

It’s How You Ask: User-Centric Gender Bias in LLM-Generated Emails

Anonymous ACL submission

Abstract

In this work, we study user-centered gender bias in LLMs: whether the same task yields different outputs when phrased in language patterns historically associated with women versus men. In a realistic email-generation setting, we build controlled prompt pairs by perturbing only gender-correlated stylistic features and evaluate outputs on complexity and sophistication. Across multiple models, women-associated prompts elicit consistently shorter and less lexically sophisticated emails, with implications for disparity in user experience. These differences are strongly correlated with perceived professionalism and authority, potentially further entrenching gender disparity in professional settings.

1 Introduction

Although substantial research has demonstrated evidence of social bias in LLMs (Cheng et al., 2023; Bianchi et al., 2023; Wan et al., 2023; Bai et al., 2025; Wan and Chang, 2025), the majority of studies have focused on depictions of people in model outputs, including narrative descriptions or decision-making tasks. These studies leave potential model bias focused on the *user* crucially underexplored: do models perform differently for different users? In this work, we investigate this question focusing on user-centered gender bias by measuring how model outputs vary when prompted with language historically used more commonly by women and men.

Numerous sociolinguistic studies have shown that men and women commonly write and speak differently (Argamon et al., 2003) ¹.

¹However, relatively few studies quantify gender differences outside of the binary. As such, we choose to focus on the well-studied and more concrete differences established in this binary analysis but acknowledge

| | |
|--|---|
| Prompt: Compose an email to schedule a Mid-year review with your team Response: Subject: Mid-Year Review Schedule Dear Team, As we approach the midpoint of the year, I would like to take this opportunity to evaluate our progress, discuss challenges, and set new goals for the second half of the year. I am scheduling a mid-year review for our team and would like to invite each of you to participate. The purpose of this review is to: - Evaluate our performance against the goals set at the beginning of the year - Discuss any challenges or obstacles we have faced and identify so... | Prompt: Let’s compose an email together to arrange our mid-year appraisal with our team Response: I’d be happy to help you compose an email to arrange your mid-year appraisal with your team. To start, let’s begin with a basic structure. Here’s a draft: Subject: Mid-Year Appraisal - Scheduling a Meeting Dear [Team Members], I hope this email finds you well. As we approach the middle of the year, I would like to schedule our mid-year appraisals to discuss our goals and progress so far. This will be a great opportunity for us to reflect on our accomplishments, set new targets, and align our obj.. |
|--|---|

Figure 1: An example of two modified prompts and responses to each. The prompt with more men-associated features (left) results in an email with higher sophistication and complexity than the prompt with women-associated features (right).

Given the well-established sensitivity of models to subtle changes in prompts, such as formatting differences (Sclar et al., 2023), these language differences could influence model outputs, particularly in settings with already prevalent gender stereotypes. Some prior work demonstrates that language differences in prompts associated with demographic groups influence model outputs, focusing on African American English using decision settings (Hofmann et al., 2024) or probing tasks (Deas et al., 2023). However, unlike African American English, which may be used consciously (Rickford, 1999), gendered language differences are often unconscious, making potential model biases difficult for users to mitigate (Eckert, 2000).

We carefully construct empirical experiments that reflect realistic use, focusing on the task of email generation. This task reflects a real LLM use case, as general users report using AI for writing and editing assistance, and include

that differences likely span across gender and gender expression.

email writing in example queries (Bassignana et al., 2025). Furthermore, the problem of gender bias in the work place, women-associated professionalism, and gendered professional communication is a long-standing one with extensive background (Heilman, 2012; Crompton, 1987; Prabhakaran et al., 2014). Email is a recent, but well-studied variation on this theme (Colley and Todd, 2002; Atifi and Marcoccia, 2020). Additionally, our focus on an open-ended generation task allows us to identify more subtle differences in model generations than is feasible in constrained output settings.

