between gender-related terms and affective expressions, such as anger, fear, joy, and sadness, to reveal imbalances in emotional associations across gender groups [11].

4.1.1.4 Prediction-level Methodologies Prediction-level methodologies investigate how gender bias emerges in model outputs during inference. A commonly adopted approach involves gender-swapped sentence pairs that differ only in gender-related terms. Differences in predicted emotion categories or intensity scores across paired inputs are interpreted as evidence of gender-based disparities in model behavior. Corpus-level and prediction-level analyses can be conducted jointly, as they provide complementary perspectives on the origins and manifestations of gender bias in AI-generated texts [11], as illustrated in Fig. 3.

4.1.2 Forms of Gender Bias in AI-Generated Texts

To ensure cross-linguistic comparability, this study distinguishes between levels of bias diagnosis and forms of bias manifestation when analyzing gender bias in both English and Chinese language models. Specifically, gender bias is categorized into three manifestation-based forms:

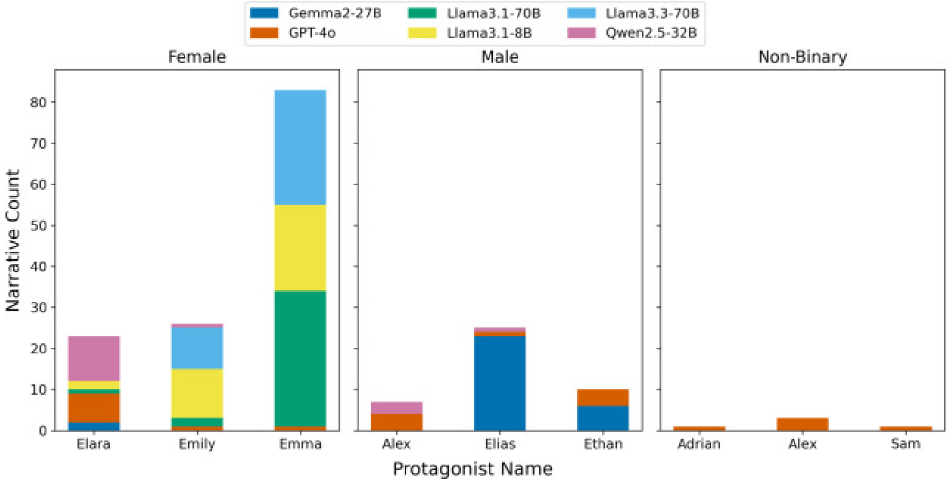


Figure 4: Protagonist names by Vanblerck *et al.* [8] across different large language models.

- ***Surface-level bias***, reflected in observable output patterns
- ***Representation-level bias***, encoded in internal gender semantic associations
- ***Behavioral bias***, manifested through systematic generative and interactional tendencies

4.1.2.1 Surface-Level Bias

Representational imbalance. One common form is representational imbalance, referring to unequal representation of different gender groups in leading character roles. Prior work has shown that certain models exhibit strong gender preferences in protagonist generation [8]. Consistent with these findings, our results reveal substantial disparities in the gender distribution of protagonists across different LLMs, as illustrated in Fig. 4.

Specifically, Gemma2-27B demonstrates a pronounced preference for male protagonists, while the LLaMA models, particularly Llama3.1-70B and Llama3.3-70B, overwhelmingly favor female protagonists, with little to no representation of male or non-binary characters. In contrast, GPT-4o and Qwen2.5-32B exhibit comparatively more balanced gender distributions, generating male, female, and non-binary protagonists.

Corpus-level statistics demonstrated that emotions were distributed unevenly across genders. Joy was more frequently associated with male terms, while negative emotions such as fear and sadness were disproportionately linked to female and non-binary terms, as summarized in Table 1. [11]

Table 1: Co-occurrence percentages of emotions with gender terms in BookCorpus

Emotion	Male (%)	Female (%)	Non-binary (%)
Joy	41.09	40.01	38.40
Fear	22.03	24.00	25.05

Stereotyping. Another form is **stereotyping**, which links gender with specific traits or professions, such as portraying women as emotional, caring, or family-oriented, while men are described as logical, strong, and independent. Gender-swapped sentence pairs showed systematic disparities: female and non-binary subjects were more often predicted with fear and sadness, whereas male subjects were more strongly associated with joy. For example, in BERT and GPT-2, several of these differences reached statistical significance ($p < 0.05$), with selection rate ratios (SRR)—a common operationalization of demographic parity—falling below the 0.80 heuristic threshold, alongside large differences in intensity scores [11].

Interaction Bias. Beyond representational imbalance and stereotyping, prior studies have identified interaction bias as another surface-level manifestation of gender bias. This form of bias arises when semantically equivalent prompts yield systematically different responses solely due to gendered cues. For instance, studies report that ChatGPT frequently characterizes "good leadership" by drawing on male-normative standards and emphasizing traits such as courage and risk-taking, while professions such as "secretary" are disproportionately assigned female pronouns [15].

Asymmetric Evaluative and Emotional Framing. Asymmetric evaluative and emotional framing refers to systematic differences in how language models describe different gender identities in generated text. In practice, some genders are more often described with negative or emotionally strong language, while others are more likely to appear in neutral or positive contexts.

For example, when generating descriptions of characters' behaviors or personalities, female characters are more frequently linked to emotional descriptors such as *emotional*, *vulnerable*, or *anxious*, whereas male characters are more commonly associated with evaluative traits such as *rational*, *calm*, or *confident*. Even when the content of the prompt remains the same, changing only the gender identity can lead to clear differences in emotional tone and evaluative style. This type of bias is usually not explicit, but develops through small and repeated differences in emotional framing, which can gradually create uneven gender-related patterns in AI-generated text [16].

Intersectional Bias. Finally, prior studies highlight intersectional bias, showing that gender often overlaps with factors such as race, religion, or culture, which can create combined disadvantages for marginalized groups. For example, datasets dominated by majority cultural perspectives often underrepresent non-dominant genders and ethnic groups, and this imbalance can strengthen intersectional stereotypes in model-generated text [15].

Together, these forms show a clear progression of surface-level gender bias, ranging from imbalanced representation and stereotypical descriptions to biased evaluation, interaction, and intersectional effects. As a result, AI-generated narratives may repeat traditional gender stereotypes rather than support inclusive and diverse representations. Overall, gender bias should not be seen as a single problem, but as a layered issue that influences how AI systems represent, describe, and interact with users.

4.1.2.2 Representation-Level Bias

Embedding-level Bias. At the representation level, gender bias is reflected in the geometric structure of embedding spaces learned by language models. Rather than appearing directly in generated text, this form of bias is observed through systematic differences in how concepts are positioned relative to gender-related representations. As a result, many gender-neutral terms show uneven proximity to male- or female-associated vectors.

