THE PSYCHOLOGY OF GEN AI

SUBTLE SIGNALS, GENDERED OUTPUTS: BIAS IN GENERATIVE AI

INTRODUCTION

Language is never neutral. Tone, indirectness, emotionality and communication style all carry social meaning, including gender signals. This project examines whether **gender stereotypes can emerge from pattern-based generation processes within large language models (LLMs) in response to subtle linguistic cues,** even when gender is not explicitly mentioned. Put differently, do gender-coded differences in how users phrase prompts influence whether LLMs produce gender-coded outputs?

Prior research demonstrates that LLMs can reproduce social biases present in their training data, including gender stereotypes in occupational and associative descriptions (Kotek et al., 2023; Soundararajan & Delany, 2024). More recent studies suggest that biased outputs may also emerge without explicit demographic labels, as contextual and linguistic features of prompts activate learned patterns during generation, producing effects consistent with implicit bias (Torres et al., 2025).

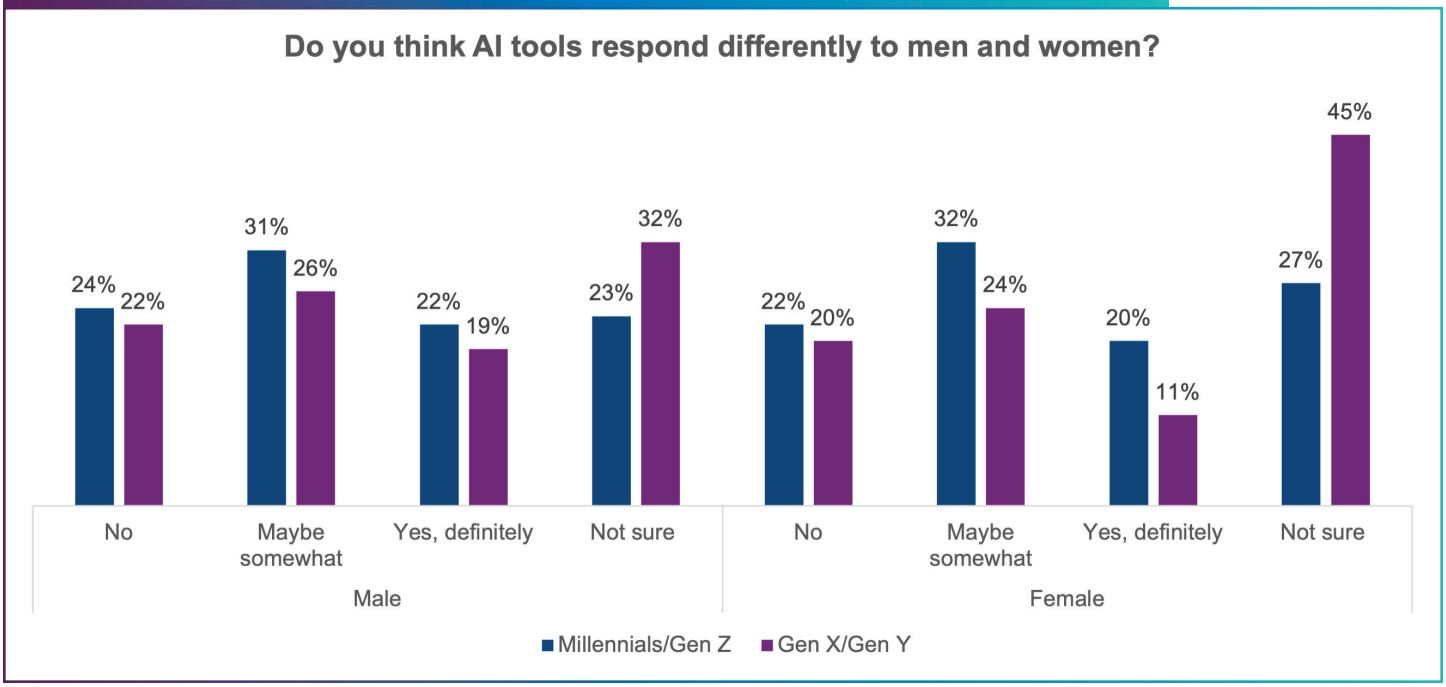
This study builds on that work by systematically varying prompt tone using psychologically grounded language traits, to examine how gender is operationalized at the input level—both explicitly and implicitly. The broader study is structured in two phases. Phase 1, reported here, examines differences in prompt construction generated by the model itself. Phase 2 examines how those prompt differences shape downstream model outputs, specifically shopping and product recommendation outcomes.