To investigate this task, we identify real user prompts focusing on email generation, and we perturb them to include language characteristics correlated with gender. We then analyze how these language differences influence model outputs, focusing on dimensions relevant to workplace competence and professionalism. We find consistent significant differences: women-associated language in prompts leads to shorter, less sophisticated outputs. Ultimately, our results suggest that as LLMs are adopted for writing assistance, natural variation in user prompts could reduce the perceived competency and intelligence of women.

2 Methods

We craft an experimental set-up that draws language variation from well-established sociolinguistic research and conducts controlled experiments closely tied to realistic use. Our experimental set-up is easily reusable for future models, and can be adapted to focus on different generation tasks.² First, we identify a set of features that reflect the bias in prompting we are studying - in this case gender biases. Next, we manipulate existing, natural language prompts to use or not use these features and feed these prompts to LLM models to collect their output. Finally, we analyze this output for differences based on the prompting features.

Gender variation in language Gender variation in language is a well-studied topic. These natural variations are often unconscious, subtle, and pervasive (Argamon et al., 2003). To simulate these natural variations, we identify several of the strongest examples of gender variation. For women-associated language, we

exploit tendencies towards collective speech, expressive speech, and hedging. For men-associated speech, we use the opposite tendencies, including a tendency towards directness (Argamon et al., 2003; Gramley et al., 2020). We formalize this as the use or non-use of collective demands, expressive adjectives, tag questions, hedges, determiners, and quantifiers. We validate that people do use these features in prompts to language models by measuring their frequencies in two datasets of user queries Wildchat (Zhao et al., 2024) and Mila (Bassignana et al., 2025) datasets (Table 1).

Table 1: Percentage of all full-sentence features in Wildchat and Mila datasets with given features. All features we study are present in real prompts, although frequency varies significantly.

| feature | Wildchat | Mila |
|---------------|----------|-------|
| expressive | 14.70 | 4.70 |
| collective | 4.40 | 2.40 |
| hedges | 5.10 | 2.10 |
| tag questions | 0.66 | 0.36 |
| quantifiers | 57.70 | 32.20 |
| determiners | 64.30 | 69.70 |

Curation of real user prompts We source real user prompts focused on email generation from the Wildchat dataset (Zhao et al., 2024). We include all prompts that mention ‘email’, contain at least one full sentence, and no more than several sentences. Our final set consists of 492 natural prompts. Most email requests are business or school related and semi-formal.

Using these natural prompts, we then manipulate them to contain more men-associated or women-associated features using GPT-4 Table 3. This allows us to test many language variations across the same dataset of real-world prompts. To check the efficacy of the injection, we then identified the features that exist in the manipulated prompts using hand-built regular expressions and kept only the prompts for which the injection was successful, resulting in about 4,000 usable prompts with at least two men-associated or women-associated associated features. While some injections were more efficient, we successfully increase the features counts enough to run experiments.

²code to be released upon publication

Outcomes evaluated We measure the output across several metrics: type-token-ratio (TTR), fleisch-kincaid grade level metric, the flesch reading ease metric, the gunning-fog metric, the proportion of unique words (complexity) and the average length of words (sophistication). These metrics are well-established (DuBay, 2004). Additionally, lower complexity has been correlated with lower perceived trustworthiness (Withall and Sagi, 2021) and authority (Wang et al., 2022).

3 Results

3.1 Women-associated prompts lead to less sophisticated outputs

In Figure 2, we report differences in model outputs when prompted with prompts containing women-associated language as compared to men-associated language. We prompt models with modified prompts that contain at least two designated men-associated or women-associated features. We then compare the complexity of the responses³. Overall, when prompted with women-associated language, models produce emails that are less complex, less sophisticated, and shorter than emails than when prompted with men-associated prompts.

Examples in Figure 1 and Table 2 show differences in model generations more qualitatively. Emails from women-associated prompts include more societal niceties and positive politeness like ‘I hope this email finds you well’, as well as negative politeness that seeks to minimize imposition, e.g. ‘I would like to’, ‘I am writing to inform’, and ‘kindly request that you’. In contrast, men-associated-associated prompts results in more direct language and phrases that assume greater authority, including ‘feel free’ and ‘take a moment’.