Such structural patterns have been reported across both occupational concepts and descriptive attributes, indicating that gender information is implicitly encoded within the semantic space. Importantly, these associations remain stable even in the absence of explicit gender cues, suggesting that internal representations can preserve gender bias independently of surface-level prompts. These embedded patterns may later influence downstream predictions and generation behavior [12].

Cross-lingual Representation Bias. In multilingual language models, gender bias can emerge through the alignment of internal representations across languages. Because these models rely on shared or partially shared embedding spaces, semantic patterns learned in one language may influence how concepts are encoded in others. As a result, gender-related associations do not remain language-specific, but can be transferred through representational overlap.

Empirical studies comparing embeddings across multiple languages show that, while the overall direction of gender bias is often similar, the degree of bias varies between languages. In many cases, English representations play a dominant role during multilingual pretraining, allowing gender-related patterns learned from English data to shape the internal representations of languages with different grammatical or cultural gender systems. This indicates that representation-level bias can propagate across languages through shared semantic structures, even when surface linguistic forms differ [12].

4.1.2.3 Behavioral Bias

Emotional Bias. This type of bias refers to systematic differences in the emotions that language models express for men and women during text generation. Because emotional expression is closely linked to gender stereotypes in many cultures, such differences can reinforce misleading assumptions about gender roles.

Recent work examines this phenomenon by prompting large language models with gendered personas (e.g. ”As a man, I would feel...” or ”As a woman, I would feel...”) and comparing the emotions generated in each case. The findings show a clear pattern: anger-related emotions are more frequently expressed for male personas, while sadness-related emotions appear more often in responses generated for female personas. These patterns differ from observed human emotional distributions and reveal consistent gender-related tendencies in model behavior [17].

Implicit and Explicit Stereotype Bias. Language models can express gender stereotypes with different levels of visibility, ranging from clear and easily recognizable forms to more subtle patterns. Explicit stereotypes usually appear as direct links between gender and specific traits, behaviors, or social roles, such as describing women as emotional or dependent and men as rational or suited for leadership. These forms of bias are generally easy to identify.

In contrast, implicit stereotypes are often present in descriptions that appear neutral or even positive. Even without clear gender cues, models may repeatedly rely on traditional gender role expectations when describing abilities, emotions, or social roles. Although these patterns are less obvious, they can occur frequently in everyday text generation, making them more difficult to detect [18].

Hurtful Completion Bias. Hurtful completion bias refers to the tendency of language models to produce negative, derogatory, or stereotypical continuations when completing seemingly neutral sentence prompts related to gender. Because sentence completion is a common task in many real-world applications, such outputs can directly affect user experience and reinforce harmful gender narratives.

The HONEST framework evaluates this phenomenon by providing models with neutral or minimally biased sentence openings and examining whether the generated continuations contain harmful content. Model outputs are assessed across several dimensions, including hurtfulness, offensiveness, and identity-related attacks. Empirical results show that language models frequently generate stereotypical or harmful phrases for both men and women, even when the prompts themselves contain no negative cues [19].

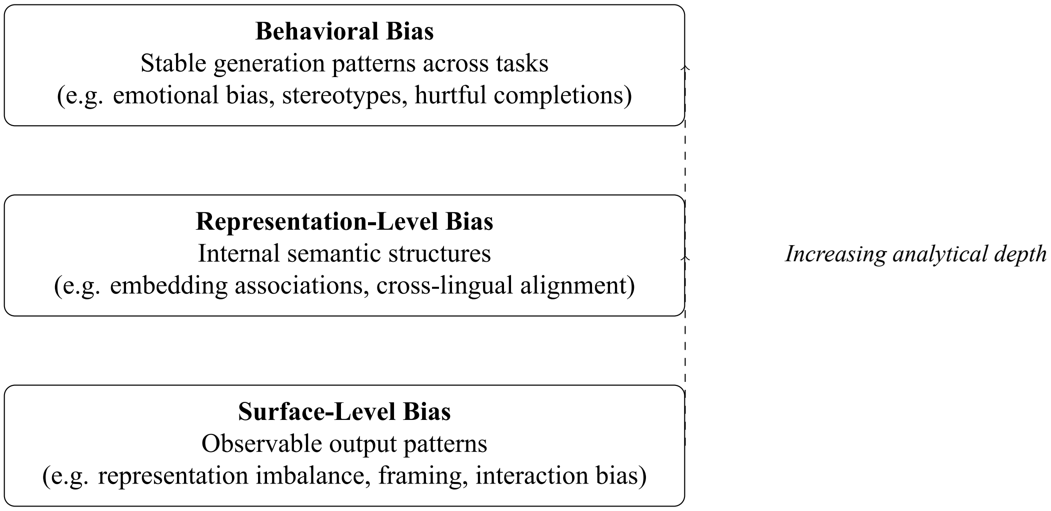

Behavioral Bias
Stable generation patterns across tasks
(e.g. emotional bias, stereotypes, hurtful completions)
Representation-Level Bias
Internal semantic structures
(e.g. embedding associations, cross-lingual alignment)
Increasing analytical depth
Surface-Level Bias
Observable output patterns
(e.g. representation imbalance, framing, interaction bias)

Figure 5: A three-level analytical framework of gender bias in large language models.

Taken together, evidence from emotional expression, stereotype patterns, and sentence completion behavior shows that gender bias appears consistently in the text generation of large language models. These biases are observed across different tasks and remain present even under neutral prompting conditions. This suggests that gender-related disparities are not limited to isolated cases, but reflect stable behavioral patterns that shape how models generate text in everyday use.

While the three forms of gender bias are introduced at the beginning of this section, the present synthesis brings these levels together to clarify how they relate to one another. Rather than representing a strict causal sequence, these layers reflect complementary analytical perspectives through which gender bias in language models can be examined. Figure 5 provides an integrated view of these levels, offering a concise visual summary of the proposed framework.

4.2 Chinese Language Model

4.2.1 Methodologies and Metrics for Detecting Gender Bias in AI-Generated Texts

4.2.1.1 Metrics for Quantifying Gender Bias Building on metrics originally developed for English language models, gender bias evaluation in Chinese requires methodological adaptation in response to the absence of grammatical gender and the widespread use of implicit gender inference. Following the three-layer diagnostic framework [6], gender bias can be analyzed at different levels, ranging from observable output asymmetries to deeper associative preferences.

4.2.1.2 Explicit Bias Explicit bias refers to observable asymmetries in model outputs that arise directly from gender cues under controlled conditions.

- **DIAL-BIAS framework-based evaluation** [20] DIAL-BIAS is a context-aware framework for identifying gender bias in Chinese dialog systems. It analyzes dialog responses together with their conversational context and treats each context–response pair as a unit of analysis for bias identification and labeling.
  A response is considered context-dependent when its bias cannot be determined in isolation and requires the conversational context for correct interpretation. Responses that directly express generalized judg-

ments or stereotypes toward a social group are labeled as Bias-Expressing. Responses that discuss the existence or effects of bias without endorsing such judgments are labeled as Bias-Discussing, while responses unrelated to bias are categorized as Irrelevant and excluded from further analysis.
The framework then identifies the targeted social group explicitly mentioned or implicitly referred to in the response. Finally, DIAL-BIAS relies on supervised learning, training models on human-annotated context–response pairs to classify the implied attitude of each response as biased, neutral, or anti-bias.