BACKGROUND

Recent consumer research suggests that people experience AI as a socially responsive system, one that mirrors how they communicate. Yet, in some cases, it appears to treat users differently based on perceived identity. In a nationally representative, omnibus survey of 2,000 U.S. adults, 52% of respondents reported that AI does not make any assumptions about user identity. At the same time, 28% believe it does and 20% are unsure, indicating that identity inference remains a concern or source of uncertainty for nearly half of users (MRXPros et al., 2025).

Nearly half of respondents (47%) also reported believing that AI tools exhibit at least some gender bias, with perceptions strongest among younger users (see Figure 1). These perceptions vary systematically with usage intensity. Among frequent AI users, 55% believe AI responds differently to men and women, compared with 47% of infrequent users and 28% of non-users. In other words, the more time people spend interacting with AI, the more likely they are to notice differences in tone, style or content that feels gendered.

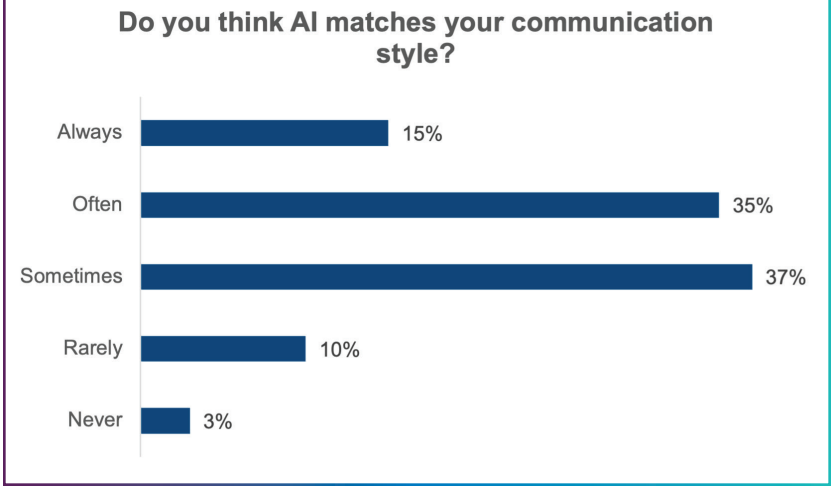
Figure 1: Perceptions of Whether AI Responds Differently to Men and Women, by Gender and Generation



Source: MRxPros Omnibus Survey, in partnership with MFour, November 2025, n=2,000 P18+ (MRXPros et al., 2025)

These perceptions are closely tied to how **users interact with AI systems**. Most consumers do not rely on structured prompts or templates but instead engage conversationally. A large majority (87%) report that AI matches their communication style at least sometimes, and those who consistently perceive this style-match are significantly more likely to believe that AI responds differently to men and women (64%) than those who do not (27%) (See Figure 2).

Figure 2: Perceived Frequency of AI Matching Users' Communication Style



Source: MRxPros Omnibus Survey, in partnership with MFour, November 2025, n=2,000 P18+ (MRXPros et al., 2025)

Taken together, these findings suggest that concerns about gender bias are not limited to whether systems produce different outputs for women and men. Instead, this is a perceived **bias experienced through interaction, where a user feels that the system responds differently because of how they communicate.** At the same time, survey responses indicate substantial variation in how such responsiveness is interpreted, underscoring the ambiguity of these interactions.

This raises an important question: **Can gendered patterns enter AI systems through the language of interaction itself, even in the absence of explicit demographic information?**

STUDY OVERVIEW

This research is designed as a two-phase experimental program. Phase 1 focuses on how gender enters AI systems at the point of interaction by examining prompt construction under implicit and explicit gender framing. Phase 2 examines how these prompt-level differences translate into model outputs, including AI-generated shopping and product recommendations.