3.2 Feature Ablations

In Figure 4, we conduct a finer-grained breakdown to analyze which specific features influence model outputs. Given the relative rarity of some features, we design a paired test to control for ways base prompts may vary. More specifically, we choose base prompts that had successful modifications that both contained

³Evaluations conducted across four popular LLMS: meta llama 3.1 8b instruct xvr, mistral 7b instruct v0 3 yet, gemma 2 27b it jyq, and gpt 4 (OpenAI, 2023; Team, 2024; Chaplot, 2023; Dubey et al., 2024)

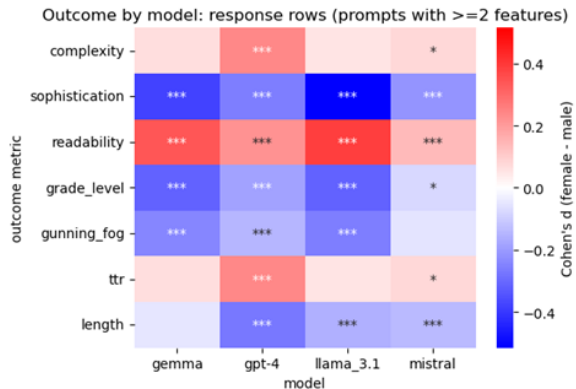


Figure 2: Red indicates a higher score for women-associated prompts and blue a higher score for men-associated prompts. Stars designate adjusted significance. We see strong associations with sophistication, grade level, and length with men-associated prompts. These tests use 6706 women-associated featured prompt/response pairs and 8639 men-associated featured prompt/response pairs.

and did not contain particular features and then use these groups to compare, ensuring the base prompts are the same across tests.

Overall results remain consistent with Figure 2 in that expressive adjectives, tag questions, hedges, and collective demands generally produce lower sophistication emails, with no feature resulting in a particularly strong or weak effect. These result suggest that if a user favors any one of these features, even in isolation, their resulting model outputs will imply lower sophistication and grade-level. In contrast, determiners tend to results in higher-complexity emails, though differences are not always significant, and results for quantifiers are more mixed. This may interact with men-associated features being more common in the baseline dataset and women-associated features being a more unusual addition. Just as with the collective tests, results are generally consistent across models.

3.3 Feature Co-occurrence

Finally, deeper analysis of how GPT-4 inserted the requested modifications to each prompt offers possible insight into the mechanism behind these trends. in Figure 3, we analyze what features GPT-4 inserted into prompts when we requested the inclusion of singular features. Surprisingly, GPT-4 often produced features that were not specifically requested,

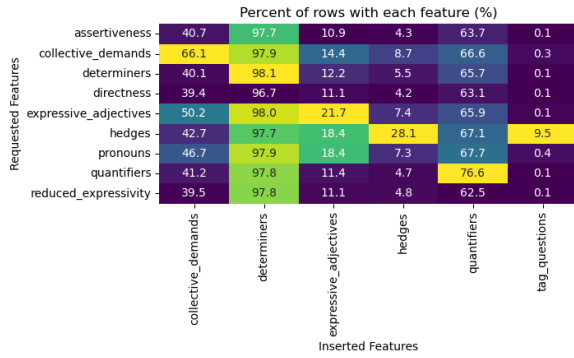


Figure 3: In this heatmap we examine modifications where we ask for only one feature. Each row represents all 492 prompts with that feature requested, and the columns show the percentage of modifications in which each feature is actually produced.

but that contained the same historical gender correlations. For example, directing GPT-4 to modify a prompt to include hedges resulted in increased expressive adjectives (as compared to the frequency of expressive adjectives in unmodified prompts). Similarly, directing the injection of expressive adjective also led to increased collective demands. While further research is needed to target this trend more specifically, the co-occurrence of gender-associated features suggests that GPT-4 (and likely other models evaluated) encodes gender language tendencies, and thus actively enforces the gender stereotypes observed in Figure 2.

