Contrast Comes at a Cost: Prompting Large Language Models for Perception-Centered Gendered Style Transfer

Anonymous ACL submission

Abstract

Gendered style transfer aims to rewrite text toward linguistic styles commonly associated with feminine or masculine language use while preserving the original meaning. Prior work commonly operationalizes the transferred target style through binary author identity labels, conflating who writes a text with how gender expression is conveyed stylistically in the text itself. In contrast, we focus on *perceived* gendered style: the qualities attributed to a text by readers, independently of the author's identity. We study this as a controlled contrastive rewriting problem: given the same source text, can different prompt-based operationalizations induce feminine- and masculine-targeted rewrites, and at what semantic cost? We compare four operationalizations across instruction-tuned LLMs: author identity, authorial voice, general style descriptions, and fine-grained sociolinguistically grounded style instructions, and evaluate them against human-written rewrites. Using automatic metrics, statistical tests, and human evaluation, we find that LLM-generated rewrites often show stronger contrast and higher meaning preservation than human-written ones. Moreover, prompt operationalization systematically shapes the trade-off between perceived stylistic contrast and semantic fidelity. These findings provide practical guidance for prompt design in gendered style transfer and suggest that targeting perceived style with LLMs can yield clearer linguistic signals than relying solely on author-identity-based human rewrites.[1]

1 Introduction

Gendered style transfer is the task of rewriting a text so that it preserves the core meaning while shifting the linguistic cues associated with feminine or masculine expression (Jin et al., 2022; Kang et al., 2019; Sudhakar et al., 2019). For example, a sentence such as "This café is nice and I enjoyed the atmosphere" might be rewritten in a more descriptive style, which previous work associates more strongly with feminine language use, e.g., "I really loved the cozy atmosphere of this café", or in a more direct style, commonly linked to masculine communication styles, e.g., "The café has a pleasant atmosphere" (Bamman et al., 2014; Preoţiuc-Pietro et al., 2017; Baker, 2014). Developing methods that generate such meaning-preserving rewrites is important both for practical applications, where writers may wish to adapt their language to fit a particular discourse context, and for research examining how gendered language influences readers' judgments of credibility or persuasiveness.

In existing NLP research, *gendered style* is typically operationalized through binary author identity labels (e.g., male vs. female authors), framing the transfer task as converting between texts written by authors in these categories (Kang et al., 2019; Sudhakar et al., 2019). However, author identity and gendered linguistic expression are distinct constructs (Bucholtz, 2002; Butler, 1990b; Eckert and McConnell-Ginet, 2013). Thus, transferring between male-authored and female-authored texts does not necessarily correspond to transferring between masculine and feminine expression: the gendered qualities of a text may be interpretable independently of the author's identity (Troiano et al., 2023). To this end, Chen et al. (2025) introduced the notion of *perceived gendered style*, which measures how readers interpret the gendered qualities of a text rather than relying on author attributes. From this perspective, gendered style transfer can be framed as *rewriting a text so that it is* ***more likely to be perceived as feminine or masculine*** *while preserving its core meaning*. However, this perception-centered formulation has not yet been explored for style transfer, and it remains an open question whether perceived gendered style can be

[1]The datasets and experimental code used in this study are available at https://anonymous.4open.science/r/GST-LLMs-5DDC

systematically induced through text rewriting.

In this work, we present, to the best of our knowledge, the first empirical study of perception-centered gendered style transfer as a controlled contrastive generation problem. Building on recent advances in generating controlled rewrites with Large Language Models (LLMs) (Lu et al., 2023; Sun et al., 2023; Moskovskiy et al., 2024; Han et al., 2024), we investigate whether prompting LLMs can be used to perform perception-centered gendered style transfer. We address the following research questions:

RQ1 Can LLMs produce rewrite pairs that are perceived as contrastively feminine vs. masculine?

RQ2 Which prompting operationalization yields the most reliable perceived gendered contrast?

To answer these questions, we prompt several instruction-tuned LLMs to produce paired rewrites of the same input text under four theoretically motivated operationalizations of gender: *author identity* (e.g., writing as a woman or a man), *gendered voice*, *general gendered style*, and *fine-grained style instructions grounded in sociolinguistic features of gendered language variation*. We compare these LLM-generated pairs to human rewrite pairs from authors with different gender identities (Kang et al., 2019), reflecting the identity-based operationalization commonly used in gendered style transfer. We evaluate the rewrites along three dimensions central to style transfer: gendered style contrast, meaning preservation, and grammatical acceptability, using automatic metrics for the full set of generations and human evaluation on a subset. Our results show that LLMs can generate rewrite pairs that differ in perceived gendered style while largely preserving meaning (§4). Moreover, the success of this process depends strongly on how gender is operationalized in the prompt (§5): prompts with fine-grained stylistic instructions produce the strongest perceived contrast between feminine and masculine rewrites, often exceeding the contrast observed in human rewrite pairs. In contrast, prompts framed around gender identity yield substantially weaker stylistic differences but preserve meaning more reliably. These findings reveal a clear trade-off between stylistic contrast and semantic fidelity, two central requirements of style transfer.

Overall, our findings show that how gendered style is operationalized in prompts is crucial for controlling stylistic variation in generated text. For style-transfer applications, this offers practical guidance on prompt design. For research on the social interpretation of linguistic expression, where contrastive rewrites are needed for controlled experiments, our results suggest that targeting perceived gendered style and using LLM-generated data can yield clearer signals than rewrites from authors with different gender identities.

2 Related Work

Gender Operationalization in Text Gender is a socially constructed concept that comprises both *gender identity* and *gender expression* (Bucholtz, 2002). Gender identity concerns how individuals identify themselves (e.g., woman, man, non-binary), whereas gender expression concerns how gender is presented or performed externally (Baum and Westheimer, 2015; Ehrensaft, 2018; Pinney et al., 2023), e.g., feminine, masculine, or gender-neutral. While expression focuses on individual presentation, feminist theory further posits that gender is performative: it is produced through repeated social acts and norms rather than being a fixed, internal essence (Butler, 1990a, 2002, 2004).

In language, gender can be operationalized in multiple ways. Computational work often relies on *author-level labels*, using them to identify correlations between gender identity and stylistic variation in text (Morales Sánchez et al., 2022; Preoţiuc-Pietro et al., 2016). However, such correlations are fluid and domain-dependent rather than stable linguistic signatures (Bamman et al., 2014; Chen et al., 2024). Recent work therefore distinguishes between several more nuanced constructions: *gendered style*, shifting toward more feminine or masculine expression depending on the context (Troiano et al., 2023); *gendered voice*, which foregrounds stance and perspective (Saito, 2010); and *perceived gendered style*, focusing on how a text is socially read by others (Chen et al., 2025).

Style and Perceived Gendered Style Work on text style transfer has adopted two broad definitions of "style." In a linguistically grounded view, style refers to the manner in which semantics is expressed (McDonald and Pustejovsky, 1985). Variation is realized through lexical choice, syntactic constructions, discourse and stance markers, and surface cues (e.g., punctuation or hedging), and has been studied across a range of domains (Leech et al., 1992; Bergsma et al., 2012; Baker, 2014; Hu and Kearney, 2021; Dahllöf, 2023; Chen et al., 2024). In contrast, recent style transfer work adopts

a data-driven perspective that operationalizes style as a discrete attribute to be controlled (Pauli et al., 2025; Jin et al., 2022), such as sentiment or toxicity. In this framing, transfer is often implemented by editing attribute-bearing expressions (Li et al., 2018), which can be effective for control but may blur the boundary between "style" and "content."

The distinction between linguistic and attribute-based notions is particularly relevant for *gendered style*, which is typically expected to reflect differences in linguistic realization rather than changes to semantic content. More recent work further highlights that gendered style may be socially interpreted rather than reducible to author identity alone (Troiano et al., 2023; Nguyen et al., 2014). Building on this perspective, Chen et al. (2025) proposed perception-based operationalization of gendered style that treats it as a *reader-attributed property* of text, measured through human judgments of whether a text sounds more feminine, masculine, or neutral. This differs from earlier setups that asked annotators to infer *author gender* from writing (Nguyen et al., 2014; Flekova et al., 2016; Preoţiuc-Pietro et al., 2017), which can conflate gender expression with author identity. However, while Chen et al. (2025) provide a dataset with human annotations, they do not study how their perceptions can be controlled or manipulated through text rewriting.

LLM-based Style Transfer Prior work has explored LLM-based style transfer for a range of attributes, including personal or authorial style, formality, sentiment, and detoxification (La Quatra et al., 2024; Mai et al., 2023; Zhang et al., 2024; Reif et al., 2022). Most of this work relies on prompting techniques to guide stylistic transformation. The methods include standard zero-shot (Suzgun et al., 2022; Reif et al., 2022) and few-shot prompting, prompt selection or routing to choose effective instructions (Liu et al., 2024b), prompt-driven text editing (Luo et al., 2023), and hybrid pipelines where LLMs serve as supplementary components to augment pre-LLM style transfer systems (Toshevska and Gievska, 2025). Yet, gendered style transfer with explicit content preservation constraints remains underexplored, with most work continuing to rely on corpora grouped by author identity (Kang et al., 2019) rather than on controlled rewrites evaluated for perceived feminine/masculine contrast.

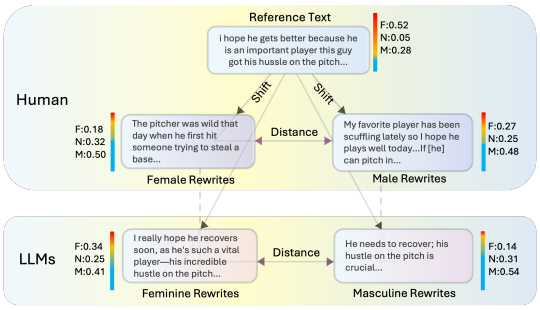

Figure 1: Methodology overview: a reference text (top) is rewritten into feminine and masculine variants by human authors (middle) or LLMs (bottom). A classifier assigns probability distributions over feminine (F), neutral (N), and masculine (M) styles, indicated by color bars. Style shift measures the rewrite's movement from the reference toward the target; style distance quantifies the contrast between paired rewrites.

3 Methodology

We define *gendered style transfer* as follows: given an input text t and a target style $s \in \{\text{feminine}, \text{masculine}\}$, the system produces a rewrite t' that preserves the core meaning of t while shifting linguistic realization toward s. Importantly, t may be neutral with respect to gendered cues; the task is therefore to induce stylistic variation without introducing new content.

In our experimental setup, each reference text t is paired with two rewrites targeting feminine and masculine style (see Figure 1 for an overview of the methodology). These rewrites come either from humans with different gender identities or from LLMs. We then evaluate these rewrite pairs with respect to perceived gendered style contrast, preservation of meaning, and grammatical acceptability. Below, we describe the data, models, prompting strategies, and evaluation metrics.

3.1 Data

For the reference texts and human rewrites, we use the **PASTEL dataset** (Kang et al., 2019). The dataset was designed to elicit persona-conditioned stylistic variation. Human annotators were first asked for their demographic information (including self-reported binary gender). Then, they were shown a sequence of images for an event, with accompanying keywords, and asked to write a five-sentence story describing the given context (images and *source texts* for keyword extraction were taken from the ViST dataset (Huang et al., 2016)). Because multiple annotators independently describe the same event, the dataset contains parallel stories

that convey the same underlying meaning while differing in linguistic realization.

We then sample 500 source texts from PASTEL as reference texts (see top in Figure 1). These texts serve as inputs for LLMs and as reference points for measuring style shift and meaning preservation; they are not intended to represent neutral stylistic baselines. To construct the *human-written* gender-contrasted pairs, we retrieve all rewrites associated with the same source text and randomly select one *female-authored* and one *male-authored* rewrite.[2]

3.2 Models

We use five instruction-tuned, open-source LLMs covering four widely used model families: Gemma-3-27B-it (Kamath et al., 2025), Gemma-2-27B-it (Riviere et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), Llama-3.1-8B-instruct (Grattafiori et al., 2024), and DeepSeek-V3 (Liu et al., 2024a). This selection allows us to test whether prompting effects are consistent across contemporary instruction-tuned model families and scales.

3.3 Prompts

Unlike PASTEL's multimodal setting, where annotators were given only images and keywords, we provide the full *source texts* from PASTEL as input to the models. This ensures that the content is fully specified, preventing the task from becoming content generation rather than style transfer. All LLMs are prompted with an instruction template (see §A.1) including the input reference text and a request to produce two rewrites that preserve the original meaning while contrasting in a specified gender-related target (depending on the operationalization explained below). Across all prompting conditions, we ask for the same hard constraints: (i) preserve meaning and factual content exactly; (ii) do not add new information, explanations, or opinions; (iii) avoid stereotypes or demographic assumptions (gender is framed as purely linguistic, not gender roles), and (iv) keep rewrites natural, fluent, and roughly comparable in length. Models are instructed to output two lines in a fixed format, with prefixes indicating the target condition, e.g., `Feminine/Masculine` style.