While Phase 2 is more directly outcome-oriented, Phase 1 plays a critical foundational role, by isolating how linguistic style and persona specification shape the inputs that ultimately drive model behavior.

Method: Phase 1 — Meta-Prompt Construction

Phase 1 examines whether an LLM produces systematically different prompts when given implicit versus explicit gender cues, simulating how differences in user linguistic style may shape AI interactions. To do this, we instructed the model (GPT-5) to **generate prompts written in the voice of hypothetical users.**

Across all conditions, the underlying task was constant. Each prompt requested a birthday gift shopping list with concrete, brand-specific recommendations. What varied was how the request was phrased, specifically the prompt's **tone and linguistic framing.**

Meta-Prompt Design

Meta-prompts instructed the model to generate brief, LLM-ready prompts reflecting different gender framings. Gender was operationalized in two ways:

- **Implicit Gender Framing:** Gender was operationalized through psychologically grounded language traits commonly associated with gendered communication styles, including agentic vs. communal orientation, directness vs. indirectness, hedging, politeness and emotional expressiveness (Hochschild, 1979; Lakoff, 1973)[1]. Meta-prompts defined these traits without mentioning gender and specified relative levels along each dimension. For simplicity and control, the same percentage value was applied across all traits within each implicit condition.
- **Explicit Gender Framing:** Meta-prompts specified gender explicitly along a masculine–feminine spectrum (e.g., "0% feminine / 100% masculine" to "100% feminine / 0% masculine"). The model was instructed to generate a prompt reflecting the specified position while keeping task instructions identical.

Six percentage values (0%, 20%, 40%, 60%, 80%, 100%) were used for both implicit and explicit framing. For each value, 50 prompts were generated, yielding 300 prompts per framing condition.

By isolating prompt tone and linguistic structure while holding the task constant, Phase 1 establishes a controlled set of prompts that differ only in how gender is expressed.

[1] See Appendix for a list of all language traits and corresponding definitions.

0% Feminine Explicit Prompt

"Make me a 10-item birthday gift shopping list for an adult friend. Each pick must name a specific, well-known brand and the product type (not just a generic category). Cover a mix of EDC/gear, tech, apparel, home/bar, and self-care. Include a one-line why it's a solid pick and a rough USD price. Use widely available brands only-no niche labels or knockoffs. Mix budgets and cap it at one high-end splurge. Keep it tight and practical, no fluff. Output as a numbered list."

100% Feminine Explicit Prompt

"Hi love! Could you curate a chic, cozy, and thoughtful 10-item birthday gift shopping list for my adult friend? For each pick, include the exact brand and product name, a one-sentence why-it's-great, and an approximate price. Please mix categories (self-care, home, food/drink, tech, accessories), span price points (~$25-$300), keep it gender-neutral and classy, and choose items that are easy to buy online in the U.S. No gag gifts, no overly niche hobby gear. Format as a clean bullet list."

0% Feminine Implicit Prompt

"Generate a 10-item shopping list of birthday gift ideas. Each item must specify a distinct, real brand. Format as a numbered list with: item type - brand name. No commentary."

100% Feminine Implicit Prompt

"Hi there-could we maybe, if it's not too much trouble, gently brainstorm a 10-item birthday gift shopping list for a loved one? I'm hoping each suggestion might come from a specific brand, with a brief, warm note about why it could feel special and an approximate price range, if that's possible. It would be lovely if the list were clearly numbered and included a nice mix of interests-thank you so much for helping me make this feel caring and celebratory."

Linguistic Analysis of Generated Prompts

To evaluate whether generated prompts differed systematically by framing condition, we analyzed them using LIWC-22, a validated linguistic analysis tool that quantifies psychological and stylistic features of text.

We examined LIWC's four Summary Measures—**Analytical Thinking, Clout, Authenticity** and **Emotional Tone.** These capture higher-level dimensions of reasoning style, social orientation and affect. The measures were selected because they are algorithmically derived and suitable for short texts.

We also analyzed basic structural features, including **word count, punctuation use** and **the frequency of longer words** (seven characters or more), to assess differences in verbosity and surface structure.