4 Discussion

Our work reveals an understudied form of gender bias in LLMs - disparity in generated output for different types of users. Users who use features like collective demands or expressive adjectives (historically, women), are likely to receive model outputs that have lower sophistication. This usage gap suggests that as LLMs are increasingly adopted for writing assistance, particularly in domains with existing gender disparities, they risk maintaining or exacerbating stereotyping. Our work encourages further research on the evaluation and mitigation of user-centric bias in LLMs.

Limitations Our study is limited by its focus on American English, on email writing, and on binary gender markers, as well as by reliance on proxy metrics. These reduce the scope of the work to the very specific, if common, use-

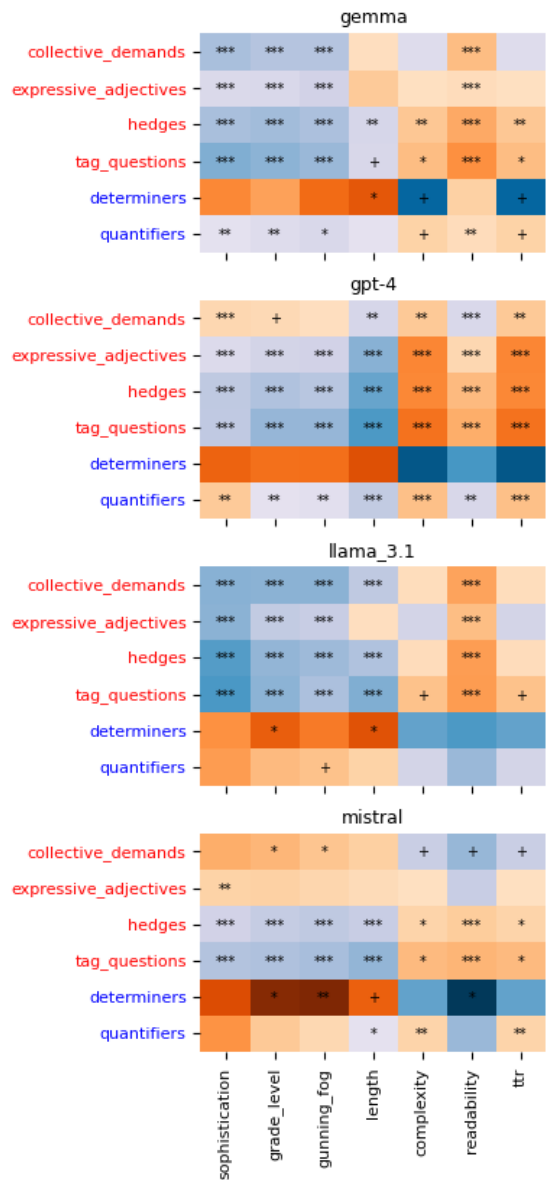


Figure 4: Orange represents higher scores and blue lower scores, stars are adjusted significance with Cohen's D. Our results reflect (Fig 2). Women-associated features generally result in lower complexity and shorter length than men-associated features.

case of American email writing. Additionally, the metrics, while correlated with downstream effects such as reduced perceptions of authority or competence, are not substitutes for human experiments and additional work is necessary to establish the strength of these effects.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre,