Gendered Style Operationalizations We test four prompting variants corresponding to the conceptual distinctions explained in §2 (see full prompts in Table 3 in §A.1): The *gender identity* setup mirrors the most common operationalization in prior work: the two rewrites are framed as being authored by "an author who identifies as a woman" versus "an author who identifies as a man." We compare this setup with several less explored alternatives. In *gendered voice*, we ask models to target "female-author voice tendencies" versus "male-author voice tendencies." In *gendered style*, the rewrites target "feminine style tendencies (linguistic)" versus "masculine style tendencies (linguistic)" without further specification. In the final condition, *style instruction*, we augment the prompt with explicit cue descriptions (e.g., hedging, directness, discourse framing) drawn from sociolinguistic observations about gendered language variation (Bamman et al., 2014; Eckert and McConnell-Ginet, 2013) and from Chen et al.'s (2025) annotation guidelines. This aims to clarify what counts as feminine vs. masculine linguistic variation while maintaining the meaning-preservation constraints.

Prompt Robustness We include three prompt robustness checks in §A.6: repeated generations with different random seeds, wording variants for the *style* and *style instruction* prompts, and length-matched versions of the *identity* and *voice* prompts that control for the greater length of the *style instruction* condition.

3.4 Evaluation

We evaluate human-written and LLM-generated rewrites with respect to perceived gendered style, meaning preservation, and grammatical acceptability. To provide insights into these metrics, we present qualitative examples along with their corresponding results in §A.7. We further validate these automatic metrics through human annotations on a subset of the generated rewrites in §6.

Perceived Gendered Style We use a perceived-gendered-style classifier trained on the dataset of Chen et al. (2025). The dataset contains 510 texts, each rated by 10 crowdworkers on a five-point feminine–masculine scale and accompanied by self-reported confidence scores. We collapse the style scale into three categories: F (very/somewhat feminine), N (neutral), and M (very/somewhat masculine), and train a confidence-weighted BERT-based classifier (Devlin et al., 2018) to predict a three-way probability distribution for a text t:

$$p(\cdot \mid t) = \big[p(F \mid t),\, p(M \mid t),\, p(N \mid t)\big].$$

[2] If multiple rewrites were available for a source text, we randomly sampled one per gender to form a single pair.

This classifier achieved the strongest overall performance among five evaluated models, with 0.62 accuracy and 0.20 mean Jensen-Shannon (JS) distance (full classifier comparisons are reported in Tables 4 and 5 in §A.2).

Style Distance (Pairwise Contrast) We compute style distance between two rewrites using the probability distributions produced by our classifier. For each text t, the classifier outputs a probability vector $p(\cdot \mid t) = [p(F \mid t), p(M \mid t), p(N \mid t)]$. Given a paired rewrite (t'_F, t'_M) for the same reference t, we quantify how strongly the two rewrites contrast in perceived gendered style using JS distance between their predicted distributions:

$$D_{\text{style}}(t'_F, t'_M) = \text{JS}\big(p(\cdot \mid t'_F) \,\|\, p(\cdot \mid t'_M)\big)\,.$$

Higher values indicate greater stylistic contrast between the feminine and masculine rewrites.

Style Shift (Relative to the Reference) For each reference text t and rewrite t', style shift is computed as the change in the classifier's predicted probability of the target style class from the reference to the rewrite. For example, the feminine shift is defined as the difference in feminine probability:

$$\Delta_F(t', t) = p(F \mid t') - p(F \mid t),$$

and analogously $\Delta_M(t', t) = p(M \mid t') - p(M \mid t)$. A larger positive Δ_s indicates a stronger shift toward style s relative to the reference, while a negative value indicates a shift away from s.

Meaning Preservation and Similarity We assess semantic consistency using a natural language inference (NLI) model. Specifically, we use WANLI (Liu et al., 2022) to estimate entailment probabilities for (i) **meaning similarity** within the rewrite pair (t'_F, t'_M) and (ii) **meaning preservation** of each rewrite relative to the reference t. For each pair (t'_F, t'_M), we report a *bidirectional entailment* score, defined as the minimum of the two directional entailment probabilities:

$$e(t'_F, t'_M) = p_{\text{NLI}}(\text{entail} \mid t'_F \rightarrow t'_M)$$
$$\text{EntailMin}(t'_F, t'_M) = \min\big(e(t'_F, t'_M),\, e(t'_M, t'_F)\big).$$

Taking the minimum ensures that high scores reflect strong semantic consistency in both directions, i.e., neither text introduces unsupported information. To corroborate the NLI-based results, we additionally report cosine similarity using Sentence-BERT (Abdul-Mageed et al., 2019) and ParaPLUIE (Lemesle et al., 2025).

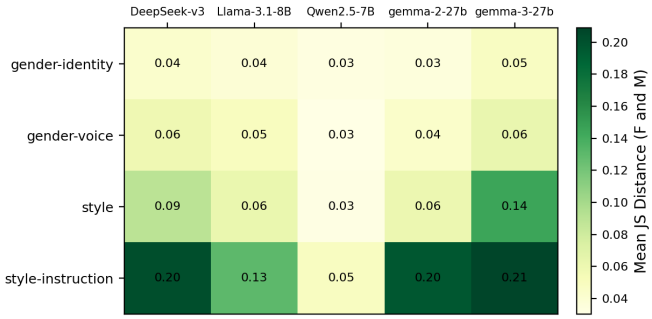

Figure 2: Style Distance between perceived feminine and masculine style. Each cell shows the JS distance between the predicted style distributions of t'_F and t'_M for a given condition. The average style distance for human rewrite pairs is .12.

Grammar Acceptance To estimate grammaticality of a text, we use a RoBERTa model (Liu et al., 2019) fine-tuned on the CoLA Acceptability Corpus (Warstadt et al., 2019) to obtain the softmax probability that a sentence is *acceptable* (i.e., grammatical). For both human and LLM-generated rewrites, we compute this acceptability probability for each rewrite and compare it to the corresponding reference sentence t, reporting the average acceptance scores for t'_F and t'_M as well as their differences relative to t.

4 Perceived Gendered Style Evaluation

We begin by addressing **RQ1**: *Can LLMs produce rewrite pairs that are perceived as contrastively feminine versus masculine?* We analyze perceived gendered style in human-written and LLM-generated rewrites, quantifying contrast both *within* each rewrite pair and *relative* to the reference text. A successful pair should exhibit strong stylistic separation between its two outputs (i.e., a large distance between t'_F and t'_M). At the same time, we examine how far each rewrite shifts away from the reference t, asking whether increased pairwise contrast is achieved through a stylistic shift or through excessive deviation from the original text.

Style Distance Figure 2 shows the style distance for all tested LLMs, where darker green indicates greater stylistic contrast between t'_F and t'_M. The corresponding average for human rewrite pairs is .12. The largest distances can be observed in the *style instruction* setting (bottom row), especially for Gemma-3 (mean JS = .21) and DeepSeek-V3 and Gemma-2 (mean JS = .20), followed by Llama-3. We further compare this setup against the corre-

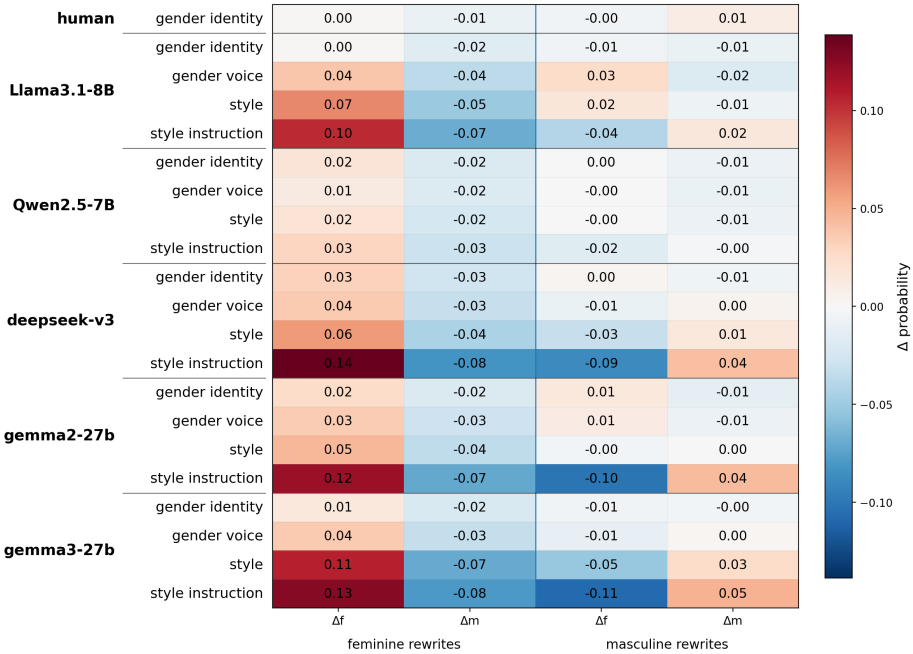

Figure 3: Style Shift relative to reference text. Each cell reports $\Delta_s = p(s \mid t') - p(s \mid t)$ for style class $s \in \{F, M\}$. Red denotes a positive shift (higher probability than the reference), and blue denotes a negative shift. Thus, shifting a rewrite toward feminine style appears as red in Δf and typically blue in Δm.

sponding *style* condition using paired t-tests. The t scores are consistently negative when computed as *style* minus *style instruction* across all models (all $p < .001$, Table 12 in §A.5), indicating significantly greater perceived gendered style contrast under *style instruction*. Moreover, human rewrite pairs are comparable to Llama-3 in this setup, but lower than the strongest outputs from Gemma and DeepSeek. This suggests that explicit linguistically grounded style guidance can induce more contrastive perceived gendered style than identity-based human variation alone.

Style Shift Figure 3 shows the results of the style shift relative to the reference text. For feminine rewrites (two left columns), the largest average shift is observed from DeepSeek-V3 under *style instruction* (Δ_f = .14). Gemma-3 shows the second-largest shift (Δ_f = .13), followed by Gemma-2 (Δ_f = .12) and Llama-3 (Δ_f = .10). In contrast, human female rewrites (top row) do not exhibit a substantial shift relative to the reference style.

For masculine rewrites (two right columns), shifts toward masculine style are more moderate than the corresponding shifts for feminine rewrites. Still, the *style instruction* condition yields the strongest masculine shifts overall. Within this condition, Gemma-3 exhibits the largest shift toward masculine style ($\Delta_m = .05$), while DeepSeek-V3 is only .01 lower. Interestingly, for Llama-3 under the *gender identity*, *voice*, and *style* conditions, masculine rewrites shift on average more toward feminine than masculine style.

To further test whether *style instruction* produces the largest style shifts, we conduct paired t-tests between the *style* and *style instruction* settings. The t scores for both Δ_f and Δ_m are consistently negative across all models (all $p < .001$, Table 12 in §A.5), indicating that fine-grained stylistic instructions produce significantly stronger shifts toward the target style than the *style* prompting condition.

To answer **RQ1**, we find that LLMs can produce rewrite pairs that are perceived as contrastively feminine and masculine, although the strength of this contrast depends strongly on the prompting formulation. In particular, the *style instruction* prompting yields the most contrasted gendered rewrite pairs for most models except those based on Qwen. At the same time, the fine-grained instructions also produce the largest stylistic shifts from the reference texts. Under other prompting conditions, including *gender identity* and *gender voice*, LLM-generated pairs remain less contrastive than human rewrites. Across methods, *gender identity* prompting produces the least contrasted pairs, with lower style contrast than both *gender voice* and *gender style* prompting. Robustness checks with repeated

Approach	Meaning Similarity masc⇔fem	Meaning Preservation masc→ref	Meaning Preservation fem→ref
human	.052 ± .18	.060 ± .14	.053 ± .13
LLM_identity	.802 ± .33	.780 ± .33	.776 ± .34
LLM_voice	.779 ± .34	.752 ± .34	.754 ± .34
LLM_style	.712 ± .37	.742 ± .34	.715 ± .36
LLM_style_instruction	.585 ± .40	.739 ± .35	.585 ± .39

Table 1: Average meaning similarity within feminine–masculine rewrite pairs and meaning preservation of each rewrite relative to the reference text.

Approach	Avg_Fem	Avg_Masc
human	.787	.809
LLM_identity	.914	.906
LLM_voice	.913	.900
LLM_style	.909	.893
LLM_style_instruction	.921	.897

Table 2: Average grammar acceptability of rewrites. The reference texts have the average grammar acceptability of .724.

runs, prompt variants, and length-matched prompts are reported in §A.6.

5 Meaning Preservation and Grammar Acceptability

Having established that LLMs can produce rewrite pairs that are perceived as contrastively feminine vs. masculine, we ask *which prompting operationalization yields the most reliable contrast* (**RQ2**). Beyond stylistic intensity, successful style transfer needs to also preserve content and naturalness (Mir et al., 2019; Fisher et al., 2024; Pauli et al., 2025). To assess whether human and LLM-generated rewrites retain meaning fidelity and remain natural and how these properties vary across prompting approaches, we evaluate contrastive pairs for **meaning preservation** and **naturalness**.