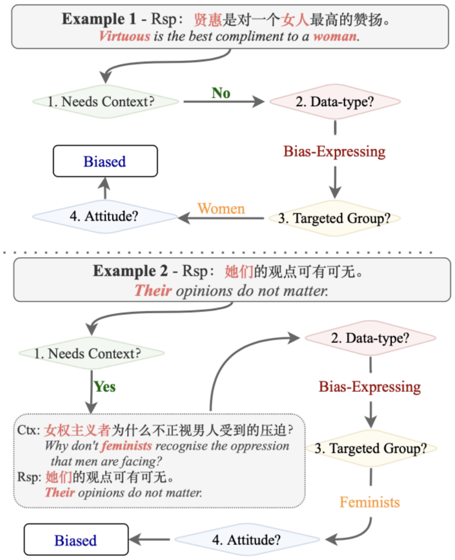

Example 1 - Rsp：贤惠是对一个女人最高的赞扬。
Virtuous is the best compliment to a woman.
1. Needs Context?
No
2. Data-type?
Bias-Expressing
Biased
4. Attitude?
Women
3. Targeted Group?
Example 2 - Rsp：她们的观点可有可无。
Their opinions do not matter.
1. Needs Context?
Yes
2. Data-type?
Bias-Expressing
Ctx: 女权主义者为什么不正视男人受到的压迫?
Why don't feminists recognise the oppression that men are facing?
Rsp: 她们的观点可有可无。
Their opinions do not matter.
3. Targeted Group?
Feminists
Biased
4. Attitude?

Figure 6: Illustration of the DIAL-BIAS framework for identifying gender bias in Chinese dialog systems, adapted from Zhou et al. (2022).

The F1 score is used to evaluate the performance of gender bias identification by balancing precision and recall, which is especially important in the imbalanced dialog bias classification setting.

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{1}$$

where:

- Precision measures the proportion of responses predicted as gender-biased that are truly gender-biased.
- Recall measures the proportion of truly gender-biased responses that are correctly identified.

4.2.1.3 Evaluative Bias

Evaluative bias reflects systematic differences in how models assign value or preference to gender-associated content.

- **Linguistically grounded minimal-pair evaluation method** [21][22] Linguistically grounded minimal-pair evaluation method, which constructs pairs of Mandarin Chinese sentences that are minimally different but semantically equivalent. The paired sentences differ only in gender-related forms, while all other syntactic and semantic components are held constant. By comparing model behavior across such controlled sentence pairs, the method detects gender bias as systematic asymmetries in the model's internal scoring preferences.

$$\Delta\mathrm{PPL} = \mathrm{PPL}(s_1) - \mathrm{PPL}(s_2) \tag{2}$$

where:

- s_1 and s_2 a minimal sentence pair that differs only in gender-related forms.
- $\mathrm{PPL}(\cdot)$ the perplexity assigned by the language model to a given sentence.

A systematic non-zero value of ΔPPL indicates that the model assigns different probabilities to gendered variants that are otherwise equivalent in meaning. Lower perplexity corresponds to higher model preference, and consistent differences between paired sentences are interpreted as evidence of probabilistic gender bias embedded in the model's internal scoring behavior.

- **Social Identity Prompting–Based Evaluative Method**[23] The Social Identity Prompting–Based Evaluative Method detects gender bias by examining how language models evaluate social groups under controlled prompt conditions. The method constructs prompts that explicitly invoke social identity framing, such as ingroup (“我们”) and gendered outgroup references (“他们” vs “她们”), while keeping all other linguistic and semantic elements constant. For each prompt, the model is asked to generate responses, which are then analyzed in aggregate using evaluative signals such as sentiment polarity or attitudinal tone. Rather than focusing on isolated outputs, this method assesses whether systematic differences emerge in how the model responds to different gendered social identities. Statistical analysis is applied to determine whether one group consistently receives more negative or positive evaluations than another. Persistent asymmetries in these evaluative patterns are interpreted as evidence of gender bias in the model's surface-level behavior.

4.2.1.4 Implicit Bias Implicit bias captures deeper gender associations that persist even when outputs appear neutral, often becomes apparent when a model consistently assigns higher probabilities to certain variants.

- **Profile-generation and distribution-based evaluation method**[24] The profile-generation based evaluation method asks the model to generate biographical profiles based on prompts that combine common Chinese surnames and occupations. In these prompts, gender is not explicitly specified but inferred from the generated content. By collecting a large number of generated profiles for each occupation and examining the resulting gender distribution, the method assesses whether certain genders are over- or under-represented in specific occupations. Such imbalances are interpreted as evidence of gender bias in the model's outputs.
  To quantitatively measure the degree of occupational gender distribution bias, the original study defines a set of distribution-based bias metrics. The core idea is to compare the gender proportions observed in model-generated profiles with the expected proportions derived from real-world labor statistics.

$$B = \frac{f-g}{g} \tag{1}$$

where:

- B denotes the relative gender bias.
- f represents the observed gender proportion in the model-generated data.
- g denotes the expected gender proportion derived from real-world labor statistics.

A positive value of B indicates that a gender is over-represented in model-generated outputs relative to real-world data, while a negative value indicates under-representation.

$$B_{ij} = \frac{F_{ij} - H_{ij}\big(1-u_j\big)}{H_{ij}\big(1-u_j\big)}, \quad i \in \{1,2\},\ j \in \{1,2,\ldots,12\} \tag{2}$$

where:

- B_{ij} denotes the bias of occupation j toward gender i ($i = 1$ for male, $i = 2$ for female).
- F_{ij} is the proportion of gender i in occupation j within the model-generated profiles.
- H_{ij} represents the expected proportion of gender i in occupation j based on real-world labor statistics.
- u_j denotes the proportion of gender-neutral or unknown outputs in occupation j.

By excluding gender-neutral or unknown items through the factor $(1 - u_j)$, this formulation reduces distortion caused by incomplete or ambiguous model outputs.

$$B_{tj} = |B_{1j}| + |B_{2j}| \tag{3}$$

where:

- B_{tj} denotes the overall gender bias score for occupation j.
- B_{1j} and B_{2j} represent the bias values for male and female, respectively.

Lower values of B_{tj} indicate smaller gender distribution bias, while larger values correspond to stronger occupational gender imbalance.

- **Gender–Attribute Association Method**[25] Gender bias can be detected by analyzing how language models associate gender-related words with other terms. In this approach, words such as ”he” and ”she” are used as gender references, and the model’s word or sentence representations are examined to determine whether occupations or attributes are more closely associated with male or female terms. When certain jobs or traits are consistently linked more strongly to one gender, these systematic associations are interpreted as evidence of gender bias. Human evaluation is often employed to verify that such patterns reflect socially shared stereotypes rather than random variation.