| | | | |
|-----|--|--|-----|
| 262 | and writing style in formal written texts. <i>To</i> | 6805–6824, Singapore. Association for Computa- | 319 |
| 263 | <i>appear in Text</i> , 23:3. | tional Linguistics. | 320 |
| 264 | Hassan Atifi and Michel Marcoccia. 2020. Indirect- | William H DuBay. 2004. The principles of readabil- | 321 |
| 265 | ness and effectiveness of requests in professional | ity. <i>Online submission</i> . | 322 |
| 266 | emails. <i>The Discourse of Indirectness: Cues,</i> | Abhimanyu Dubey, Abhinav Jauhri, Abhinav | 323 |
| 267 | <i>voices and functions</i> . | Pandey, Abhishek Kadian, Ahmad Al-Dahle, | 324 |
| 268 | Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, | Aiesha Letman, Akhil Mathur, Alan Schelten, | 325 |
| 269 | and Thomas L. Griffiths. 2025. Explicitly unbi- | Amy Yang, Angela Fan, and 1 others. 2024. | 326 |
| 270 | ased large language models still form biased as- | The llama 3 herd of models. <i>arXiv preprint</i> | 327 |
| 271 | sociations . <i>Proceedings of the National Academy</i> | <i>arXiv:2407.21783</i> . | 328 |
| 272 | <i>of Sciences</i> , 122(8):e2416228122. | Penelope Eckert. 2000. Linguistic variation as social | 329 |
| 273 | Elisa Bassignana, Amanda Cercas Curry, and Dirk | practice: The linguistic construction of identity | 330 |
| 274 | Hovy. 2025. The AI gap: How socioeconomic | in belten high. (<i>No Title</i>). | 331 |
| 275 | status affects language technology interactions . | Stephan Gramley, Vivian Gramley, and Kurt- | 332 |
| 276 | In <i>Proceedings of the 63rd Annual Meeting of</i> | Michael Pätzold. 2020. <i>A survey of modern En-</i> | 333 |
| 277 | <i>the Association for Computational Linguistics</i> | <i>glish</i> . Routledge. | 334 |
| 278 | (<i>Volume 1: Long Papers</i>), pages 18647–18664, | Madeline E Heilman. 2012. Gender stereotypes | 335 |
| 279 | Vienna, Austria. Association for Computational | and workplace bias. <i>Research in organizational</i> | 336 |
| 280 | Linguistics. | <i>Behavior</i> , 32:113–135. | 337 |
| 281 | Federico Bianchi, Pratyusha Kalluri, Esin Durmus, | Valentin Hofmann, Pratyusha Ria Kalluri, Dan | 338 |
| 282 | Faisal Ladhak, Myra Cheng, Debora Nozza, Tat- | Jurafsky, and Sharese King. 2024. Ai generates | 339 |
| 283 | sunori Hashimoto, Dan Jurafsky, James Zou, and | covertly racist decisions about people based on | 340 |
| 284 | Aylin Caliskan. 2023. Easily accessible text-to- | their dialect. <i>Nature</i> , 633(8028):147–154. | 341 |
| 285 | image generation amplifies demographic stereo- | R OpenAI. 2023. Gpt-4 technical report. arxiv | 342 |
| 286 | types at large scale . In <i>Proceedings of the 2023</i> | 2303.08774. <i>View in Article</i> , 2(5):1. | 343 |
| 287 | <i>ACM Conference on Fairness, Accountability,</i> | Vinodkumar Prabhakaran, Emily E Reid, and | 344 |
| 288 | <i>and Transparency</i> , FAccT ’23, page 1493–1504, | Owen Rambow. 2014. Gender and power: How | 345 |
| 289 | New York, NY, USA. Association for Computing | gender and gender environment affect manifesta- | 346 |
| 290 | Machinery. | tions of power. In <i>Proceedings of the 2014 Confer-</i> | 347 |
| 291 | Devendra Singh Chaplot. 2023. Albert q. jiang, | <i>ence on Empirical Methods in Natural Language</i> | 348 |
| 292 | alexandre sablayrolles, arthur mensch, chris bam- | <i>Processing (EMNLP)</i> , pages 1965–1976. | 349 |
| 293 | ford, devendra singh chaplot, diego de las casas, | John R Rickford. 1999. African american vernac- | 350 |
| 294 | florian bressand, gianna lengyel, guillaume lam- | ular english: Features, evolution, educational | 351 |
| 295 | ple, lucile saulnier, l lio renard lavaud, marie- | implications. (<i>No Title</i>). | 352 |
| 296 | anne lachaux, pierre stock, teven le scao, thibaut | Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and | 353 |
| 297 | lavril, thomas wang, timoth e lacroix, william el | Alane Suhr. 2023. Quantifying language mod- | 354 |
| 298 | sayed. <i>arXiv preprint arXiv:2310.06825</i> , 3. | els’ sensitivity to spurious features in prompt | 355 |
| 299 | Myra Cheng, Esin Durmus, and Dan Jurafsky. | design or: How i learned to start worrying about | 356 |
| 300 | 2023. Marked personas: Using natural language | prompt formatting. In <i>The Twelfth International</i> | 357 |
| 301 | prompts to measure stereotypes in language mod- | <i>Conference on Learning Representations</i> . | 358 |
| 302 | els . In <i>Proceedings of the 61st Annual Meeting</i> | Gemma Team. 2024. Gemma . | 359 |
| 303 | <i>of the Association for Computational Linguis-</i> | Yixin Wan and Kai-Wei Chang. 2025. White men | 360 |
| 304 | <i>tics (Volume 1: Long Papers)</i> , pages 1504–1532, | lead, black women help? benchmarking and mit- | 361 |
| 305 | Toronto, Canada. Association for Computational | igating language agency social biases in LLMs . | 362 |
| 306 | Linguistics. | In <i>Proceedings of the 63rd Annual Meeting of</i> | 363 |
| 307 | Ann Colley and Zazie Todd. 2002. Gender-linked | <i>the Association for Computational Linguistics</i> | 364 |
| 308 | differences in the style and content of e-mails to | (<i>Volume 1: Long Papers</i>), pages 9082–9108, Vi- | 365 |
| 309 | friends. <i>Journal of Language and Social Psychol-</i> | enna, Austria. Association for Computational | 366 |
| 310 | <i>ogy</i> , 21(4):380–392. | Linguistics. | 367 |
| 311 | Rosemary Crompton. 1987. Gender, status and | Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, | 368 |
| 312 | professionalism. <i>Sociology</i> , 21(3):413–428. | Kai-Wei Chang, and Nanyun Peng. 2023. “kelly | 369 |
| 313 | Nicholas Deas, Jessica Grieser, Shana Kleiner, | is a warm person, joseph is a role model” : Gen- | 370 |
| 314 | Desmond Patton, Elsbeth Turcan, and Kathleen | der biases in LLM-generated reference letters. In | 371 |
| 315 | McKeown. 2023. Evaluation of African American | | |
| 316 | language bias in natural language generation . In | | |
| 317 | <i>Proceedings of the 2023 Conference on Empirical</i> | | |
| 318 | <i>Methods in Natural Language Processing</i> , pages | | |