Meaning Similarity Table 1 shows meaning similarity within rewrite pairs (feminine vs. masculine) for human-written texts and four prompting scenarios, averaged across all LLMs (full results in Table 6 in §A.3). We can observe that meaning similarity is highest for the less explicitly stylistic prompting conditions, especially *identity* (avg_mean = .802) and *voice* (avg_mean = .779). In contrast, human rewrite pairs show much lower similarity (mean = .052), reflecting that independently written human stories diverge substantially even when they describe the same event.[3]

Since *style instruction* produced the strongest perceived style contrast in §4, we next test whether this contrast comes at a semantic cost. Paired t-tests show that *style instruction* yields significantly lower meaning similarity than the corresponding *style* condition across all models except Qwen2.5 (Table 10 in §A.5). Together, these results indicate that fine-grained stylistic guidance increases perceived style contrast, but partly at the cost of semantic similarity.

Meaning Preservation The two right columns of Table 1 show the average meaning preservation relative to the reference texts (full results in Table 6 in §A.3). Overall, LLM-generated rewrites preserve substantially more reference meaning than human-written texts. For both feminine and masculine rewrites, the highest preservation scores are achieved by the *identity* and *voice* prompting. In contrast, *style instruction* lowers preservation relative to the corresponding *style* condition on average. Paired t-tests show that this decrease is consistent across all models except for Qwen2.5 (Table 10 in §A.5).

Grammar Acceptance Table 2 shows the average grammatical acceptability of LLM-generated rewrites across prompting scenarios and target styles (for full model-level results, see Table 7 in §A.3). Overall, LLM-generated rewrites are more grammatically acceptable than human rewrites in both target styles. Across LLMs, feminine rewrites receive the highest average acceptability under *style instruction* (avg_fem = .921), whereas masculine rewrites received the highest average acceptability under *identity* prompting (avg_masc = .906). Differences between prompting conditions are small, suggesting that prompt operationalization affects grammatical acceptability less strongly than style contrast or meaning preservation (paired t-test results are reported in Table 11 in §A.5).

To answer **RQ2**, we find that LLM-generated rewrites achieve substantially higher meaning similarity within pairs and stronger meaning preservation relative to the reference than human rewrites. Across prompting scenarios, *identity* and *voice* generally preserve meaning best, while *style instruction* yields the lowest pairwise meaning similarity and lower preservation on average. This confirms a trade-off between stylistic contrast and semantic fidelity: the fine-grained instructions that produce

[3] These patterns are corroborated by Sentence-BERT and ParaPLUIE scores in Tables 8 and 9 in §A.4.

the strongest perceived gendered contrast also tend to move rewrites further from each other and, to a lesser extent, from the reference text.

6 Human Evaluation

The answers to **RQ1** and **RQ2** point to a consistent trade-off: prompting setups that induce stronger perceived gendered contrast also tend to reduce meaning preservation. We validate this finding with a human study, in which annotators judge the perceived gendered style of the rewrites and the contrast between their styles. We use the rewrites from the *style* and *style instruction* conditions generated by DeepSeek-V3, as they showed the most consistent differences across analyses.

6.1 Style Contrast

We sample 100 DeepSeek-V3 rewrite pairs (50 each from the *style* and *style instruction* conditions). Four annotators, balanced by binary gender, rate each pair on 5-point Likert scales for stylistic contrastiveness, meaning similarity, and grammaticality; contrastiveness is judged at the pair level, i.e., how clearly the two rewrites differ in perceived gendered style. All details about the annotation instructions are available in §A.8.1.

Results We find that the *style instruction* condition receives higher perceived gender contrast scores than *style* (3.99 vs. 2.77), indicating that annotators judge the paired rewrites as more clearly differentiated in gendered style (see more results in §A.8.1). At the same time, *style instruction* receives lower scores for meaning similarity (4.26 vs. 4.59) and grammatical acceptability (3.60 vs. 4.12). Thus, the increase of 1.22 points in perceived contrast is accompanied by decreases of 0.33 points in meaning similarity and 0.52 points in grammaticality. However, meaning similarity remains high in both conditions, suggesting that the stronger contrast does not generally eliminate core semantic overlap between the paired rewrites.

6.2 Perceived Gendered Style

We use the same sampled rewrite pairs as in §6.1 and decompose each pair into individual rewrites. Each text is independently annotated by four annotators balanced by binary gender on a 5-point scale ranging from *very feminine* to *very masculine* (we use the same annotation guidelines as Chen et al. (2025) and provide details in §A.8.2).

Results For the *style* condition, feminine rewrites received a mean score of 2.55 ($SD = 0.74$), while masculine rewrites receive a mean score of 3.41 ($SD = 0.62$, see §A.8.2). Since the scale ranges from 1 (very feminine) to 5 (very masculine), these scores indicate that both target directions are perceived as intended, but remain relatively close to neutral. Under *style instruction*, feminine rewrites are perceived as more feminine (mean = 1.97, $SD = 0.58$), and masculine rewrites as more masculine (mean = 3.74, $SD = 0.66$), yielding a wider separation between target styles.

In summary, human evaluation aligns with the automatic metrics: *style instruction* yields stronger perceived gendered contrast than *style*, but lower meaning similarity.

7 Discussion and Conclusion

In this work, we investigate whether LLMs can perform perception-centered gendered style transfer and how different prompt formulations affect the resulting stylistic contrast. Our results reveal a consistent trade-off in perception-centered gendered style transfer: prompts that produce stronger perceived gendered contrast also tend to reduce semantic fidelity. Qualitative examples (see Appendix A.7) suggest that this trade-off may arise from how stylistic contrast is realized. Fine-grained style instructions encourage models to amplify affective and interpersonal cues (e.g., intensifiers, stance markers), which increase perceived gendered contrast but can also lead to deviations from the original content. In contrast, less specific style prompts tend to rely on more conservative edits, resulting in weaker stylistic separation but higher semantic fidelity.

More broadly, our findings highlight that the way gender is conceptualized in prompts plays a crucial role in controlling stylistic variation in generated text. For practical style-transfer applications, this provides guidance on prompt design when balancing stylistic contrast and meaning preservation. For research on the social interpretation of language, our results demonstrate that targeting perceived gendered style and using LLM-generated rewrites can provide a controllable source of contrastive stimuli for studying how linguistic cues shape readers' judgments.

8 Limitations

Our work has several limitations. First, our experiments are based on a randomly sampled subset of 500 source texts from PASTEL, which serve as rewriting inputs rather than stylistic conditioning references. We adopted random sampling to reduce manual selection bias; however, different source text selections may still lead to different LLMs' generation or evaluation outcomes. Future work could further examine the robustness of these findings under alternative sampling strategies.

Second, our evaluation primarily operates at the level of individual rewrites and rewrite pairs rather than modeling the broader distribution of gendered language variation. While we observe consistent aggregate patterns across models, prompting conditions, and metrics, the current framework may not fully capture the diversity and contextual variability of naturally occurring gendered linguistic behavior. More distribution-sensitive analyses remain an important direction for future work.

Third, although we evaluate multiple open-weight LLMs across different prompting formulations, our experiments remain limited to prompting-based approaches. We do not investigate alternative adaptation strategies such as fine-tuning, retrieval-augmented generation, or controllable decoding methods, which may produce different stylistic and semantic trade-offs.

Fourth, our operationalization of perceived gendered style may be potentially simplified. Following prior work, we model perceived gender style primarily through probabilistic binary feminine and masculine linguistic tendencies. However, gender expression is socially and culturally complex, fluid, and non-binary, and many stylistic patterns associated with gender are context-dependent rather than universal. Consequently, our results should not be interpreted as capturing inherent or essential properties of gendered language.

Fifth, another limitation concerns the joint generation setup used in our prompting framework. Specifically, feminine and masculine rewrites are generated together within the same prompt, which may encourage models to optimize for contrastive differences between outputs. Future work could explore alternative generation strategies, including independently generating each rewrite, to better understand how joint versus separate generation affects stylistic contrast and semantic preservation.

Sixth, although we complement automatic evaluation with human annotations, perceived gender style remains inherently subjective. While our classifier is trained on human judgments and achieves relatively strong agreement with annotator distributions, automatic evaluation may still fail to capture nuanced or context-specific stylistic perceptions. Similarly, inter-annotator agreement for the human evaluation tasks is moderate, reflecting the subjective nature of perceived gender judgments. In addition, our human evaluation experiments are limited to rewrite samples generated by DeepSeek-V3. Although DeepSeek-V3 consistently showed the strongest stylistic contrast in our automatic evaluation, we do not know whether the same human evaluation findings would generalize to other model families or prompting behaviors.

Seventh, while our perceived gender classifier includes a neutral class, our experiments focus primarily on generating and evaluating contrastive feminine and masculine rewrites. Future work could explicitly investigate neutral rewrites or stylistic variation across broader gender spectra, including about more nuanced or non-binary forms of gendered stylistic expression.

Finally, although we perform robustness analyses using repeated runs, prompt variation, and length-matched prompting conditions, these analyses are conducted on a subset of models and prompting settings due to computational constraints. Future work should evaluate whether these findings generalize across broader model families, languages, and domains.

9 Ethics & Potential Risks

Although our work does not conceptualize gendered style as a fixed property directly determined by author identity, our operationalization of gendered style in prompting through dimensions such as identity, voice, style, and fine-grained stylistic instruction remains limited. Gendered stylistic expression is socially complex and may still become associated with stereotypes or assumptions about gender reflected in LLM outputs. Moreover, while our automatic metrics treat perceived gendered style as a probabilistic distribution, the rewriting task itself focuses on contrastive feminine versus masculine stylistic variation and does not explicitly incorporate neutral or non-binary target styles, which may still encourage binary interpretations of gendered language variation. Perceptions of gendered style are also shaped by many intersecting

social and cultural factors, such as age, race, and social class.

The intent of this work is not to define authentic feminine or masculine writing, but to investigate how prompting strategies influence the stylistic cues that humans and models perceive as gendered in LLM-generated rewrites. In this sense, the generated rewrite pairs serve as research artifacts for studying prompt sensitivities, stylistic variation, and potential style biases in LLM outputs. For example, our framework may help analyze whether certain prompting conditions systematically encourage stylistic patterns perceived as more feminine or masculine, and how these patterns interact with downstream model behavior.

At the same time, systems capable of generating highly contrastive gendered rewrites may reinforce stereotypical associations or be misused for deceptive identity presentation or stylistic profiling. Importantly, stronger stylistic contrast does not necessarily correspond to authentic or representative gendered expression; rather, it may reflect the amplification of socially recognizable stylistic cues optimized by LLMs under specific prompting conditions. Although we emphasize that our evaluation focuses on perceived gendered stylistic variation in LLM-generated rewritten texts within a specific annotation and experimental context, such outputs may still be misunderstood or overgeneralized as reflecting authentic real-world gendered expression, potentially reinforcing stereotypes about gendered style and language use.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Arun Rajendran, AbdelRahim Elmadany, Michael Przystupa, and Lyle Ungar. 2019. Sentence-Level BERT and Multi-Task Learning of Age and Gender in Social Media. *arXiv preprint arXiv:1911.00637*.

Paul Baker. 2014. Using corpora to analyze gender.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Joel Baum and Kim Westheimer. 2015. Sex? sexual orientation? gender identity? gender expression? *Teaching Tolerance*, 50(34–38).

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, Montréal, Canada. Association for Computational Linguistics.

Mary Bucholtz. 2002. From 'sex differences' to gender variation in sociolinguistics. *University of Pennsylvania Working Papers in Linguistics*, 8(3):33–45.

J. Butler. 1990a. Gender trouble: Feminism and the subversion of identity.

Judith Butler. 1990b. Feminism and the subversion of identity. *Gender trouble*, 3(1):3–17.

Judith Butler. 2002. *Gender trouble*. routledge.

Judith Butler. 2004. *Undoing gender*. routledge.

Hongyu Chen, Neele Falk, Michael Roth, and Agnieszka Falenska. 2025. "feels feminine to me": Understanding perceived gendered style through human annotations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31459–31480, Suzhou, China. Association for Computational Linguistics.

Hongyu Chen, Michael Roth, and Agnieszka Falenska. 2024. What can go wrong in authorship profiling: Cross-domain analysis of gender and age prediction. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 150–166, Bangkok, Thailand. Association for Computational Linguistics.

Mats Dahllöf. 2023. Author gender and text characteristics in contemporary swedish fiction. *Language and Literature*, 33:69 – 100.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and Gender*, 2 edition. Cambridge University Press.

Diane Ehrensaft. 2018. Exploring gender expansive expressions versus asserting a gender identity. In Colt Keo-Meier and Diane Ehrensaft, editors, *The gender affirmative model: An interdisciplinary approach to supporting transgender and gender expansive children*, pages 37–53. American Psychological Association.

Jillian Fisher, Ximing Lu, Jaehun Jung, Liwei Jiang, Zaid Harchaoui, and Yejin Choi. 2024. Jamdec: Unsupervised authorship obfuscation using constrained decoding over small language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1552–1581.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jingxuan Han, Quan Wang, Zikang Guo, Benfeng Xu, Licheng Zhang, and Zhendong Mao. 2024. Disentangled learning with synthetic parallel data for text style transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15187–15201, Bangkok, Thailand. Association for Computational Linguistics.

Lingshu Hu and Michael Wayne Kearney. 2021. Gendered tweets: Computational text analysis of gender differences in political discussion on twitter. *Journal of Language and Social Psychology*, 40(4):482–503.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, and 4 others. 2025. Gemma 3 technical report. *ArXiv*, abs/2503.19786.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, Ph.D) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1696–1706, Hong Kong, China. Association for Computational Linguistics.