FINDINGS

LIWC Summary Measures

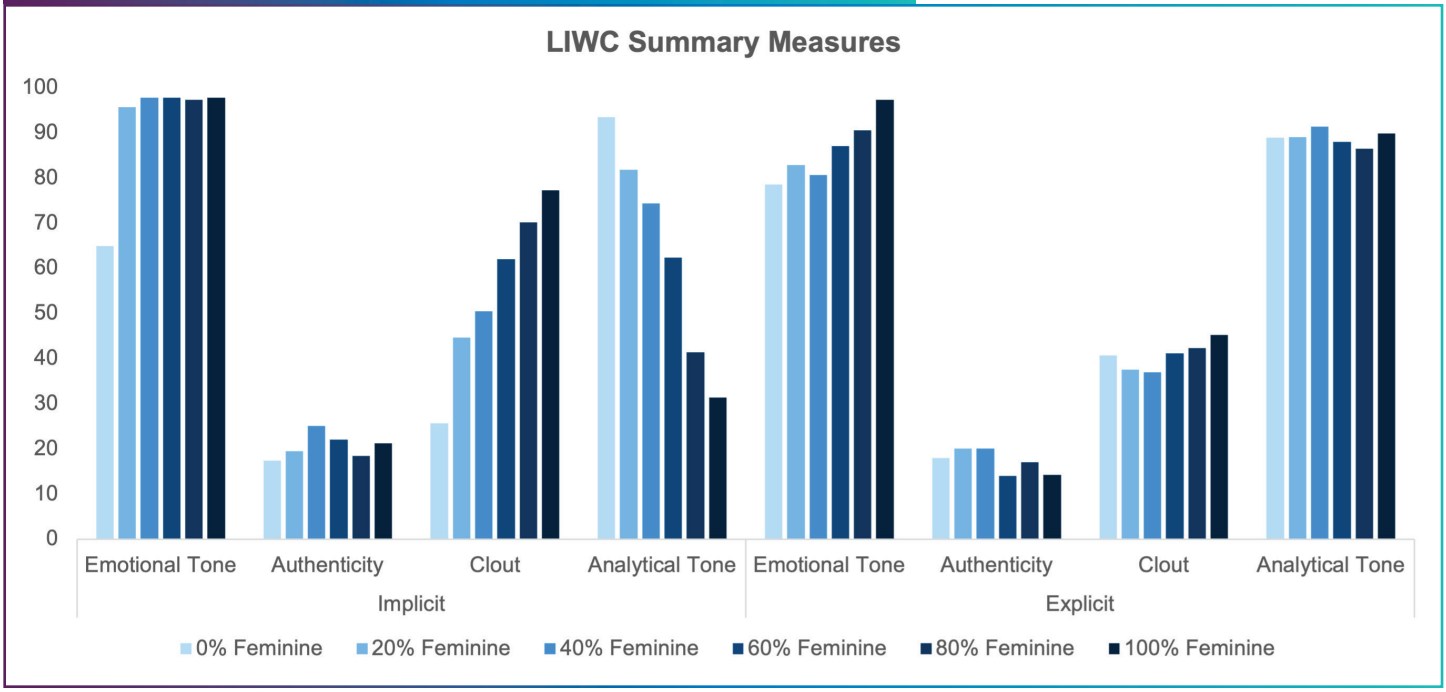
Under **implicit gender framing,** three of LIWC's four summary measures showed **consistent trends as femininity increased** (see Figure 3). **The language became less analytical (the analytic score decreased). It became more positive in tone (the emotional tone score increased), and higher in clout.** Authenticity, however, did not show a consistent pattern.

As femininity increased, the LLM's language became less analytical, more positive in tone and higher in clout.

The changes seen in the LIWC scores may reflect the way gender was built into the prompts themselves. Because the meta-prompts in the implicit gender framing criteria defined femininity and masculinity using language traits (such as indirectness, emotional expressiveness or confidence), those traits were picked up by LIWC's summary measures.

In contrast, prompts generated under **explicit gender** framing showed far fewer systematic changes. Only emotional tone increased with femininity, while analytical thinking, clout and authenticity remained relatively stable.