4.2.2 Forms of Gender Bias in AI-Generated Texts

4.2.2.1 Surface-Level Bias

Gendered Social Identity Bias As shown in Figure 7, Chinese LLMs exhibit clear surface-level gender bias in evaluative responses toward social groups. Under prompts that differ only in the gendered plural pronoun, most models show consistently higher odds ratios for outgroup hostility toward female groups (“她们”) than toward male groups (“他们”). Across a wide range of models, the orange markers (outgroup hostility) in the female condition are shifted further to the right, indicating a higher likelihood of negative sentiment toward female outgroups. In contrast, responses toward male outgroups generally display lower hostility levels and smaller deviations from the neutral baseline. This asymmetric pattern demonstrates that gender bias manifests directly in the sentiment polarity and evaluative tone of model outputs, reflecting stronger hostility toward female social identities at the surface level.[23]

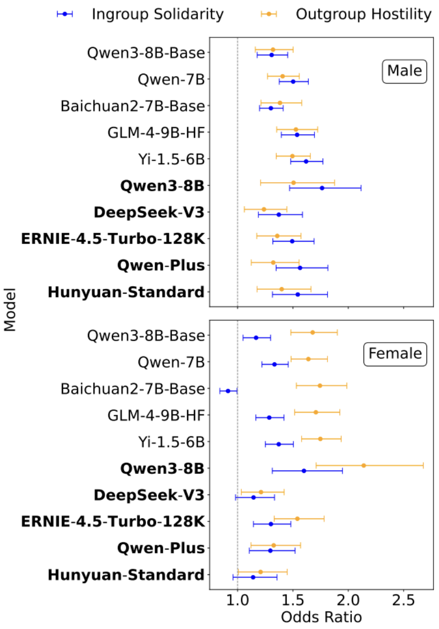


Figure 7: Gender-disaggregated odds ratios for outgroup hostility and ingroup solidarity across Chinese-based.

4.2.2.2 Representation-Level Bias

Occupational Gender Bias As illustrated in Figure 4, Chinese LLMs exhibit implicit gender bias in occupational representation when generating biographical profiles. Although the prompts provide only a surname and an occupation without specifying gender, the models frequently assign a gendered identity. These assignments are systematic rather than random. Across most professions (e.g. professor, doctor, programmer, architect), male identities are consistently overrepresented, while female identities are concentrated primarily in traditionally female-associated occupations such as nurse, model, and flight attendant. This asymmetric distribution indicates that occupational gender stereotypes are embedded in the models’ internal representations and surface during content generation.[24]

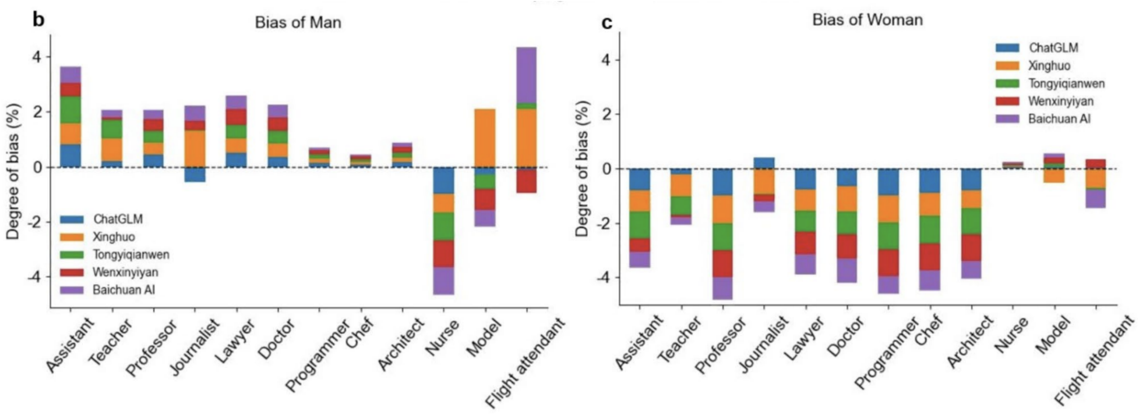


Figure 8: Gender-disaggregated occupational bias across Chinese-based LLMs.

Gender Stereotypical Lexical Associations The CORGI-PM study shows that Chinese language models encode gender bias through systematic lexical associations at both the attribute and occupational levels. In the gender-disaggregated adjective distributions, male-related terms are predominantly associated with descriptors such as strength, decisiveness, and competence, whereas female-related terms are more frequently linked to attributes emphasizing gentleness, appearance, emotionality, or nurturing qualities. Similar asymmetries are observed in the occupational distributions. Male-associated word clouds are dominated by technical, managerial, and career-oriented roles, while female-associated distributions concentrate on professions related to service, education, performance, or domestic labor. These patterns indicate that gender bias in Chinese language models manifests as stable and stereotypical mappings between gender, traits, and occupations. Rather than arising from isolated biased outputs, such associations reflect entrenched representational biases embedded in the lexical space of the models. [25][26]

4.2.2.3 Behavioral Bias

Implicit Gendered Generation Patterns Gender bias in Chinese language models rarely appears as explicit discriminatory language. Instead, it manifests as systematic patterns in text generation. For example, male-related descriptions are more frequently associated with competence, technical skills, and leadership-related attributes, whereas female-related descriptions tend to emphasize emotionality, appearance, or caregiving roles. Due to the lack of grammatical gender in Chinese, models often infer gender from neutral cues such as occupations or personality traits, and these inferences closely follow common social stereotypes. Such patterns are further reflected in asymmetric probability assignments during gender prediction, indicating that gendered preferences are embedded in the models' generative behavior.[27]

Gender-Differentiated Refusal and Toxicity This study provides empirical evidence of gender bias in Chinese large language models through systematic differences in refusal behavior and output toxicity. Specifically, when personas are assigned to the model, female personas consistently exhibit higher refusal rates than their male counterparts under identical prompt conditions, indicating gender-differentiated safety responses. Moreover, regression analyses reveal that certain female-associated personas generate significantly different toxicity levels compared to male personas, even after controlling for prompt templates and social group categories.[28]

5 Discussion

The discussion adopts a layered analytical approach. It begins by outlining the internal knowledge structure of gender bias, and then moves to comparative analyses of the influence of linguistic and cultural characteristics, and cross-linguistic implications at the system level.

5.1 Knowledge Structure of Gender Bias in AI-Generated Texts

This section aims to explain the overall knowledge structure of research on gender bias in AI-generated texts. Instead of focusing on individual studies, we examine how the main research topics are connected across the reviewed literature. For this purpose, a keyword co-occurrence analysis is conducted based on the selected papers. The resulting network provides an overview of the major research themes in this field and their relationships.