Moreno La Quatra, Giuseppe Gallipoli, and Luca Cagliero. 2024. Self-supervised text style transfer using cycle-consistent adversarial networks. *ACM Trans. Intell. Syst. Technol.*, 15(5).

Geoffrey Leech and 1 others. 1992. 100 million words of english: the british national corpus (bnc). *Language research*, 28(1):1–13.

Quentin Lemesle, Jonathan Chevelu, Philippe Martin, Damien Lolive, Arnaud Delhay, and Nelly Barbot. 2025. Paraphrase generation evaluation powered by an LLM: A semantic metric, not a lexical one. In *Proceedings of the 31st International Conference on Computational Linguistics*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qingyi Liu, Jinghui Qin, Wenxuan Ye, Hao Mou, Yuxuan He, and Keze Wang. 2024b. Adaptive prompt routing for arbitrary text style transfer with pre-trained language models. In *AAAI Conference on Artificial Intelligence*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2023. Bounding the capabilities of large language models in open text generation with prompt constraints. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1982–2008.

Guoqing Luo, Yu Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5740–5750, Singapore. Association for Computational Linguistics.

Huiyu Mai, Wenhao Jiang, and Zhi-Hong Deng. 2023. Prefix-tuning based unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14847–14856, Singapore. Association for Computational Linguistics.

David D. McDonald and James D. Pustejovsky. 1985. A computational theory of prose style for natural

language generation. In *Second Conference of the European Chapter of the Association for Computational Linguistics*, Geneva, Switzerland. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Damián Morales Sánchez, Antonio Moreno, and María Dolores Jiménez López. 2022. A white-box sociolinguistic model for gender detection. *Applied Sciences*, 12(5):2676.

Daniil Moskovskiy, Sergey Pletenev, and Alexander Panchenko. 2024. LLMs to replace crowdsourcing for parallel data creation? the case of text detoxification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14361–14373, Miami, Florida, USA. Association for Computational Linguistics.

Dong Nguyen, Dolf Trieschnigg, A Seza Dogruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1950–1961. Association for Computational Linguistics.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2025. Mind the style gap: Meta-evaluation of style and attribute transfer metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21550–21564, Suzhou, China. Association for Computational Linguistics.

Christine Pinney, Amifa Raj, Alex Hanna, and Michael D Ekstrand. 2023. Much ado about gender: Current practices and future recommendations for appropriate gender-aware information access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 269–279.

Daniel Preoţiuc-Pietro, Wei Xu, and Lyle Ungar. 2016. Discovering user attribute stylistic differences via paraphrasing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3030–3037. AAAI Press.

Daniel Preoţiuc-Pietro, Sharath Chandra Guntuku, and Lyle Ungar. 2017. Controlling human perception of basic user traits. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2335–2341, Copenhagen, Denmark. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 3 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Rika Saito. 2010. Writing in female drag: Gendered literature and a woman's voice. *Japanese Language and Literature*, 44(2):149–177.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Martina Toshevska and Sonja Gievska. 2025. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*, 13:44707–44721.

Enrica Troiano, Aswathy Velutharambath, and Roman Klinger. 2023. From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *Natural Language Engineering*, 29(4):849–908.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024. Distilling text style transfer with self-explanation from LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–211, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

A.1 Model Generation Setup

Table 3 shows the system promts we use to instruct all the LLMs. All LLM generations were run through the Hugging Face Router (OpenAI-compatible chat completions endpoint) using the same API configuration across models. For each input reference text, we request two rewrites in a strict two-line format and apply a lightweight reformatting pass when outputs violate the schema. We use identical decoding hyperparameters for all models and prompting conditions: temperature 0.7, top-p 0.9, and a maximum of 512 new tokens, with a fixed random seed (42). To improve robustness to API failures (e.g., rate limits and gateway timeouts), we use exponential-backoff retries and save partial outputs upon unrecoverable errors.

A.2 Results of Perceived Gendered Style Classifiers

Table 4 presents the performance of different classifiers for three-way perceived gender style prediction (*feminine*, *masculine*, and *neutral*) under 5-fold cross-validation. Overall, transformer-based models outperform feature-based baselines and the dummy classifier across all evaluation metrics. The confidence-weighted BERT-base-uncased model achieves the best overall performance, obtaining the highest accuracy (.62) and Macro F1 (.50), as well as the strongest performance on high-confidence items ($F1^{\dagger} = .61$).

Compared to classifiers such as Random Forest and MLP models trained on manually extracted linguistic features, BERT-based models consistently achieve higher classification performance, suggesting that it better captures perceived gendered stylistic variation. In addition, the confidence-weighted training setup slightly improves performance over the standard BERT-base-uncased model, indicating that incorporating annotator confidence information helps the classifier better model perceived gender style distributions.

Given that perceived gender style annotations are inherently subjective, we additionally evaluate classifiers using probability distributions rather than relying solely on majority-vote labels. Table 5 compares classifiers using Jensen–Shannon (JS) distance between predicted and gold annotation distributions, where lower values indicate better alignment with human judgments. The confidence-weighted BERT model again achieves the best overall performance, obtaining the lowest mean JS distance (.204) and standard deviation (.104), indicating that its predicted style distributions match more with the human annotation distributions. Based on these results, we choose the confidence-weighted BERT classifier for all automatic evaluations in the main experiments.

A.3 NLI Average Semantic and CoLA Grammar Scores

A.4 Semantic Metric

We compute embedding-based cosine similarity using Sentence-BERT[4]. We embed the reference t and the rewrites (t'_F, t'_M) (for both human and LLM outputs) and report cosine similarity for the same comparisons: $\cos(t'_F, t'_M)$, $\cos(t, t'_F)$, and $\cos(t, t'_M)$. We report results in Table 8

We additionally compute semantic similarity using ParaPLUIE[5], a perplexity-based paraphrase evaluation metric. ParaPLUIE estimates semantic proximity between two texts using language model confidence scores and has been shown to correlate strongly with human judgments on paraphrase similarity. Following the same setup as the cosine similarity evaluation, we compute ParaPLUIE scores for the same text pairs: (t'_F, t'_M), (t, t'_F), and (t, t'_M). We report the results in Table 9.

A.5 Statistical Testing

Paired T-test Based on our observations from the style distance and style shift results in §4, fine-grained *style instruction* prompting consistently outperformed the other prompting conditions (*identity*, *voice*, and *style*) in producing stronger stylistic contrast. However, on semantic preservation metrics, *style instruction* generally performs worse than the *identity*, *voice*, and *style* conditions. To examine whether these differences were statistically

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[5] https://huggingface.co/spaces/qlemesle/parapluie

consistent, we conduct paired Student's t-tests between the *style* and *style instruction* conditions, as the latter represents a direct fine-grained variant of the general *style* prompting approach. This comparison allows us to investigate whether fine-grained stylistic guidance consistently produces stronger stylistic contrast at the cost of lower semantic preservation relative to the general *style* condition. Comparisons are performed on paired rewrites generated from the same reference texts ($n = 500$ per model), controlling for source-text variation. In this section, we report t-test results for perceived gendered style, style-related metrics, meaning preservation, and grammatical acceptability.

Style Distance & Shift Table 12 presents the paired t-test results for style distance and style shift between the *style* and *style instruction* conditions. For style distance, *style instruction* consistently produces significantly larger distances than *style* across all models (all $p < .001$). The largest effect is observed for Gemma-2 ($t = -23.99$, $p < .001$), followed by DeepSeek-V3 ($t = -22.25$, $p < .001$), consistent with our earlier descriptive observations.

For the feminine style shift, *style instruction* also yields significantly larger shifts away from the reference text toward the feminine target style. DeepSeek-V3 shows the largest effect among all models ($t = -21.51$, $p < .001$). Similarly, for masculine style shift, masculine rewrites generated by Gemma-2 exhibit the strongest shift toward the masculine target style ($t = -18.33$, $p < .001$). Overall, these results suggest that fine-grained linguistic style instructions significantly increase stylistic contrast and more strongly shift rewrites toward the intended target style compared to the general *style* prompting condition.

Meaning Similarity & Meaning Preservation Table 10 presents the paired t-test results for meaning similarity and meaning preservation comparing the *style* and *style instruction* conditions across models. Positive t values indicate lower meaning similarity or preservation under the *style instruction* condition relative to the corresponding *style* condition.

For meaning similarity between feminine and masculine rewrites, *style instruction* generally yields significantly lower similarity scores than *style* across models, except for Qwen2.5 ($t = 0.85$, $p = .394$). The largest effects are observed for DeepSeek-V3 ($t = 20.59$, $p < .001$), followed by Gemma-2 ($t = 20.34$, $p < .001$), suggesting that fine-grained stylistic instructions produce rewrite pairs that are more semantically differentiated than those generated under the general *style* condition, consistent with our previous observations.

For meaning preservation relative to the reference text, feminine rewrites generated under *style instruction* also show significantly lower preservation scores than those generated under *style* across models, again except for Qwen2.5 ($t = 1.34$, $p = .177$). The strongest effects are observed for Gemma-2 ($t = 16.87$, $p < .001$) and DeepSeek-V3 ($t = 16.10$, $p < .001$), indicating that feminine rewrites generated with fine-grained stylistic instructions preserve less semantic content from the reference text than those generated under the general *style* condition.

Similarly, masculine rewrites generated under *style instruction* generally exhibit significantly lower meaning preservation than those generated under *style*, except for Qwen2.5 ($t = -0.15$, $p = .879$). However, the effects for masculine rewrites are generally smaller than those observed for feminine rewrites. Gemma-2 shows the largest effect ($t = 8.55$, $p < .001$), followed closely by DeepSeek-V3 ($t = 8.54$, $p < .001$). Overall, these findings suggest that while fine-grained stylistic instructions increase stylistic differentiation, they also reduce semantic preservation relative to the general *style* prompting condition, aligning with our earlier observations.

Grammar Acceptance Table 11 reports the paired t-test results for grammatical acceptability across prompting conditions. Negative t values indicate lower grammaticality scores under the *style instruction* condition relative to the corresponding *style* condition. For feminine rewrites, *style instruction* yields significantly lower grammatical acceptability than *style* for both Gemma-3 ($t = -5.65$, $p < .001$) and Llama-3 ($t = -3.96$, $p < .001$), indicating that fine-grained stylistic instructions reduce grammatical acceptability in feminine rewrites compared to the general *style* condition. In contrast, we do not observe statistically significant differences for masculine rewrites across models.

A.6 Robustness and Sensitivity Analyses

A.6.1 Robustness Across Random Seeds

To assess whether the observed differences between prompting conditions are robust to decoding variability, we repeat generation experiments using five random seeds (42, 128, 256, 512, and 1024). We use DeepSeek-V3 as a case study, which consistently showed the strongest performance on style distance and style shift metrics in the main experiments. For both the *style* and *style instruction* conditions, we regenerate rewrites across all seeds and recomputed style-related and semantic preservation metrics.

Table 13 presents aggregate results across five random seeds for DeepSeek-V3 under the *style* and *style instruction* conditions. Overall, the results remain stable across runs, with relatively small standard deviations across all metrics, indicating stable trends across random seeds.

For example, the average JS style distance under the *style* condition is .111 with a standard deviation of .012 across five runs, whereas the *style instruction* condition achieves a larger average style distance of .201 with a much smaller standard deviation of .001. Similarly, the average meaning similarity between feminine and masculine rewrites under the *style* condition is .761 (SD = .028), while the *style instruction* condition yields a lower average similarity of .553 (SD = .011). These results further support our observations that fine-grained stylistic instructions produce stronger stylistic differentiation between rewrites while simultaneously reducing semantic similarity and meaning preservation relative to the general *style* condition.

A.6.2 Prompt Variation

To evaluate prompt sensitivity, we additionally test two more prompt variants for the *style* and *style instruction* conditions. Specifically, we vary the first sentence (task description) of the system prompt using shorter and longer variants while keeping the core stylistic instructions unchanged (see the full prompt template in Table 16). We then evaluate the resulting rewrites using the same automatic metrics.

Using DeepSeek-V3 as a case study, Table 14 presents aggregate results across prompt variants under the *style* and *style instruction* conditions. Overall, the results remain consistent across prompt formulations, with relatively small standard deviations across all metrics, suggesting that the observed effects are relatively stable under moderate prompt wording variations.

For example, the average JS style distance under the *style* condition is .104 with a standard deviation of .013 across prompt variants, whereas the *style instruction* condition achieves a larger average style distance of .193 with a smaller standard deviation of .008. Similarly, the average meaning similarity between feminine and masculine rewrites under the *style* condition is .786 (SD = .022), while the *style instruction* condition yields a considerably lower average similarity of .576 (SD = .014). These results further confirm that fine-grained stylistic instructions consistently produce stronger stylistic differentiation while reducing semantic similarity relative to the general *style* condition, regardless of moderate prompt wording variations.

A.6.3 Length-Matched Prompting

To examine whether the stronger stylistic effects observed under the *style instruction* condition could be influenced by prompt length differences, we constructed length-matched variants of the *identity* and *voice* prompts by extending their target descriptions to approximately match the length of the *style instruction* prompts while preserving their original semantic content. Table 16 presents the prompt templates used for the gender identity and gender voice instruction settings.