Figure 3: LIWC Summary Measures by Gender Framing Type and Femininity Level



Source: MRxPros Omnibus Survey, in partnership with MFour, November 2025, n=2,000 P18+ (MRXPros et al., 2025)

Structural Linguistic Features

Implicit framing also produced systematic differences in basic structure. As femininity increased, prompts became longer and used fewer long words. These trends did not appear under explicit framing.

Explicitly gendered prompts showed limited surface variation, primarily through punctuation use or stereotypical descriptors (e.g., "chic," "classy"). Implicitly gendered prompts, by contrast, exhibited broader shifts in structure and style.

Summary of Findings

Implicit, trait-based gender framing produced more pronounced and multidimensional linguistic variation than explicit gender labeling. While explicit framing primarily affected surface features such as punctuation or stereotypical gendered descriptors, implicit framing altered structure, reasoning style and emotional tone. Consistent with the LIWC analysis, these results suggest that **user-style differences encoded implicitly may drive more substantial linguistic variation than explicit gender labels alone.**

What Do These Findings Mean?

The findings demonstrate that **gendered patterns can enter AI systems** before any content is generated, **through the way requests are phrased.** By isolating prompt construction, the study identifies an upstream mechanism through which gendered responses may emerge, rooted in **interaction style rather than declared identity.**

This shifts how gender bias in AI should be understood.

> Rather than treating bias solely as a property of outputs, the results suggest that everyday user–system interaction, shaped by socially gendered communication norms, may influence system behavior.

Importantly, while this experiment does not directly measure how users interpret AI outputs, it lays the groundwork for further research linking user perceptions of bias to measurable linguistic differences in prompt construction.

IMPLICATIONS FOR MARKETING AND ADVERTISING

Generative AI tools are increasingly used upstream, to draft briefs, generate creative prompts, simulate consumer perspectives and support research design. If gendered patterns are present at the prompt level, bias can be introduced before strategy, creative or analysis begins.

Gender-based prompt variances can affect every facet of AI-driven marketing, including:

- **Creative development:** Subtle differences in prompt tone may steer AI-generated ideas toward gender-stereotypical products, benefits or narratives.
- **Personalization, targeting and product recommendations:** In shopping and e-commerce contexts, AI systems increasingly shape which products or offers are surfaced. If systems adapt to user style rather than explicit preferences, prompt-level gendering may influence recommendation paths before observed behavior is incorporated.
- **AI-assisted research:** In applications such as synthetic respondents, concept testing or insight generation, prompt-level gendering may influence findings independently of real consumer data.
- **Brand trust, governance and reputational risk:** Behaviors often framed as "good UX," such as conversational tone matching, could heighten perceptions of gender profiling.

NEXT STEPS

Phase 1 identified gendercoded language in prompts. Phase 2 will examine how those differences influence the content, categories and framing of AI-generated shopping recommendations.

Together, the two phases will allow us to trace how bias enters and propagates the system, from interaction style to prompt construction to final output, providing an end-to-end perspective essential for responsible AI use in marketing and advertising.

APPENDIX

Implicit Gender Traits

- Communal vs. Agentic Orientation: Agentic = self-reliant, competitive, goal/achievement-focused; Communal = warm, helpful, relationship-focused, collaborative framing.
- Indirectness vs. Directness: Use of imperatives/declaratives vs. mitigated/indirect requests (e.g., "Could we", "Might you). Elaborated and explanatory vs. concise and directive, likeliness to interrupt
- Hedging vs. Non-Hedging: Hedges ("maybe," "sort of," "I think"), modal verbs ("could you, "would you mind").
- Politeness vs. Impoliteness: Greetings, apologies, honorifics, please and thank you.
- Emotional Expressiveness vs. Non-expressiveness: Managing others' feelings, supplying warmth and apologetic or soothing tone. Emotive adjectives, perspective-taking, supportive and affiliative speech.