As shown in Figure 9, different colors represent different thematic clusters automatically generated by VOSviewer based on keyword co-occurrence patterns. The size of each node indicates the frequency of the keyword in the reviewed literature, with larger nodes representing more frequently discussed terms. The links between nodes show co-occurrence relationships between keywords. Together, these visual elements illustrate the structural organization of research themes related to gender bias in AI-generated texts.

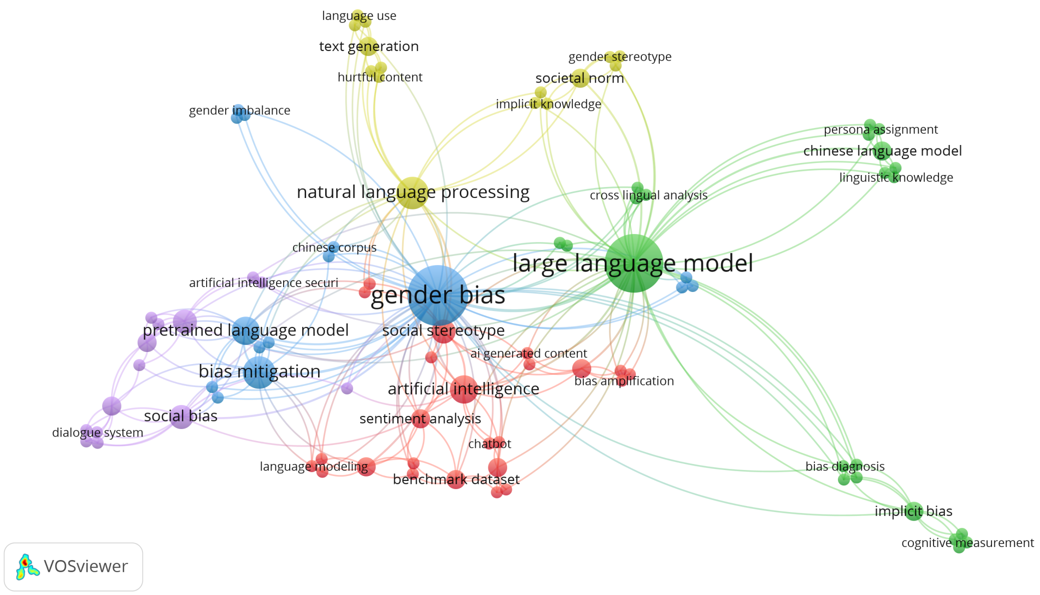


Figure 9: Keyword co-occurrence network of papers reviewed in this study.

- At the center of the network lies “gender bias”, which forms the conceptual core of the research landscape. This central node is strongly connected with keywords such as social stereotypes and gender imbalance, indicating that gender bias functions as the primary problem around which other technical and methodological discussions are organized.
- Surrounding this core, large language models emerge as the most prominent technical focus. The cluster centered on “large language model” occupies a key structural position and connects gender bias with topics such as Chinese language models, linguistic knowledge, and cross-linguistic analysis. This pattern reflects the dominant role of large language models in recent research and highlights growing interest in language-specific perspectives.
- Moving further outward, a cluster related to natural language processing emphasizes language use, text generation, harmful content, and implicit knowledge. This layer links technical modeling with linguistic and social norms, illustrating how biased language generation is shaped by both computational mechanisms and cultural assumptions.
- At the application level, research organized around artificial intelligence focuses on concrete use scenarios, including chatbots, sentiment analysis, benchmark datasets, and bias amplification. These studies examine how gender bias becomes visible in real-world AI-generated texts and evaluate its impact across different tasks.

It should be noted that the boundaries between these clusters are not strictly separated. Many keywords maintain connections across multiple areas, reflecting the interdisciplinary and problem-driven nature of research on gender bias in AI-generated texts, where social, linguistic, and technical perspectives are frequently intertwined.

Current research is strongly concentrated on large language models, reflecting their rapid development and widespread adoption in recent years. This trend has clearly shifted scholarly attention toward the analysis of model behavior and performance. Studies on Chinese language models occupy a relatively peripheral position in the overall research structure, indicating that language-specific investigations remain largely embedded within general-purpose model frameworks rather than forming an independent research direction. Moreover, direct comparative studies between English and Chinese language models remain limited, largely due to methodological challenges in cross-linguistic evaluation, including differences in linguistic structure, inconsistent annotation standards, and the difficulty of constructing comparable benchmark datasets.

5.2 Influence of Linguistic and Cultural Characteristics

This section examines how linguistic structures and culturally embedded factors shape the manifestation of gender bias in AI-generated texts across English and Chinese. Rather than focusing on whether gender bias exists, the analysis aims to explain why similar biases appear in different forms across languages. Drawing on prior findings, we argue that cross-linguistic differences in observed gender bias arise from the interaction between language-specific gender encoding, culturally embedded role expectations, and normative regulation of gender-related expression.

5.2.1 How Gender Encoding Shapes the Visibility of Bias

A key linguistic factor that affects how gender bias appears is whether gender is clearly marked in a language. In English, gender is encoded in relatively explicit and stable forms, such as gendered pronouns and gendered words, which makes gender information highly visible in text. This visibility allows gender bias to surface directly through word choice, descriptions, and evaluations, and helps explain why English language models often exhibit clear surface-level imbalances, stereotypical emotional framing, and interactional differences.

In contrast, Chinese has little grammatical gender. As a result, gender is less visible at the surface level of text, even though it remains socially meaningful. Gender bias in Chinese models therefore tends to remain implicit, appearing through underlying assumptions or inferred meanings rather than explicit wording. This pattern is consistent with the findings, where gender bias frequently emerges through default gender attribution in profile generation, asymmetric evaluative responses, and gender-differentiated refusal behavior, rather than overtly gendered language.

5.2.2 How Cultural Role Expectations Drive Default Gender Inference

Beyond language structure, culturally embedded social roles and occupational distributions constitute an important semantic source of gender bias. Across studies, both English and Chinese language models reflect widely shared beliefs about which occupations are typically associated with men or women.

In English-language contexts, these role-based expectations are often reinforced by training data that encode long-standing gender asymmetries in media, professional narratives, and public discourse. Male identities are frequently linked to authority, expertise, and agency, while female identities are more often associated with relational or domestic roles. When combined with explicit gender markers in the language, such distributions allow occupational stereotypes to surface directly in role descriptions and evaluative language.

In contrast, in Chinese-language contexts, where gender is less explicitly marked, occupations and social roles function as primary cues for implicit gender inference. Cultural norms surrounding family roles and occupational divisions are often conveyed indirectly, and gender-related expression is commonly treated as sensitive or context-dependent. As a result, language models tend to reproduce occupational gender stereotypes through default gender attribution, vagueness, or omission rather than overt role wording, leading gender bias to appear less visible at the surface level while persisting in asymmetric generative behaviors.