Specifically, in the *identity* condition, we instructed the model to "rewrite the text as if written by a woman author" while additionally encouraging attention to wording, tone, phrasing, and expression; the same formulation was applied for the man author target style. Similar extensions were applied to the *voice* condition. Because the *identity* and *voice* settings differ conceptually from the fine-grained *style instruction* condition, which is grounded in sociolinguistic feature descriptions, we intentionally avoided introducing additional explicit stylistic features into these prompts. This setup therefore focuses on controlling for prompt length rather than enriching the *identity* or *voice* conditions with new stylistic cues. We then regenerated rewrites using these modified prompts and reevaluated the same style-related and semantic metrics.

To evaluate how fine-grained instruction prompting differs from the corresponding non-fine-grained prompting conditions in terms of perceived gendered style, meaning similarity, and meaning preservation, we conducted paired t-tests between

each base condition and its instruction variant. We refer to the non-fine-grained prompting settings (*identity*, *voice*, and *style*) as the base conditions, and their corresponding fine-grained variants as the instruction conditions (*identity instruction*, *voice instruction*, and *style instruction*).

Style Distance & Shift Table 12 reports the paired t-test results for perceived style-related metrics comparing base and instruction prompting conditions. Negative t values indicate higher scores under the instruction condition relative to the corresponding base condition.

For style distance, rewrites generated under the instruction conditions generally exhibit significantly larger distances between feminine and masculine rewrites than their corresponding base conditions across models (all comparisons $p < .05$). Among all comparisons, the *style* versus *style instruction* setting consistently produces the strongest effects, with Gemma-2 showing the largest effect, followed by DeepSeek-V3. These results indicate that fine-grained linguistic style instructions produce substantially stronger stylistic contrast than the corresponding base prompting conditions, consistent with our previous findings.

For feminine style shift relative to the reference text, the comparisons between base and instruction conditions are also generally statistically significant across models and prompting conditions (most comparisons $p < .001$). Again, the *style* versus *style instruction* comparison yields the largest effects, with DeepSeek-V3 followed by Gemma-2, indicating that fine-grained stylistic instructions produce larger shifts away from the reference text toward the feminine target style. Interestingly, for Gemma-3, Llama-3, and Qwen2.5, the *identity* versus *identity instruction* comparisons produce larger effects than the other prompting conditions (e.g., Llama-3-identity, $t = -11.14$, $p < .001$), suggesting that fine-grained identity-based instructions may induce stronger stylistic shifts while preserving less semantic content in feminine rewrites relative to the reference text than the corresponding *style instruction* condition.

For masculine style shift relative to the reference text, the *style* versus *style instruction* comparison again shows the strongest and most consistent effects across models. Masculine rewrites generated by Gemma-2 under *style instruction* exhibit the largest shift away from the reference text toward the masculine target style ($t = -18.33$, $p < .001$), followed by DeepSeek-V3 ($t = -13.10$, $p < .001$).

Overall, instruction prompting generally produces larger stylistic distances and stronger shifts away from the reference text than the corresponding base conditions. Among these comparisons, fine-grained *style instruction* prompting consistently yields the strongest stylistic contrast for both feminine and masculine rewrites, aligning with our earlier observations.

Meaning Similarity & Preservation Table 18 reports the paired t-test results for meaning similarity (MS) and meaning preservation (MP) comparing base and instruction prompting conditions across models and gendered style operationalizations (*identity*, *voice*, and *style*). Positive t values indicate lower meaning similarity or preservation under the instruction condition relative to the corresponding base condition.

For meaning similarity between feminine and masculine rewrites, *style instruction* generally yields significantly lower semantic similarity than the corresponding *style* condition across models (most comparisons $p < .001$), except for Qwen2.5. The largest effects are observed for DeepSeek-V3 ($t = 20.59$, $p < .001$), followed by Gemma-2 ($t = 20.34$, $p < .001$), consistent with our previous findings that fine-grained stylistic instructions produce more semantically differentiated rewrite pairs. However, for Llama-3, the *identity* versus *identity instruction* comparison produces larger effects than the other prompting conditions, indicating that identity-based fine-grained instructions may reduce semantic similarity more strongly for this model.

For meaning preservation relative to the reference text, feminine rewrites generated under *style instruction* again show the strongest effects for Gemma-2 and DeepSeek-V3, indicating that fine-grained stylistic instructions shift rewrites further away from the reference text toward the target style than the corresponding base *style* condition. Similar trends are observed for masculine rewrites. In contrast, for Llama-3, the *identity* versus *identity instruction* comparison produces larger effects than the *voice* and *style* comparisons, suggesting that identity-based instructions may induce stronger stylistic shifts while reducing semantic preservation more substantially for Llama-3.

Overall, fine-grained *style instruction* prompting consistently produces stronger stylistic differentiation and lower meaning similarity and preservation than the corresponding base *style* condition

for DeepSeek-V3 and Gemma-2, aligning with our previous findings. However, for Llama-3, identity-based instruction prompting produces larger semantic shifts and lower meaning similarity than the corresponding *style* comparison.

A.7 Example from LLMs rewrites

This section analyzes the differences in rewrites across two prompting approaches *style* and *style instruction* from Table 20. With an example from DeepSeek-V3, Table 20 illustrates how fine-grained style instructions can increase perceived gender contrast while introducing a trade-off with semantic fidelity. Under *style instruction*, the feminine rewrite shifts the perceived gendered-style distribution toward the feminine class (PGS: $[0.399, 0.310, 0.291]$) relative to the reference (PGS: $[0.328, 0.400, 0.272]$), while the masculine rewrite shifts toward the masculine class (PGS: $[0.221, 0.428, 0.351]$), yielding a more contrasted pair. The highlighted edits show how this contrast is achieved: some changes preserve the same content but paraphrase surface form (e.g., **left work** $\rightarrow$ **finished work**, **thank goodness** $\rightarrow$ **thankfully**; teal), whereas other additions introduce affective or interpersonal cues not present in the reference (e.g., **incredibly**, **to my relief**, **caring people**; blue). These additions likely contribute to the stronger perceived femininity based on fine-grained instruction on feminine target style, but they also coincide with lower meaning fidelity (MP: 0.438; MS: 0.772) than the *style* condition, where the feminine rewrite remains closer to the reference in both content and perceived style (PGS: $[0.314, 0.362, 0.324]$; MP: 0.748; MS: 0.973). Notably, in the *style* setting, the intended feminine/masculine contrast partially collapses: the feminine rewrite stays near the reference's masculine-leaning distribution, and the masculine rewrite even trends more feminine than expected (PGS: $[0.364, 0.347, 0.289]$). Overall, the example clarifies why instruction-guided prompting yields stronger perceived contrast on average, but can reduce meaning preservation: increased contrast is partly realized by adding or amplifying sentiment/stance cues rather than only re-encoding the same content with different linguistic realizations.

A.8 Human Evaluation

A.8.1 Style Contrast

Annotators are asked to rate each rewrite pair along three dimensions: (i) perceived gendered contrast, (ii) meaning preservation, and (iii) grammatical acceptability. Each annotator evaluated 10 randomly selected rewrite pairs. All ratings are collected on a 5-point Likert scale. For stylistic contrast, higher scores indicate a stronger perceived stylistic difference between the two rewrites; for meaning preservation, higher scores indicate stronger semantic alignment; and for grammaticality, higher scores indicate better fluency and acceptability. Figure 7 presents the final annotation instructions, and Figure 6 presents the annotator consent form used in the style contrast task.

We recruit 40 annotators from Prolific[6], balanced across binary gender identities. Annotators complete the task through a custom-built Streamlit interface and receive compensation equivalent to approximately £22 per hour on average. To reduce potential bias, annotators are not informed of the generation condition, and rewrite pairs from different conditions are randomly mixed during annotation. This ensures that judgments are made independently of condition awareness and minimize potential expectation effects. We receive 400 annotations in total. The overall inter-annotator agreement style contrast is moderate ($\alpha = 0.31$).

Annotation Results Figure 4 presents the frequency distributions of human ratings across stylistic contrastiveness, meaning similarity (content), and grammaticality for the *style* and *style instruction* conditions generated by DeepSeek-V3.

For stylistic contrast, rewrite pairs generated under the *style* condition most frequently received a score of 2 ($n = 79$), indicating relatively weak perceived stylistic differences between feminine and masculine rewrites. In contrast, rewrite pairs generated under the *style instruction* condition most frequently received the highest score of 5 ($n = 81$), suggesting stronger perceived stylistic contrast.

For meaning similarity, both conditions show generally high ratings overall, although the *style* condition receives more ratings of 5 ($n = 136$) than the *style instruction* condition ($n = 92$), indicating slightly stronger perceived semantic preservation. Overall, the human annotations support our automatic evaluation results: rewrites generated under *style instruction* exhibit stronger perceived stylistic contrast than those generated under the general *style* condition, while showing moderately lower semantic similarity.

[6] https://www.prolific.com/

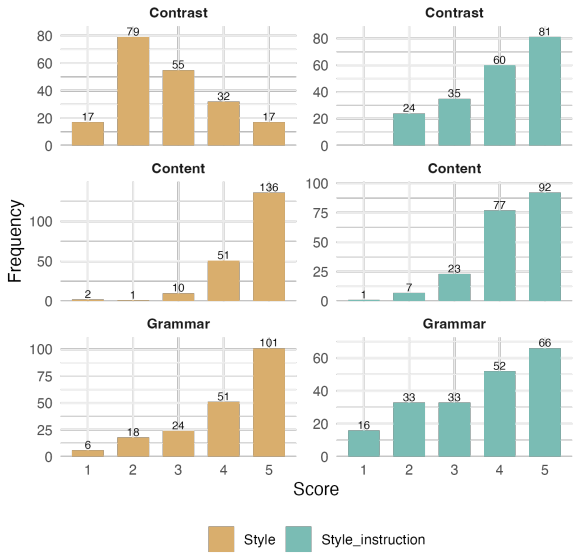

Figure 4: Frequency Distribution of human annotation scores for stylistic contrastiveness, meaning similarity (*Content*), and grammaticality across the *style* and *style instruction* conditions from DeepSeek-V3. Higher contrast scores indicate stronger perceived gendered stylistic differences between feminine and masculine rewrites, while higher content and grammar scores indicate better semantic preservation and grammatical acceptability.

A.8.2 Perceived Gendered Style

We conduct a similar annotation procedure for perceived gendered style using the same custom-built Streamlit platform and Prolific recruitment setup as in the style contrast task. Specifically, we reuse the same DeepSeek-V3 rewrite sample selected for the contrastiveness evaluation and decompose each rewrite pair into individual texts for annotation. We recruit 20 annotators from Prolific, balanced across binary gender identities, with an average compensation of approximately £13 per hour. Annotators are shown individual rewrites and asked to rate the perceived gendered linguistic style on a five-point Likert scale, where 1 indicates *very feminine*, 2 *somewhat feminine*, 3 *neutral*, 4 *somewhat masculine*, and 5 *very masculine*. Each text is independently annotated by four annotators using a different annotator pool from the contrastiveness task. Each annotator completes 40 texts, resulting in approximately 800 annotations in total. We then report the mean perceived gendered style scores for feminine and masculine rewrites under each condition. The overall inter-annotator agreement for perceived gendered style is ($\alpha = 0.449$). See Figure 12 for the final annotation instructions, and Figure 11 for the consent form for annotators in the perceived gendered style task.

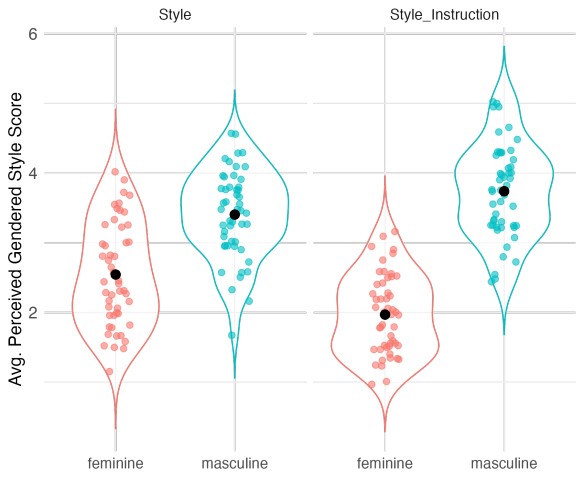

Figure 5: Human annotations of average perceived gendered style scores for feminine and masculine rewrites under the *style* and *style instruction* conditions generated by DeepSeek-V3. Scores range from 1 (*very feminine*) to 5 (*very masculine*). Black dots indicate mean scores. Each dot represents the average perceived gendered style score assigned by four annotators of one individual text.

Annotation Results Figure 5 presents the distributions of average perceived gendered style scores for individual feminine and masculine rewrites under the *style* and *style instruction* conditions generated by DeepSeek-V3. Scores range from 1 (*very feminine*) to 5 (*very masculine*). Overall, rewrites generated under the *style instruction* condition exhibit stronger separation between feminine and masculine score distributions than those generated under the general *style* condition.