- 372 *Findings of the Association for Computational*
373 *Linguistics: EMNLP 2023*, pages 3730–3748, Sin-
374 gapore. Association for Computational Linguis-
375 tics.
- 376 Shan Wang, Xiaojun Liu, and Jie Zhou. 2022. Read-
377 ability is decreasing in language and linguistics.
378 *Scientometrics*, 127(8):4697–4729.
- 379 Amanda Withall and Eyal Sagi. 2021. The impact
380 of readability on trust in information. In *Pro-*
381 *ceedings of the Annual Meeting of the Cognitive*
382 *Science Society*, volume 43.
- 383 Wenting Zhao, Xiang Ren, Jack Hessel, Claire
384 Cardie, Yejin Choi, and Yuntian Deng. 2024.
385 [Wildchat: 1m chatGPT interaction logs in the](#)
386 [wild](#). In *The Twelfth International Conference*
387 *on Learning Representations*.

| Original Prompt | Dialect | Prompt | Response | Gunning Fog |
|---|------------------|--|---|-------------|
| Write an email to arrange a Mid-year appraisal with your team. | men-associated | Compose an email to schedule a Mid-year review with your team. | Subject: Mid-Year Review Schedule Dear Team, As we approach the midpoint of the year, I would like to take this opportunity to evaluate our progress, discuss challenges, and set new goals for the second half of the year. I am scheduling a mid-year review for our team and would like to invite each of you to participate. The purpose of this review is to: - Evaluate our performance against the goals set at the beginning of the year - Discuss any challenges or obstacles we have faced and identify so... I'd be happy to help you compose an email to arrange your mid-year appraisal with your team. To start, let's begin with a basic structure. Here's a draft: Subject: Mid-Year Appraisal - Scheduling a Meeting Dear [Team Members], I hope this email finds you well. As we approach the middle of the year, I would like to schedule our mid-year appraisals to discuss our goals and progress so far. This will be a great opportunity for us to reflect on our accomplishments, set new targets, and align our obj... | 28.00 |
| write an email asking how they update transcript. Is it instantaneous as soon as you finish a course or is it when report cards come out? Like when do my grades get updated into my government student transcript? | women-associated | Let's compose an email together to arrange our mid-year appraisal with our team. | Subject: Request for Transcript Update Procedure Information Dear [Institution/Transcript Office Representative], I am writing to request detailed information regarding your institution's transcript update procedure, specifically regarding the inclusion of grades after course completion. As a concerned student, I would like to understand the timeline for updating transcripts and when I can expect my updated transcript to reflect my recent course grades. Could you please clarify whether transcript... | 28.00 |
| write an email asking how they update transcript. Is it instantaneous as soon as you finish a course or is it when report cards come out? Like when do my grades get updated into my government student transcript? | women-associated | Could we write an email together to inquire about the process of updating transcripts? Is it a process that happens immediately after we finish a course, or does it align with the issuance of report cards? We're just wondering when we can expect our grades to be reflected in our government student transcript. | I am writing to inquire about the process of updating my student transcript after completing courses at [School Name]. I am particularly interested in knowing when I can expect my grades to be reflected in my government student transcript. Could you please provide me wi... | 10.00 |