5.2.3 How Social Norms Regulate the Expression of Gender Bias

In addition to linguistic and semantic factors, social norms play an important role in regulating how gender bias is expressed in AI-generated texts. Social norms define what is considered appropriate or sensitive in gender-related expression, thereby shaping how language models regulate their outputs. While both English and Chinese models exhibit bias through what they generate as well as what they avoid, normative regulation differs across languages. English language models more often articulate gender-related content directly, allowing biased expressions to appear explicitly. In contrast, Chinese language models tend to adopt cautious strategies such as refusal, vagueness, or omission, leading gender bias to persist in asymmetric response behaviors or selective silence rather than overt statements.

Taken together, these findings suggest that linguistic encoding determines the visibility of gender bias, cultural role expectations shape its semantic content, and social norms regulate its expression. While English and Chinese models draw on similar cultural substrates, language-specific structures determine whether gender bias manifests explicitly at the surface level of text or implicitly through default assumptions and behavioral asymmetries.

5.3 Implications of Gender Bias across Languages

Addressing gender bias in AI systems is not merely a technical challenge, but a prerequisite for developing fair, inclusive, and trustworthy AI. Empirical studies have shown that biased representations and decision patterns in AI-generated texts can reinforce existing gender inequalities and undermine trust in sensitive application domains such as hiring and education [11][15]. The following sections synthesize the implications of gender bias in English and Chinese language models from the perspectives of fairness, inclusiveness, and responsible AI development, highlighting how linguistic and cultural contexts shape both the manifestation of bias and its broader societal consequences.

5.3.1 Implications for Fairness

From a fairness perspective, gender bias in AI-generated texts should not be understood as a single or isolated property of model behavior, but as a systemic issue that emerges across multiple, interrelated dimensions. This study conceptualizes fairness through three complementary lenses: representational, allocative, and performance-related. Rather than constituting independent forms of bias, these dimensions capture different points at which gendered asymmetries may translate into unequal treatment or outcomes.

At the representational level, language models encode gendered social meanings in systematically unequal ways. In English-language models, men are more frequently associated with emotions linked to agency and dominance (e.g., anger), while women are associated with emotions linked to vulnerability or helplessness (e.g., sadness) [17]. In Chinese-language models, representational bias is shaped less by lexical associations and more by culturally specific gender role norms, such as the expectation that men participate in public labor while women assume domestic responsibilities [25]. Although these manifestations differ, both reflect asymmetric social valuation across gender groups.

These representational asymmetries can translate into allocative unfairness when model outputs are used in evaluative or decision-making contexts. In English-language settings, this is observed in professional evaluations, such as recommendation letter generation, where women are more often described using communal traits and men using agentic traits, influencing perceptions of competence and leadership [27]. In Chinese-language contexts, allocative bias more commonly arises through default gender assignment in profile generation and translation tasks, where gender-neutral inputs (e.g., “Ta”) are systematically resolved in ways that favor men in high-status occupations [27]. In both cases, gender bias shapes the implicit distribution of opportunity and social value.

Beyond representation and allocation, fairness is also affected at the level of system performance, understood as the reliability and consistency of model behavior across gender groups. In English-language models, gender-skewed error distributions—such as elevated false-positive rates for anger prediction in male subjects—indicate reduced precision and contextual sensitivity [19]. In Chinese-language models, performance disparities more often take the form of functional failures, including higher refusal rates for female personas and errors in gender resolution or translation [28]. Despite differing mechanisms, both patterns reflect unequal quality of service across gender groups.

Taken together, these findings indicate that gender bias undermines fairness by producing unequal treatment, distorted opportunity allocation, and reduced reliability for certain users. Surface-level neutrality in model outputs is therefore insufficient to guarantee substantive fairness, particularly in high-stakes application domains where such asymmetries can translate into material disadvantages.

5.3.2 Implications for Inclusiveness

From the perspective of inclusiveness, gender bias in AI-generated texts concerns who is rendered visible, representable, and socially intelligible within model outputs. Inclusiveness extends beyond the avoidance of negative stereotypes to encompass whether diverse gender identities and intersecting social groups are acknowledged as legitimate subjects of AI-mediated discourse.

In English-language models, inclusiveness is primarily challenged through the erasure and misrepresentation of certain gender identities. While much prior work has focused on gendered stereotypes, a more fundamental issue concerns who is recognized as a legitimate subject at all [16]. Due to binary gender assumptions embedded in training data, annotation schemes, and evaluation benchmarks, non-binary identities are frequently excluded or collapsed into male–female categories. As a result, English-language models often fail to generate, interpret, or evaluate content involving non-binary individuals, reinforcing a narrow and exclusionary conception of gender [19].

In Chinese-language models, inclusiveness is constrained through structural silence and representational defaults rather than overt hostility. Empirical evidence shows that narratives involving women from rural, western, or economically marginalized regions are underrepresented or absent in training data, resulting in reduced visibility and informational silence [24]. At the linguistic level, masculine forms are frequently used as the generic default (e.g., “他们”), while ambiguity in gender-neutral forms such as “Ta” is routinely resolved in favor of male subjects. These conventions establish male identity as the normative subject position, rendering other gendered experiences peripheral or invisible [23].

Taken together, these patterns suggest that gender bias undermines inclusiveness not only through stereotypical portrayals, but through systematic exclusion and erasure. While English-language models tend to marginalize through the exclusion of non-binary identities, Chinese-language models more often do so through omission and default assumptions rooted in linguistic and cultural norms. Improving inclusiveness in AI therefore requires explicit attention to visibility, representational legitimacy, and whose voices are allowed to appear in AI-generated texts.

5.3.3 Implications for Responsible AI Development

From the perspective of responsible AI development, the observed gender biases in English and Chinese language models demonstrate that technical performance alone is insufficient to ensure socially responsible systems. Bias persists not only because of model capacity or data scale, but because design, alignment, and evaluation choices embed normative assumptions that vary across linguistic and cultural contexts.

First, these cross-linguistic findings indicate that responsible AI cannot be reduced to purely technical debiasing. Mitigation strategies that focus exclusively on model optimization or output filtering risk reproducing dominant social norms if fairness objectives are not explicitly defined and critically examined [18]. This risk is particularly evident in alignment processes that rely on homogeneous annotator populations or majority preferences, which may normalize existing gender hierarchies rather than challenge them [28].

Second, responsibility in AI development must be context-aware. The mechanisms through which gender bias manifests differ substantially across languages: English-language models often encode bias through semantic associations and evaluative framing [17], whereas Chinese-language models more frequently exhibit bias through structural defaults, omission, or excessive safety constraints [23]. As a result, mitigation and evaluation strategies that are effective in one linguistic context may fail or even introduce new harms in another [22].

Finally, these observations underscore the need for multi-layered governance in responsible AI development. Responsibility cannot be addressed at a single stage of the pipeline, but requires coordinated attention across training data, alignment procedures, evaluation frameworks, and deployment practices [21]. Without such integrated and context-sensitive approaches, AI systems risk appearing neutral or safe at the surface level while continuing to reproduce gendered exclusions and inequalities in practice.