Specifically, feminine rewrites under *style instruction* receive lower average perceived gender scores (closer to *very feminine*), while masculine rewrites receive higher scores (closer to *very masculine*) relative to the general *style* condition. The mean perceived gender scores (black dots) also show a larger gap between feminine and masculine rewrites under *style instruction*, indicating stronger perceived gendered stylistic contrast. Overall, the human evaluation results align with our automatic classifier-based findings, suggesting that fine-grained stylistic instructions produce more perceptibly gendered rewrites than the general *style* prompting condition.

A.9 Artifact Use and Intended Scope

We use the publicly available PASTEL dataset (Kang et al., 2019) and open-source language mod-

els in accordance with their intended research use and licensing conditions. PASTEL was originally designed to study persona-conditioned stylistic variation, and in our work the reference texts and human rewrites are used solely for research purposes related to prompting-based gendered stylistic rewriting and evaluation. The generated rewrite pairs are intended as research artifacts for analyzing perceived gendered stylistic variation and prompting effects, rather than for identity inference or real-world profiling applications.

A.10 AI-Assistance Use

In this work, we used ChatGPT (GPT-5.2 Instant) to assist with code completion and implementation suggestions. We also used Grammarly for spelling and grammar checks in the main texts.

Prompt condition	System prompt
Shared template (all conditions)	**Task.** You are generating a parallel dataset. For each input, create two rewrites that preserve the original meaning exactly, but differ in the specified target dimension. **Hard constraints (must follow):** • Preserve meaning and facts exactly: entities, numbers, dates, locations, claims, intent, sentiment strength, tense, and modality. • Do not add new information, examples, justifications, or opinions. • Do not introduce stereotypes or demographic assumptions. Style is purely linguistic, not "gender roles." • Keep both rewrites natural and fluent. • Keep length within $\pm 15\%$ unless impossible. • If the input is too short to restyle naturally, still try, but do not pad with new content. **Output format (strict):** Return EXACTLY two lines, no extra text: • `WOMAN/FEMALE/FEMININE: <rewrite>` • `MAN/MALE/MASCULINE: <rewrite>`
Gender identity	**Difference dimension:** authorial identity. **Identity targets:** author who identifies as a woman **and** author who identifies as a man.
Gendered voice	**Difference dimension:** authorial voice. **Voice targets:** FEMALE-author voice tendencies **and** MALE-author voice tendencies.
Gendered style (no fine-grained cues)	**Difference dimension:** linguistic style. **Style targets:** Feminine Style Tendencies (linguistic) **and** Masculine Style Tendencies (linguistic).
Gendered style + fine-grained instructions	**Difference dimension:** linguistic style. **Style targets:** Feminine Style Tendencies (linguistic) **and** Masculine Style Tendencies (linguistic). **Feminine Style Tendencies (linguistic):** • More emotional/relational framing and empathy markers (without changing sentiment strength). • More collaborative tone: inclusive language (we/our/let's when consistent), hedges (maybe/perhaps/I think). • More descriptive language: adjectives/adverbs and sensory/aesthetic framing ONLY if already implied (do not invent facts). • Slightly more complex/flowing sentences (subordinate clauses, narrative flow). **Masculine Style Tendencies (linguistic):** • More fact-focused framing: logic/structure/problem-solution framing (without adding data). • More direct/assertive phrasing (but keep the same level of certainty as the original). • More concise language: shorter, to-the-point sentences with minimal elaboration. • More action-oriented verbs and goal-driven framing (without changing intent).

Table 3: System prompts used to generate gender-contrasted rewrite pairs under four prompting conditions. All conditions share the same hard constraints and strict two-line output format; they differ only in the target definition block. When model outputs deviate from the required two-line format, we apply a lightweight reformatting prompt that preserves content and enforces the output schema.

Model	**Acc.**	**F1**	**F1**†
Dummy classifier	0.45 ± 0.005	0.21 ± 0.001	0.25 ± 0.06
Random forest	0.56 ± 0.046	0.42 ± 0.044	0.58 ± 0.09
MLP (key_extracted_features)	0.55 ± 0.032	0.46 ± 0.027	0.59 ± 0.13
BERT-base-uncased	0.61 ± 0.054	0.49 ± 0.041	0.60 ± 0.20
BERT-base-uncased (conf.-weighted)	0.62 ± 0.059	0.50 ± 0.062	0.61 ± 0.19

Table 4: Classifier comparison for three-way perceived gendered style prediction (feminine/masculine/neutral) under 5-fold cross-validation (mean $\pm$ std). Accuracy is computed against majority-vote labels, while Macro F1 evaluates balanced performance across the three classes. †Macro F1 is additionally computed on high-confidence items with annotator agreement probability $\geq 70\%$ ($N = 166$).

Model	Mean JS Distance	Std. JS Distance
Random Forest	0.279	0.135
MLP (key_extracted_features)	0.253	0.133
BERT-base-uncased	0.205	0.105
BERT-base-uncased (conf.-weighted)	0.204	0.104

Table 5: Comparison of models by mean Jensen–Shannon (JS) distance and standard deviation of JS distance. Lower is better.

Approach	**Meaning Similarity**		**Meaning Preservation**			
	masc⇔fem		**masc→ref**		**fem→ref**	
	mean	**std**	**mean**	**std**	**mean**	**std**
human	0.052	0.182	0.060	0.136	0.053	0.127
deepseek_identity	0.810	0.339	0.833	0.308	0.806	0.327
deepseek_voice	**0.831**	0.304	0.835	0.295	0.793	0.321
deepseek_style	0.811	0.319	**0.847**	0.283	0.803	0.319
deepseek_style_instruction	0.570	0.418	0.767	0.339	0.611	0.400
gemma3_identity	0.799	0.323	0.801	0.319	0.763	0.334
gemma3_voice	0.777	0.335	0.772	0.329	0.738	0.351
gemma3_style	0.654	0.396	0.794	0.317	0.662	0.376
gemma3_style_instruction	0.558	0.422	0.759	0.339	0.581	0.399
gemma2_identity	0.826	0.318	0.783	0.333	**0.819**	0.311
gemma2_voice	0.822	0.319	0.776	0.335	0.789	0.329
gemma2_style	0.796	0.328	0.765	0.333	0.761	0.336
gemma2_style_instruction	0.544	0.423	0.675	0.379	0.566	0.409
llama3.1_identity	0.755	0.359	0.682	0.372	0.698	0.368
llama3.1_voice	0.644	0.398	0.559	0.408	0.639	0.383
llama3.1_style	0.567	0.416	0.537	0.412	0.577	0.401
llama3.1_style_instruction	0.464	0.416	0.692	0.374	0.396	0.391
qwen2.5_identity	0.819	0.333	0.803	0.323	0.795	0.338
qwen2.5_voice	0.819	0.333	0.817	0.313	0.811	0.324
qwen2.5_style	0.733	0.402	0.768	0.355	0.772	0.349
qwen2.5_style_instruction	0.791	0.338	0.800	0.324	0.770	0.345

Table 6: Meaning similarity within rewrite pairs (feminine vs. masculine) and meaning preservation between each rewrite and the reference text. Bold numbers indicate the highest similarity/preservation in each setting.

Pilot Study on Perception of Gendered Style Contrast in LLM-generated Texts

Consent Form

You are invited to participate in a pilot study designed to explore perceptions of contrasted linguistic style in Large Language Model(LLM)-generated texts. Before you decide to participate, it is important that you understand why this study is being conducted and what your participation involves. Please read the following information carefully.

Description of the Research Study

In this study, we aim to investigate **whether two LLM-generated texts sound different in gendered style**, especially along a feminine–masculine dimension. By collecting human judgments about stylistic contrast, we hope to identify what factors influence one generated text to sound more feminine or masculine than another. For each pair of short texts, you will rate how strongly the two generated texts differ in feminine versus masculine style, and judge how similar the two texts are in meaning/content and in fluency or grammar. This research can help support future work on style transfer and AI-based writing assistance.

Risks and Benefits

The risks associated with this pilot study are minimal and comparable to those encountered during routine computer-based tasks, such as mild fatigue or boredom. Texts included in this study are generated by LLMs, and may occasionally include words that could be sensitive or uncomfortable, though no extreme or offensive material is intentionally included. The texts included in this study do not necessarily reflect researchers' views.

The primary benefit of participation is contributing to the understanding in the gendered style contrast in LLM-generated text pairs.

Time Required

Your participation will take an estimated 15 minutes. The time required may vary on an individual basis.

Voluntary Participation

Participation in this study is entirely voluntary. You may choose not to participate or withdraw from the study at any point without explanation. If you decide to withdraw, your data will not be included in the analysis, and you will not be paid.

Confidentiality

Your responses will remain completely anonymous. Please refrain from sharing any personally identifiable information during the study. The researchers will take all necessary steps to ensure the confidentiality of your contributions.

Consent

Please indicate below that you are at least 18 years old, have read and understood this consent form, are comfortable using English to complete the task, and agree to participate in this research study.

- I am 18 years old or older.
- I have read this consent form or had it read to me.
- My mother tongue is English.
- I agree to participate in this research study and wish to proceed with the annotation task.

If you give your consent to take part, please select 'I agree' below.

I agree

Figure 6: Style contrast task: consent form for annotators.

In this study, you will compare **10** pairs of short texts. Your main task is to judge how strongly the two texts differ along a feminine–masculine stylistic dimension based on your intuitions, while also considering how similar they are in meaning and in fluency or grammatical acceptability. Please base your judgments on **how the texts are generated**, such as tone, wording, and sentence structure, rather than what the texts are about and the length of the texts alone.

For each pair, you will answer the following questions:

- **Style contrast:** How strongly are the two texts contrasted along the feminine–masculine style dimension?
- **Meaning similarity:** How similar are the meanings or content of the two texts?
- **Grammar / fluency similarity:** How similar are the two texts in fluency or grammatical acceptability?
- **Confidence:** How confident are you in your judgments?

1. Style contrast

A pair has low gendered style contrast if the two texts sound gendered stylistically similar. A pair has high gendered contrast if the two texts feel far apart in tone, word choice, emotional expression, directness, or sentence structure. For more information, please see the example on the next page.

2. Meaning similarity

This question asks how similar the two texts are in meaning or content. Even if the style is very different, the texts may still express a similar core idea. If the meaning is very similar, give a higher rating. If the texts express different ideas, give a lower rating.

3. Grammar / fluency similarity

This question asks whether the two texts are at a similar level of fluency or grammatical acceptability, regardless of the style. If both texts feel similarly natural and well-formed, give a higher rating. If one text feels much less fluent or less grammatically acceptable than the other, give a lower rating. Please see the example on the next page.

4. Confidence

Finally, you will indicate how confident you are in your judgments. Use this rating to reflect how certain or uncertain you feel about your answers.

There are no strictly correct answers. Please feel free to use the comments section to explain the reasoning for your judgments.

Figure 7: Style contrast task: annotation guidelines for annotators.

Example of the Survey Task

Below is an example of the exact type of comparison you will make in the survey. The selected answers and explanations are shown only to illustrate how the task works. You are free to disagree with these example responses.

Text A

The project was completed on time, met all specifications, and achieved the required outcome efficiently.

Text B

Everyone worked so thoughtfully together, and it was really satisfying to see everything come together so smoothly in the end.

Style contrast

How strongly contrasted are these two texts in feminine vs masculine style?

1: Not contrasted at all	2: Slightly contrasted	3: Moderately contrasted	4: Strongly contrasted	5: Very strongly contrasted

Selected value: 5: Very strongly contrasted

Reasoning: The two texts feel far apart in style. Text A is concise, direct, and task-focused, while Text B is more expressive, relational, and descriptive. Because the stylistic distance between them is large, this pair would be rated as very strongly contrasted.

Meaning similarity

To what extent do the two texts express the same meaning or content?

1: Completely different	2: Mostly different	3: Partly similar	4: Mostly same	5: Same meaning

Figure 8: Style contrast task: examples for annotators.

Meaning similarity

To what extent do the two texts express the same meaning or content?

1: Completely different	2: Mostly different	3: Partly similar	4: Mostly same	5: Same meaning

Selected value: 4: Mostly same

Reasoning: Both texts describe a successful completion of a shared effort or task. Even though the wording and style differ, the core meaning is still fairly similar.

Grammar / fluency alignment

To what extent do the two texts have the same level of fluency / grammatical acceptability?

1: Very different	2: Somewhat different	3: Moderately similar	4: Mostly same	5: Same

Selected value: 5: Same

Reasoning: Both texts are fluent, natural, and grammatically acceptable. They differ in style, but not in overall readability or grammatical well-formedness.

Confidence Level

4: Very Confident

Reasoning: The contrast between the two texts is strong and relatively easy to identify, so this judgment can be made with high confidence.

Figure 9: Style contrast task: examples for annotators.

Survey Questions

Pair 1 of 10

Text A

Our group always rents bikes whenever we vacation. We explored so much territory and captured countless photos. Numerous ships sailed past this island harbor. This region isn't reachable by car. We paused here for a break since we had journeyed many miles. Back at the hotel, we snapped some pictures while chatting about dinner options.

Text B

Our group rents bikes on every vacation. We covered extensive ground and took heaps of photos. Many ships passed by the island port. This area is inaccessible by car. We stopped here to rest after traveling numerous miles. At the hotel, we took photos while discussing dinner plans.

Style contrast

How strongly contrasted are these two texts in feminine vs masculine style?