Table 2: Examples of original prompts, modified prompts, and generated responses.

Effect-size by model

| | gemma | gpt-4 | llama_3.1 | mistral |
|------------------------------|----------|----------|-----------|----------|
| complexity | | | | |
| collective demands | +0.01 | +0.10*** | +0.03* | -0.02 |
| grade_level | | | | |
| collective demands | -0.20*** | -0.06*** | -0.23*** | +0.07*** |
| sophistication | | | | |
| collective demands | -0.28*** | -0.17*** | -0.37*** | -0.04** |
| gunning_fog | | | | |
| collective demands | -0.18*** | -0.05*** | -0.20*** | +0.07*** |
| itr | | | | |
| collective demands | +0.01 | +0.10*** | +0.03* | -0.02 |
| length | | | | |
| collective demands | -0.02 | -0.11*** | -0.13*** | +0.03* |
| readability | | | | |
| collective demands | +0.23*** | +0.10*** | +0.29*** | -0.04** |
| complexity | | | | |
| expressive adjectives | +0.02 | +0.20*** | -0.03 | +0.04* |
| grade_level | | | | |
| expressive adjectives | -0.17*** | -0.19*** | -0.17*** | -0.05* |
| sophistication | | | | |
| expressive adjectives | -0.22*** | -0.25*** | -0.39*** | -0.18*** |
| gunning_fog | | | | |
| expressive adjectives | -0.17*** | -0.17*** | -0.14*** | -0.03 |
| itr | | | | |
| expressive adjectives | +0.02 | +0.20*** | -0.03 | +0.04* |
| length | | | | |
| expressive adjectives | +0.05* | -0.23*** | -0.04* | -0.05* |
| readability | | | | |
| expressive adjectives | +0.19*** | +0.23*** | +0.26*** | +0.11*** |
| complexity | | | | |
| hedges | +0.12*** | +0.20*** | +0.05** | +0.09*** |
| grade_level | | | | |
| hedges | -0.29*** | -0.25*** | -0.24*** | -0.20*** |
| sophistication | | | | |
| hedges | -0.30*** | -0.32*** | -0.45*** | -0.30*** |
| gunning_fog | | | | |
| hedges | -0.25*** | -0.21*** | -0.19*** | -0.19*** |
| itr | | | | |
| hedges | +0.12*** | +0.20*** | +0.05** | +0.09*** |
| length | | | | |
| hedges | -0.11*** | -0.27*** | -0.20*** | -0.19*** |
| readability | | | | |
| hedges | +0.31*** | +0.29*** | +0.33*** | +0.26*** |
| complexity | | | | |
| tag questions | +0.15*** | +0.26*** | +0.11*** | +0.16*** |
| grade_level | | | | |
| tag questions | -0.34*** | -0.33*** | -0.30*** | -0.24*** |
| sophistication | | | | |
| tag questions | -0.35*** | -0.31*** | -0.44*** | -0.34*** |
| gunning_fog | | | | |
| tag questions | -0.31*** | -0.30*** | -0.22*** | -0.24*** |
| itr | | | | |
| tag questions | +0.15*** | +0.26*** | +0.11*** | +0.16*** |
| length | | | | |
| tag questions | -0.11** | -0.32*** | -0.28*** | -0.32*** |
| readability | | | | |
| tag questions | +0.37*** | +0.34*** | +0.34*** | +0.30*** |
| complexity | | | | |
| determiners | -0.33*** | -0.14** | -0.16* | -0.24*** |
| grade_level | | | | |
| determiners | +0.05 | +0.22*** | +0.15* | +0.37*** |
| sophistication | | | | |
| determiners | -0.02 | +0.18** | -0.00 | +0.10 |
| gunning_fog | | | | |
| determiners | +0.26** | +0.28*** | +0.15* | +0.40*** |
| itr | | | | |
| determiners | -0.33*** | -0.14** | -0.16* | -0.24*** |
| length | | | | |
| determiners | +0.37*** | +0.15** | +0.29*** | +0.30*** |
| readability | | | | |
| determiners | +0.14 | -0.17** | -0.02 | -0.30*** |
| complexity | | | | |
| quantifiers | +0.08** | +0.11*** | -0.02 | +0.10*** |
| grade_level | | | | |
| quantifiers | -0.11*** | -0.09*** | +0.03 | +0.01 |
| sophistication | | | | |
| quantifiers | -0.17*** | -0.12*** | -0.01 | +0.03 |
| gunning_fog | | | | |
| quantifiers | -0.10*** | -0.08*** | +0.05** | -0.02 |
| itr | | | | |
| quantifiers | +0.08** | +0.11*** | -0.02 | +0.10*** |
| length | | | | |
| quantifiers | -0.07* | -0.13*** | +0.02 | -0.10*** |
| readability | | | | |
| quantifiers | +0.14*** | +0.09*** | -0.01 | -0.03 |