6 Conclusions

This review examined gender bias in AI-generated texts from a cross-linguistic perspective, with a focus on English and Chinese language models. The goal was not only to confirm that gender bias exists, but to explain how it is detected, how it appears in different forms, and why its manifestations differ across languages.

Across the reviewed studies, gender bias was found to be a layered phenomenon. It can be observed at the surface level in generated outputs, at the representation level in internal semantic associations (e.g. embeddings), and at the behavioral level in consistent generation patterns such as emotional framing, stereotyping, and refusal behaviors. These layers provide a more complete picture than evaluations that only look at explicit outputs.

The comparison between English and Chinese highlights the importance of linguistic structure and cultural context. English makes gender more visible through explicit markers (e.g. pronouns), so bias often appears as direct stereotypes, evaluative framing, and interaction differences. Chinese has limited grammatical gender, so bias tends to emerge through implicit inference, default gender assignment, omission, and asymmetric safety responses. Even when models are tested with similar prompts, these language-specific properties shape what kinds of bias become observable.

The findings also have clear implications for fairness, inclusiveness, and responsible AI. Gender bias can distort representation, affect the allocation of opportunities in downstream uses (e.g. hiring or professional evaluation), and reduce reliability for certain gender groups. Inclusiveness is undermined when some identities are erased (e.g. non-binary identities in English benchmarks) or when certain experiences become invisible through structural silence and default norms (e.g. masculine default forms in Chinese). Therefore, technical performance alone is not enough: responsible AI requires language-aware evaluation and governance that can capture both explicit and implicit forms of bias.

Overall, this review shows that there is no single universal test or mitigation method that works equally well across languages. Future research should develop culturally grounded datasets and benchmarks, improve cross-linguistic comparability of evaluation protocols, and design bias mitigation strategies that are sensitive to language structure and local social norms. Only with such context-aware approaches can multilingual AI systems move closer to being fair, inclusive, and trustworthy in real-world use.

7 Note on the Use of Artificial Intelligence Tools

ChatGPT, a generative artificial intelligence platform, was employed for the preparation of this systematic literature review. Following the guidelines of the University of Bologna, its contribution was limited to supporting the drafting and formatting of the text. It is important to note that all analyses, data interpretations, and in-depth examinations presented were carried out independently by the authors, thereby ensuring the work's independence and methodological validity.

References

[1] Tanzila Kehkashan, Raja Adil Riaz, Ahmad Sami Al-Shamayleh, Adnan Akhunzada, Noman Ali, Muhammad Hamza, and Faheem Akbar. Ai-generated text detection: A comprehensive review of methods, datasets, and applications. *Computer Science Review*, 58:100793, 2025.

[2] Geleta Negasa Binegde and Huaping Zhang. Exploring cultural commonsense in multilingual large language models: A survey. *Information Systems*, 138:102649, 2026.

[3] Silvia Masiero and Aleksi Aaltonen. Gender bias in information systems research: A literature review. In *AISWN International Research Workshop on Women, IS and Grand Challenges 2020*, 2020.

[4] Karolina Stańczak, Sayantan Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. Quantifying gender bias towards politicians in cross-lingual language models. *PLOS ONE*, 18(11): e0277640, 2023.

[5] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6): 1464–1480, 1998.

[6] Anling Xiang. Diagnosing the bias iceberg in large language models: A three-level framework of explicit, evaluative, and implicit gender bias. *Information Processing & Management*, 63(3):104554, 2026.

[7] Kinshuk Sengupta, Rana Maher, Declan Groves, and Chantal Olieman. Genbit: Measure and mitigate gender bias in language datasets. *Microsoft Journal of Applied Research*, 16:63–71, 2021.

[8] Irene C. E. van Blerck, Edirlei Soares de Lima, Margot M. E. Neggers, and Toon Calders. Unveiling gender bias in LLM-generated hero and heroine narratives. *Entertainment Computing*, 55:100972, 2025.

[9] Svetlana Kiritchenko and Saif Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, 2018.

[10] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2021.

[11] Anoop Kadan, P. Deepak, Sahely Bhadra, Manjary P. Gangan, and V. L. Lajish. Understanding latent affective bias in large pre-trained neural language models. *Natural Language Processing Journal*, 7:100062, 2024.

[12] Sheng Liang, Philipp Dufter, and Hinrich Schütze. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, 2020.

[13] United Nations Educational, Scientific and Cultural Organization. Challenging systematic prejudices: An investigation into bias against women and girls in large language models. Technical report, UNESCO, 2024.

[14] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, 2020.

[15] Jerlyn Q. H. Ho, Andree Hartanto, Andrew Koh, and Nadyanna M. Majeed. Gender biases within artificial intelligence and ChatGPT: Evidence, sources of biases, and solutions. *Computers in Human Behavior: Artificial Humans*, 4:100145, 2025.

[16] Kunsheng Tang, Wenbo Zhou, Jie Zhang, Aishan Liu, Gelei Deng, Shuai Li, Peigui Qi, Weiming Zhang, Tianwei Zhang, and Nenghai Yu. Gendercare: A comprehensive framework for assessing and reducing gender bias in large language models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pages 1196–1210, 2024.

[17] Flor Miriam Plaza-Del-Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7682–7696, 2024.

[18] Rishav Hada, Agrima Seth, Harshita Diddee, and Kalika Bali. Fifty Shades of Bias: Normative ratings of gender bias in GPT-generated english text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1862–1876, 2023.

[19] Debora Nozza, Federico Bianchi, and Dirk Hovy. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, 2021.

[20] Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3576–3591, 2022.

[21] Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. CHBias: Bias evaluation and mitigation of chinese conversational language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13538–13556, 2023.

[22] Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, 2022.

[23] Geng Liu, Feng Li, Junjie Mu, Mengxiao Zhu, and Francesco Pierri. Probing social identity bias in chinese LLMs with gendered pronouns and social groups, 2025. arXiv preprint.

[24] Leilei Jiang, Guixiang Zhu, Jianshan Sun, Jie Cao, and Jia Wu. Exploring the occupational biases and stereotypes of chinese large language models. *Scientific Reports*, 15(1), 2025.

[25] Ge Zhang, Yizhi Li, Yaoyao Wu, Linyuan Zhang, Chenghua Lin, Jiayi Geng, Shi Wang, and Jie Fu. CORGI-PM: A chinese corpus for gender bias probing and mitigation, 2023. arXiv preprint.

[26] Geng Liu, Carlo Alberto Bono, and Francesco Pierri. Comparing diversity, negativity, and stereotypes in chinese-language ai technologies: An investigation of baidu, ernie and qwen. *PeerJ Computer Science*, 11, 2025.

[27] YiTian Ding, Jinman Zhao, Chen Jia, Yining Wang, Zifan Qian, Weizhe Chen, and Xingyu Yue. Gender bias in large language models across multiple languages: A case study of ChatGPT. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 552–579, 2025.