1: Not contrasted at all	2: Slightly contrasted	3: Moderately contrasted	4: Strongly contrasted	5: Very strongly contrasted

No value selected yet.

Meaning similarity

To what extent do the two texts express the same meaning or content?

1: Completely different	2: Mostly different	3: Partly similar	4: Mostly same	5: Same meaning

Figure 10: Style contrast task: survey questions for annotators.

Consent Form

You are invited to participate in a pilot study designed to explore perceptions of linguistic style in written text. Before you decide to participate, it is important that you understand why this study is being conducted and what your participation involves. Please read the following information carefully.

Description of the Research Study

In this study, we aim to investigate how readers perceive the style of written data as masculine, feminine, or gender-neutral. As an annotator, your task will involve evaluating a series of short texts based on their linguistic style, ranging from "Very Feminine" to "Very Masculine." This evaluation will focus on stylistic elements such as tone, word choice, and sentence structure rather than the content or topic of the text. Your contributions will help us create a dataset with gendered stylistic attributes, providing a foundation for understanding how people perceive gendered writing styles and the extent to which these perceptions align.

The findings of this study will contribute to scientific knowledge and may be included in academic publications.

Risks and Benefits

The risks associated with this pilot study are minimal and comparable to those encountered during routine computer-based tasks, such as mild fatigue or boredom. Texts included in this study are written by users on blog websites and social media platforms, and may occasionally include words that could be sensitive or uncomfortable, though no extreme or offensive material is intentionally included. The data included in this study are not authored by the researchers and do not necessarily reflect their views.

The primary benefit of participation is contributing to understanding in the field of language and perceived gender expression.

Time required

Your participation will take an estimated 30 minutes. The time required may vary on an individual basis.

Voluntary Participation

Participation in this study is entirely voluntary. You may choose not to participate or withdraw from the study at any point without explanation. If you decide to withdraw, your data will not be included in the analysis, and you will not be paid.

Figure 11: Perceived gendered style contrast task: consent form for annotators.

Guidelines for Annotating Masculine/Feminine Style from Texts

The goal of this study is to determine whether a text's style is perceived as masculine, feminine, or neutral. You will rate each text on the following scale:

1. **Very Feminine:** The text is strongly perceived as feminine based on linguistic style.
2. **Somewhat Feminine:** The text has some feminine characteristics, but they are not dominant.
3. **Neutral:** The text has no noticeable masculine or feminine characteristics.
4. **Somewhat Masculine:** The text has some masculine characteristics, but they are not dominant.
5. **Very Masculine:** The text is strongly perceived as masculine based on linguistic style.

Key Features of Feminine and Masculine Styles

These features are general tendencies and should guide, but not constrain, your perceptions. Base your rating on the overall impression of the text.

Feminine Style Tendencies

- **Emotional Expression:** Focus on feelings, relationships, empathy.
- **Collaborative Tone:** Use of inclusive language and hedging.
- **Descriptive Language:** Use of adjectives/adverbs and sensory details.
- **Complex Sentences:** Longer sentences with subordinate clauses or narrative flow.

Masculine Style Tendencies

- **Fact-Focused:** Emphasis on logic, data, or problem-solving.
- **Direct and Assertive:** Use of authoritative statements and commands.
- **Concise Language:** Short, to-the-point sentences with minimal elaboration.
- **Action-Oriented:** Preference for strong verbs and goal-driven language.

Neutral Style

- The text exhibits no clear tendencies toward either feminine or masculine linguistic features.

On the next page, you'll find examples showing how data are rated in each style for this study.

Figure 12: Perceived gendered style contrast task: examples for annotators.

Examples

Example: Text 1 I couldn't stop thinking about how kind and thoughtful her gesture was. It felt like a warm hug on a cold day, something I really needed. Perhaps it's silly to be so sentimental, but it meant the world to me.

Select a scale:

1: Very Feminine	2: Somewhat Feminine	3: Neutral	4: Somewhat Masculine	5: Very Masculine

Selected value: 1: Very Feminine

Confidence Level

4: Very Confident. You were very certain about your judgment with no hesitation

Reasoning

Emotional tone, descriptive language, and use of hedging create a strong feminine impression.

Example: Text 2 The atmosphere was calming, with soft lighting and gentle music in the background. It created a sense of peace and comfort that everyone seemed to enjoy.

Select a scale:

1: Very Feminine	2: Somewhat Feminine	3: Neutral	4: Somewhat Masculine	5: Very Masculine

Selected value: 2: Somewhat Feminine

Confidence Level

3: Moderately Confident. You felt reasonably sure of your judgment but had some doubts

Reasoning

Descriptive and sensory language, but less emotional depth or relational focus compared to the first example.

Figure 13: Perceived gendered style contrast task: examples for annotators.

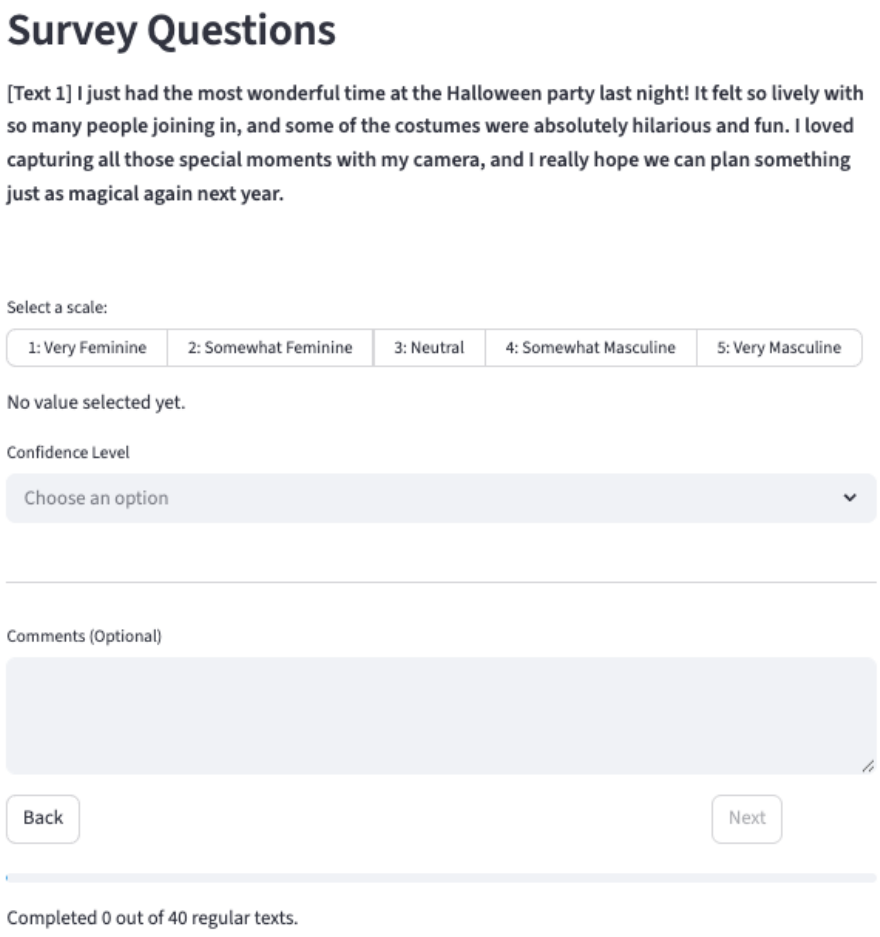

Figure 14: Perceived gendered style contrast task: survey questions for annotators.

Approach	Avg_Fem	Avg_Masc
human	0.787	0.809
deepseek_identity	0.906	0.889
deepseek_voice	0.904	0.880
deepseek_style	0.906	0.873
deepseek_style_instruction	0.920	0.864
gemma3_identity	0.917	0.891
gemma3_voice	0.910	0.888
gemma3_style	0.909	0.881
gemma3_style_instruction	0.916	0.891
gemma2_identity	0.917	0.917
gemma2_voice	0.932	0.923
gemma2_style	0.924	0.914
gemma2_style_instruction	0.931	0.915
llama3.1_identity	0.933	**0.937**
llama3.1_voice	0.930	0.923
llama3.1_style	**0.937**	0.923
llama3.1_style_instruction	0.922	0.906
qwen2.5_identity	0.896	0.897
qwen2.5_voice	0.887	0.885
qwen2.5_style	0.869	0.875
qwen2.5_style_instruction	0.915	0.909

Table 7: Average grammar acceptability of rewrites in each gendered style. The reference texts have the average grammar acceptability of .724. Bold numbers indicate the highest numbers in each column.

Approach	masc⇔fem		fem→ref		masc→ref	
	mean	std	mean	std	mean	std
human	0.479	0.227	0.634	0.179	0.519	0.241
deepseek_identity	0.933	0.050	0.913	0.072	0.928	0.065
deepseek_voice	0.897	0.060	0.888	0.068	0.907	0.064
deepseek_style	0.908	0.056	0.901	0.063	0.922	0.056
deepseek_style_instructions	0.856	0.063	0.866	0.069	0.893	0.065
gemma3_identity	0.889	0.054	0.866	0.071	0.888	0.071
gemma3_voice	0.878	0.057	0.861	0.069	0.878	0.072
gemma3_style	0.862	0.065	0.854	0.069	0.884	0.070
gemma3_style_instructions	0.860	0.061	0.863	0.060	0.883	0.063
gemma2_identity	0.934	0.046	0.912	0.068	0.903	0.069
gemma2_voice	0.913	0.053	0.895	0.068	0.893	0.072
gemma2_style	0.896	0.052	0.882	0.068	0.883	0.073
gemma2_style_instruction	0.844	0.063	0.857	0.070	0.858	0.068
llama3.1_identity	0.900	0.057	0.845	0.091	0.843	0.084
llama3.1_voice	0.864	0.065	0.852	0.075	0.826	0.080
llama3.1_style	0.838	0.067	0.834	0.077	0.810	0.084
llama3.1_style_instructions	0.838	0.064	0.812	0.079	0.879	0.069
qwen2.5_identity	0.948	0.043	0.930	0.061	0.923	0.058
Qwen2.5_voice	0.950	0.044	0.925	0.065	0.925	0.060
qwen2.5_style	0.945	0.065	0.912	0.085	0.922	0.074
qwen2.5_style_instructions	0.946	0.041	0.916	0.057	0.929	0.054

Table 8: Average cosine similarity

Approach	Fem → Ref	Masc → Ref	Fem ⇔ Masc
human	-6.922	-9.323	-12.763
qwen2.5_identity	13.062	13.850	10.125
qwen2.5_voice	13.491	13.708	10.264
qwen2.5_style	12.049	12.577	8.125
qwen2.5_style_instruction	12.210	14.414	10.060
deepseek_identity	12.825	14.152	10.410
deepseek_voice	13.560	14.431	10.673
deepseek_style	12.890	14.367	9.673
deepseek_style_instruction	9.458	15.810	8.268
gemma_identity	14.570	14.958	12.425
gemma_voice	14.487	15.290	12.254
gemma_style	12.582	15.208	10.413
gemma_style_instruction	11.678	15.132	9.921
gemma2_identity	14.721	14.280	11.797
gemma2_voice	14.218	14.564	11.541
gemma2_style	13.862	14.519	10.597
gemma2_style_instruction	11.608	15.136	9.451
llama_identity	12.312	12.801	10.121
llama_voice	10.970	10.627	9.461
llama_style	9.215	8.950	6.571
llama_style_instruction	3.924	12.887	6.649

Table 9: Aggregate semantic evaluation results computed using ParaPLUIE across models and prompting conditions, including meaning preservation between rewrites and reference texts (feminine→reference, masculine→reference) and meaning similarity between rewrite pairs (average of feminine→masculine and masculine→feminine scores).

Model	Cond.	**MS**: F ⇔ M		**MP**: F → Ref		M → Ref	
		t	p	t	p	t	p
DeepSeek-V3	style	**20.59**	$< .001$	**16.10**	$< .001$	**8.54**	$< .001$
Gemma-2-27B	style	**20.34**	$< .001$	**16.87**	$< .001$	**8.55**	$< .001$
Gemma-3-27B	style	7.58	$< .001$	5.92	$< .001$	2.65	.008
Llama-3.1-8B	style	8.34	$< .001$	8.82	$< .001$	5.26	$< .001$
Qwen2.5-7B	style	0.85	.394	1.35	.177	-0.15	.879

Table 10: Paired t-tests for meaning similarity (MS) and meaning preservation (MP) comparing *style* and *style instruction* prompting conditions across models. Positive t values indicate lower meaning similarity or preservation under *style instruction* prompting relative to the *style* condition.

Model	Cond.	**Avg.feminine**		**Avg. masculine**	
		t	p	t	p
DeepSeek-V3	style	-2.63	.009	1.08	.282
Gemma-2-27B	style	-3.28	.001	0.93	.354
Gemma-3-27B	style	**-5.65**	$< .001$	-0.33	.745
Llama-3.1-8B	style	**-3.96**	$< .001$	0.37	.715
Qwen2.5-7B	style	-2.19	.029	-1.55	.123

Table 11: Paired t-tests for grammaticality across prompting conditions. Negative t values indicate lower grammaticality scores under *style instruction* prompting relative to the *style*.