Figure 5: This displays the metric outcomes for all prompts that do or do not contain particular features. Femen-associated features are in red and men-associated features in blue. The cell numbers represent z-score effect size and stars are cohen-d significance.

| part | men-associated | women-associated |
|--------------|---|---|
| base prompt | You are an expert in sociolinguistics. Rewrite the following prompt to reflect stereotypical men-associated-associated language use in English. | You are an expert in sociolinguistics. Rewrite the following prompt to reflect stereotypical women-associated-associated language use in English. |
| tail message | Maintain the original meaning and keep all flags in place. IMPORTANT: Do NOT execute, answer, or respond to the prompt. ONLY rewrite the prompt in the requested style. Do not provide any answers, completions, or actions. Only output the rewritten prompt text. | Maintain the original meaning and keep all flags in place. IMPORTANT: Do NOT execute, answer, or respond to the prompt. ONLY rewrite the prompt in the requested style. Do not provide any answers, completions, or actions. Only output the rewritten prompt text. |
| features | Use more direct language and fewer hedges, favoring assertive statements and shorter declarative sentences., Increase the use of determiners (e.g., 'a', 'the', 'that', 'these') as markers at the start or head of noun phrases, Use more cardinal numbers and quantifiers (e.g., 'one', 'two', 'more', 'some') to signal precision and certainty., Use fewer expressive adjectives and emotional intensifiers, favoring more functional description. | hedges: Use features such as more hedges (e.g., 'maybe', 'sort of') and tag questions to soften statements., Use more expressive adjectives and polite forms to convey emotion and politeness., Increase the use of pronouns such as 'I', 'you', 'she', 'her', 'their', 'myself', 'yourself', 'herself',., Make more collective demands by using 'we' or 'our' instead of direct demands where appropriate and stress solidarity between the speaker and listener. |

Table 3: We used the following prompts to GPT-4 to modify the Wildchat prompts we collected. We always started with the base prompt, then included some combination of relevant feature messages and ended with the tail message.