LIWC Summary Measures (LIWC, n.d.):

- **Analytical Thinking**: "The analytical thinking variable is a factor-analytically derived dimension, based on several categories of function words. Originally published as the Categorical-Dynamic Index (CDI), analytic thinking captures the degree to which people use words that suggest formal, logical and hierarchical thinking patterns. People low in analytical thinking tend to write and think using language that is more intuitive and personal. Language scoring high in analytic thinking tends to be rewarded in academic settings and is correlated with things like grades and reasoning skills. Language scoring low in analytic thinking tends to be viewed as less cold and rigid, and more friendly and personable."
- **Authenticity**: "When people reveal themselves in an "authentic" or honest way, they tend to speak more spontaneously and do not self-regulate or filter what they are saying. The algorithm for authenticity was originally derived from a series of studies where people were induced to be honest or deceptive (Newman et al., 2003), as well as a summary of deception studies published in the years afterwards. However, over time we have come to understand that the authenticity measure has less to do with "deception," in a traditional sense and is, instead, more a reflection of the degree to which a person is self-monitoring. Examples of texts that score low in authenticity include prepared texts (i.e., speeches that were written ahead of time), and texts where a person is being socially cautious. Examples of texts that score high in authenticity tend to be spontaneous conversations between close friends or political leaders with little-to-no social inhibitions."
- **Clout**: "Clout refers to the relative social status, confidence or leadership that people display through their writing or talking. The clout algorithm was developed based on the results of a series of studies where people interacted with one another (e.g., Kacewicz et al., 2013). Note that clout is different from the concept of "power" (including the LIWC-22 "power" variable). Power or, more accurately, the need for power, reflects people's attention to or awareness of relative status in a social setting. You can have a confident leader who has no interest in other people's standing in the social hierarchy."
- **Emotional Tone**: "Although LIWC-22 includes both positive tone and negative tone dimensions, the tone variable puts the two dimensions into a single summary variable. The algorithm is built so that the higher the number, the more positive the tone. Numbers below 50 suggest a more negative emotional tone."

REFERENCES

Hochschild, A. R. (1979). Emotion work, feeling rules, and social structure. *American Journal of Sociology, 85*(3), 551-575.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125–143.

Kotek, H., Dockum, R., & Sun, D. (2023, November). Gender bias and stereotypes in large language models. *Proceedings of the ACM Collective Intelligence Conference* (pp. 12-24).

Lakoff, R. (1973). Language and woman's place. *Language in Society*, *2*(1), *45-79.*

LIWC. (n.d.). *LIWC analysis: Understanding the LIWC summary measures*. LIWC.

MRXPros, MFour, & Iris Flex. (2025). *MRXPros omnibus survey: U.S. consumer perceptions of AI* [Unpublished survey dataset]. MRXPros

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*(5), 665–675.

Soundararajan, S., & Delany, S. J. (2024, October). Investigating gender bias in large language models through text generation. *Proceedings of the 7th International Conference on Natural Language and Speech Processing* (ICNLSP 2024) (pp. 410-424).

Torres, N., Ulloa, C., Araya, I., Ayala, M., & Jara, S. (2025). A comprehensive analysis of gender, racial, and prompt-induced biases in large language models. *International Journal of Data Science and Analytics, 20, 3797-3894.*

Authors:

Idil Cakim is a strategist who has shaped marketing, communications and audience insights for Fortune 500 brands and nonprofits. With a background in behavioral research, audience insights and media analytics, she analyzes how people engage with content and technology. Idil is the Founder and CEO of Iris Flex, an intelligence company, and previously held senior leadership roles at Audacy, Nielsen, NM Incite, Golin, and Burson.

Tracy Adams is Senior Director of Research & Insights at the ARF, where she leads work on attention, AI and retail media networks. With a background in sociology and cultural studies, she brings a critical, interdisciplinary lens to advertising research. Tracy focuses on how technology and culture shape consumer behavior and media impact.

Samantha Zhang is a Senior Data Scientist on the Research team at the ARF, studying topics including attention measurement, artificial intelligence and consumer privacy. She has previously worked as a data scientist in the technology and media industries, hoping to bridge the gap between quantitative research and business insights.

Keith Smith is Managing Director at the Marketing Science Institute (MSI), where he leads research and collaboration at the intersection of marketing theory and practice. With experience spanning academic research, teaching and industry consulting, he focuses on connecting scholars and business leaders to generate insights that advance both knowledge and impact.