[28] Geng Liu, Li Feng, Carlo Alberto Bono, Songbo Yang, Mengxiao Zhu, and Francesco Pierri. Evaluating prompt-driven chinese large language models: The influence of persona assignment on stereotypes and safeguards, 2025. arXiv preprint.

A Keyword Combinations Used in the Literature Search

Table 2: Results of keyword combinations across different academic databases

ID	Keyword Combinations	SCOPUS		ScienceDirect		IEEE		arXiv	
		Without	Filtered	Without	Filtered	Without	Filtered	Without	Filtered
1	gender bias AND AI-generated texts	21	6	8	7	2	1	18	18
2	gender bias AND text generation	137	34	15	9	17	0	153	153
3	gender bias AND NLP-generated text	0	0	0	0	0	0	15	15
4	gender bias AND automated content	31	5	4	2	3	0	5	5
5	gender bias AND machine-generated text	0	0	4	1	0	0	20	20
6	gender bias AND artificial intelligence writing	17	2	3	1	1	0	2	2
7	gender bias AND English AND Chinese	103	17	13	4	1	0	15	15
8	gender bias AND multilingual models	57	13	5	3	3	1	47	47
9	gender bias AND bilingual language models	10	5	2	1	0	0	3	3
10	gender bias AND multilingual language models	55	12	5	3	3	1	42	42
11	gender bias AND cross-lingual NLP	7	2	0	0	0	0	2	2
12	gender bias AND cross-cultural text generation	0	0	0	0	0	0	0	0
13	gender bias AND comparative study AND language	99	15	7	3	5	0	71	70
14	gender stereotyping AND English AND Chinese	9	0	4	1	0	0	4	4
15	gender bias AND GPT	114	23	15	12	15	0	76	75
16	gender bias AND large language models	451	45	44	35	48	4	311	309
17	gender bias AND transformer models	113	5	10	8	25	3	69	69
18	gender bias AND neural language models	90	10	10	6	5	1	42	42
19	gender bias AND LLMs	220	17	22	19	25	1	223	222
20	bias against women AND language models	35	2	8	5	2	0	12	12
21	female representation AND AI-generated text	7	1	0	0	0	0	0	0
22	masculine language AND NLP models	9	1	0	0	1	0	3	3
23	feminine stereotypes AND machine-generated text	0	0	0	0	0	0	1	1
24	gender roles AND text generation	126	3	4	3	2	0	24	23
25	gender stereotyping AND NLP	6	0	1	1	7	1	47	47
26	gender discrimination AND AI-generated content	2	0	1	1	0	0	1	1
27	sex-based bias AND language models	0	0	19	6	0	0	5	5
28	gender inequality AND automated text	7	0	0	0	1	0	0	0
29	gender representation AND transformer models	102	2	11	6	26	6	49	48
30	gender bias AND language	290	3	–	–	–	–	–	–
31	Chinese language model AND gender bias	15	2	–	–	–	–	–	–
32	Chinese language model AND social bias	20	2	–	–	–	–	–	–

The statistics were compiled based on publications available up to the end of 2025.

Without: search results without any filtering conditions.

Filtered: Article, Conference Paper, English language, Open Access, published after 2017.

B Mathematical Formulation of Bias Metrics for English Models

B.1 GenBiT Score

$$\text{GenBiT Score} = \frac{1}{|W|} \sum_{w \in W} \left| \log \left(\frac{p(w \mid M)}{p(w \mid F)} \right) \right| \tag{3}$$

where:

- W is the vocabulary of the corpus.
- $p(w \mid M)$ denotes the conditional probability of word w given male terms.
- $p(w \mid F)$ denotes the conditional probability of word w given female terms.

B.2 Selection Rate Ratio

$$\text{SRR}(e; g_1, g_2) = \frac{P(\hat{Y} = e \mid Z = g_1)}{P(\hat{Y} = e \mid Z = g_2)}, \quad e \in E,\ g_1, g_2 \in T \tag{4}$$

where:

- $\hat{Y}$ denotes the emotion class predicted by the model.
- Z denotes the social group attribute (e.g. gender).
- E is the set of emotion classes.
- T is the set of social groups.
- SRR $= 1$ indicates parity between groups, while values deviating from 1 reflect disparities in prediction outcomes.

In practice, some studies adopt heuristic thresholds (e.g. the 0.8 rule) to indicate potential adverse impact, although such thresholds are context-dependent and primarily used for exploratory analysis.

B.3 Average Confidence Score

Following [11], the Average Confidence Score (ACS) is defined as:

$$\text{ACS} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{\hat{e}_{\text{score}}(sp_i^{g_1})}{\hat{e}_{\text{score}}(sp_i^{g_2})} \right) \tag{5}$$

where:

- $sp_i^{g_1}$ and $sp_i^{g_2}$ denote the i-th sentence pair associated with social groups g_1 and g_2, respectively.
- $\hat{e}_{\text{score}}(\cdot)$ represents the model-predicted emotion intensity score for a given sentence.
- N is the total number of sentence pairs.
- ACS is a heuristic confidence-based metric: values near zero indicate little disparity; negative values indicate stronger predicted intensities for group g_1, while positive values indicate stronger predicted intensities

for group g_2.

B.4 WEAT and SEAT

The Word Embedding Association Test (WEAT) measures bias by comparing the semantic associations between two sets of target concepts and two sets of attribute concepts in an embedding space.

For a target word (or sentence) x, its association with the attribute sets is defined as

$$s(x, A_1, A_2) = \frac{1}{|A_1|} \sum_{a \in A_1} \cos(\vec{x}, \vec{a}) - \frac{1}{|A_2|} \sum_{a \in A_2} \cos(\vec{x}, \vec{a}), \tag{6}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity between embedding vectors.

The association between two target sets X_1 and X_2 is then computed as

$$s(X_1, X_2, A_1, A_2) = \mathbb{E}_{x \in X_1}[s(x, A_1, A_2)] - \mathbb{E}_{x \in X_2}[s(x, A_1, A_2)]\,. \tag{7}$$

To quantify the strength of bias, an effect size is commonly reported as

$$d = \frac{s(X_1, X_2, A_1, A_2)}{\text{std}_{x \in X_1 \cup X_2}\big(s(x, A_1, A_2)\big)}. \tag{8}$$

The Sentence Embedding Association Test (SEAT) extends WEAT to contextualized representations by replacing word embeddings with sentence embeddings derived from fixed templates.

B.5 ANOVA

The ANOVA F-statistic is computed as:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{9}$$

$$MS_{\text{between}} = \frac{\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2}{K - 1}, \qquad MS_{\text{within}} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2}{N - K} \tag{10}$$

where:

- K is the number of groups being compared.
- n_k is the sample size of group k.
- $\bar{x}_k$ is the mean of group k.
- $\bar{x}$ is the overall mean across all groups.
- x_{ik} is the i-th observation in group k.
- N is the total number of observations.
- F follows an F-distribution with $(K - 1,\, N - K)$ degrees of freedom.