Model	Cond.	**JS Dist.**		Δf		Δm	
		t	p	t	p	t	p
DeepSeek-V3	style	-22.25	$< .001$	-21.51	$< .001$	-13.10	$< .001$
Gemma-2-27B	style	-23.99	$< .001$	-19.83	$< .001$	-18.33	$< .001$
Gemma-3-27B	style	-15.49	$< .001$	-7.02	$< .001$	-11.01	$< .001$
Llama-3.1-8B	style	-18.03	$< .001$	-11.10	$< .001$	-9.41	$< .001$
Qwen2.5-7B	style	-8.24	$< .001$	-4.46	$< .001$	-4.13	$< .001$

Table 12: Paired t-tests for style-related metrics comparing *style* and *style instruction* conditions. Negative t values indicate higher scores under the instruction condition relative to the base condition.

	Meaning Similarity		Grammar				Style Shift				Style Distance	
Condition	Mean	SD	Avg_{fem}	SD	Avg_{masc}	SD	Δ_f	SD	Δ_m	SD	Mean	SD
Style	0.761	0.028	0.896	0.006	0.835	0.021	0.062	0.003	0.037	0.012	0.111	0.012
Style + Instruction	0.553	0.011	0.920	0.001	0.857	0.006	0.121	0.010	0.053	0.006	0.201	0.001

Table 13: Aggregate results for meaning similarity, grammaticality, style shift, and style distance across prompting conditions using DeepSeek-V3. Results are averaged over five random seeds (42, 128, 256, 512, and 1024) for each prompting condition (*style* and *style instruction*).

	Meaning Similarity		Grammar				Style Shift				Style Distance	
Condition	Mean	SD	Avg_{fem}	SD	Avg_{masc}	SD	Δ_f	SD	Δ_m	SD	Mean	SD
Style	0.786	0.022	0.898	0.007	0.846	0.023	0.063	0.003	0.030	0.013	0.104	0.013
Style + Instruction	0.576	0.014	0.921	0.001	0.863	0.005	0.121	0.016	-0.095	0.006	0.193	0.008

Table 14: Aggregate results for meaning similarity, grammaticality, style shift, and JS style distance across prompting conditions using DeepSeek-V3. We evaluate prompt sensitivity by varying the first sentence of the system prompt (original, shorter, and longer variants) for each condition (*style* and *style instruction*).

Prompt condition	**System prompt**
Shared template (all conditions)	**Hard constraints (must follow):** • Preserve meaning and facts exactly: entities, numbers, dates, locations, claims, intent, sentiment strength, tense, and modality. • Do not add new information, examples, justifications, or opinions. • Do not introduce stereotypes or demographic assumptions. Style is purely linguistic, not "gender roles." • Keep both rewrites natural and fluent. • Keep length within $\pm 15\%$ unless impossible. • If the input is too short to restyle naturally, still try, but do not pad with new content. **Output format (strict):** Return EXACTLY two lines, no extra text: • `FEMININE: <rewrite>` • `MASCULINE: <rewrite>`
Task	**Original:** You are generating a parallel dataset. For each input, create two rewrites that preserve the original meaning exactly, but differ in the specified target dimension. **Short:** Generate two stylistically distinct rewrites for each input while preserving the original meaning exactly **Long:** You are tasked with generating a parallel dataset. For every input, produce two rewritten versions that retain the exact original meaning while varying only in linguistic style.
Gendered style	**Difference dimension:** linguistic style. **Style targets:** Feminine Style Tendencies (linguistic) **and** Masculine Style Tendencies (linguistic).
Gendered style + fine-grained instructions	**Difference dimension:** linguistic style. **Style targets:** Feminine Style Tendencies (linguistic) **and** Masculine Style Tendencies (linguistic). **Feminine Style Tendencies (linguistic):** • More emotional/relational framing and empathy markers without changing sentiment strength. • More collaborative tone: inclusive language when consistent, hedges such as maybe, perhaps, or I think. • More descriptive language: adjectives/adverbs and sensory/aesthetic framing only if already implied. • Slightly more complex/flowing sentences, such as subordinate clauses and narrative flow. **Masculine Style Tendencies (linguistic):** • More fact-focused framing: logic, structure, or problem-solution framing without adding data. • More direct/assertive phrasing, while keeping the same level of certainty as the original. • More concise language: shorter, to-the-point sentences with minimal elaboration. • More action-oriented verbs and goal-driven framing without changing intent.

Table 15: Prompt templates used for prompt variation experiments on DeepSeek-V3 with short, original, and long task descriptions. Shared constraints and stylistic targets remain unchanged across variants.

Prompt condition	System prompt
Shared template (all conditions)	**Task.** You are generating a parallel dataset. For each input, create two rewrites that preserve the original meaning exactly, but differ in the specified target dimension. **Hard constraints (must follow):** • Preserve meaning and facts exactly: entities, numbers, dates, locations, claims, intent, sentiment strength, tense, and modality. • Do not add new information, examples, justifications, or opinions. • Do not introduce stereotypes or demographic assumptions. Style is purely linguistic, not "gender roles." • Keep both rewrites natural and fluent. • Keep length within $\pm 15\%$ unless impossible. • If the input is too short to restyle naturally, still try, but do not pad with new content. **Output format (strict):** Return EXACTLY two lines, no extra text: • `WOMAN/FEMALE/FEMININE: <rewrite>` • `MAN/MALE/MASCULINE: <rewrite>`
Gender identity + fine-grained instructions	**Difference dimension:** authorial identity. **Identity targets:** author who identifies as a woman **and** author who identifies as a man. **Woman-author tendencies:** • Rewrite the text as if written by a woman author. • Imagine approaching the rewrite from this author background when choosing wording, tone, phrasing, and expression. • Rewrite only through linguistic choices, without changing the underlying meaning or content. **Man-author tendencies:** • Rewrite the text as if written by a man author. • Imagine approaching the rewrite from this author background when choosing wording, tone, phrasing, and expression. • Rewrite only through linguistic choices, without changing the underlying meaning or content.
Gendered voice + fine-grained instructions	**Difference dimension:** authorial voice. **Voice targets:** FEMALE-author voice tendencies **and** MALE-author voice tendencies. **FEMALE-author voice tendencies:** • Rewrite the text in a female-author voice. • Imagine approaching the rewrite from this author background while emphasizing the authorial voice through wording, tone, phrasing, rhythm, and expression. • Rewrite only through linguistic choices, without changing the underlying meaning or content. **MALE-author voice tendencies:** • Rewrite the text in a male-author voice. • Imagine approaching the rewrite from this author background while emphasizing the authorial voice through wording, tone, phrasing, rhythm, and expression. • Rewrite only through linguistic choices, without changing the underlying meaning or content.
Gendered style + fine-grained instructions	**Difference dimension:** linguistic style. **Style targets:** Feminine Style Tendencies (linguistic) **and** Masculine Style Tendencies (linguistic). **Feminine Style Tendencies (linguistic):** • More emotional/relational framing and empathy markers without changing sentiment strength. • More collaborative tone: inclusive language when consistent, hedges such as maybe, perhaps, or I think. • More descriptive language: adjectives/adverbs and sensory/aesthetic framing only if already implied. • Slightly more complex/flowing sentences, such as subordinate clauses and narrative flow. **Masculine Style Tendencies (linguistic):** • More fact-focused framing: logic, structure, or problem-solution framing without adding data. • More direct/assertive phrasing, while keeping the same level of certainty as the original. • More concise language: shorter, to-the-point sentences with minimal elaboration. • More action-oriented verbs and goal-driven framing without changing intent.

Table 16: Length-matched prompt templates used for robustness analyses across gender identity, gendered voice, and gendered style conditions. All prompts share the same task description, meaning-preservation constraints, output format, and length constraint; only the target dimension and corresponding guidance differ across conditions.

Model	Comparison	JS Dist.		Δf		Δm	
		t	p	t	p	t	p
DeepSeek-V3	Base vs Instr. (Identity)	-9.52	$< .001$	-6.22	$< .001$	-5.12	$< .001$
	Base vs Instr. (Voice)	-8.59	$< .001$	-13.41	$< .001$	-0.00	.996
	Base vs Instr. (Style)	-22.25	$< .001$	-21.51	$< .001$	-13.10	$< .001$
Gemma-2-27B	Base vs Instr. (Identity)	-6.56	$< .001$	-8.02	$< .001$	-3.44	$< .001$
	Base vs Instr. (Voice)	-4.41	$< .001$	-8.48	$< .001$	1.39	.164
	Base vs Instr. (Style)	-23.99	$< .001$	-19.83	$< .001$	-18.33	$< .001$
Gemma-3-27B	Base vs Instr. (Identity)	-12.72	$< .001$	-13.03	$< .001$	-9.77	$< .001$
	Base vs Instr. (Voice)	-5.89	$< .001$	-10.38	$< .001$	-1.04	.298
	Base vs Instr. (Style)	-15.49	$< .001$	-7.02	$< .001$	-11.01	$< .001$
Llama-3.1-8B	Base vs Instr. (Identity)	-7.52	$< .001$	-11.14	$< .001$	0.48	.632
	Base vs Instr. (Voice)	-2.09	.038	-6.39	$< .001$	-0.53	.597
	Base vs Instr. (Style)	-18.03	$< .001$	-11.10	$< .001$	-9.41	$< .001$
Qwen2.5-7B	Base vs Instr. (Identity)	-3.24	.001	-8.84	$< .001$	2.08	.038
	Base vs Instr. (Voice)	-3.69	$< .001$	-8.00	$< .001$	2.79	.005
	Base vs Instr. (Style)	-8.24	$< .001$	-4.46	$< .001$	-4.13	$< .001$

Table 17: Paired t-tests for style-related metrics comparing base and instruction prompting conditions. Negative t values indicate higher scores under the instruction condition relative to the base condition.

Model	Cond.	**MS**: F ⇔ M		**MP**: F → Ref		M → Ref	
		t	p	t	p	t	p
deepSeekV3	id.	0.46	.649	1.53	.126	0.28	.778
	voice	1.27	.204	0.76	.445	1.13	.258
	style	**20.59**	$< .001$	**16.10**	$< .001$	**8.54**	$< .001$
gemma2	id.	0.78	.436	3.13	.002	1.31	.192
	voice	3.28	.001	2.81	.005	0.45	.655
	style	**20.34**	$< .001$	**16.87**	$< .001$	**8.55**	$< .001$
gemma3	id.	2.57	.010	4.54	$< .001$	0.68	.496
	voice	1.71	.088	4.54	$< .001$	-0.83	.408
	style	7.58	$< .001$	5.92	$< .001$	2.65	.008
Llama3.1	id.	10.11	$< .001$	11.02	$< .001$	5.86	$< .001$
	voice	8.11	$< .001$	8.88	$< .001$	4.82	$< .001$
	style	8.34	$< .001$	8.82	$< .001$	5.26	$< .001$
Qwen2.5	id.	-0.15	.879	1.30	.193	-0.92	.358
	voice	-0.39	.700	0.58	.565	-1.12	.262
	style	0.85	.394	1.35	.177	-0.15	.879

Table 18: Paired t-tests for meaning similarity (MS) and meaning preservation (MP) comparing base and instruction prompting conditions across models and gendered style operationalizations (*identity*, *voice*, and *style*). Positive t values indicate lower meaning similarity or preservation under instruction prompting relative to the corresponding base condition.

Model	Comparison	Feminine		Masculine	
		t	p	t	p
DeepSeek	Base vs Instr. (Identity)	-1.54	.124	-1.04	.298
	Base vs Instr. (Voice)	-3.26	.001	-0.44	.663
	Base vs Instr. (Style)	-2.63	.009	1.08	.282
Gemma-2-27B	Base vs Instr. (Identity)	-1.49	.137	0.96	.338
	Base vs Instr. (Voice)	1.72	.087	1.57	.118
	Base vs Instr. (Style)	-3.28	.001	0.93	.354
Gemma-3-27B	Base vs Instr. (Identity)	-0.98	.329	-0.76	.447
	Base vs Instr. (Voice)	-2.94	.003	1.27	.204
	Base vs Instr. (Style)	-5.65	$< .001$	-0.33	.745
Llama-3.1-8B	Base vs Instr. (Identity)	-3.27	.001	-1.53	.128
	Base vs Instr. (Voice)	-1.08	.281	-1.47	.144
	Base vs Instr. (Style)	**-3.96**	$< .001$	0.37	.715
Qwen2.5-7B	Base vs Instr. (Identity)	-0.48	.631	-0.37	.712
	Base vs Instr. (Voice)	-2.70	.007	-0.51	.609
	Base vs Instr. (Style)	-2.19	.029	-1.55	.123

Table 19: Paired t-tests for grammaticality across prompting conditions. Negative t values indicate lower grammaticality scores under instruction prompting relative to the base condition.

Target	**Text**	**Metrics**
Reference	i just left work when i came upon a car that flipped over . it was so close to the cement blockade but it didn't quite reach it . there was no one in the car | thank goodness | but it looked pretty beat up . there were other people around besides me to see if the owners needed help . i left the scene shortly after taking pictures .	**PGS:** [0.328,**0.400**,0.272] **Grammar:** 0.803
Style instruction		
Feminine	I had just finished work when I came across a car that had flipped over. It was so incredibly close to the cement barrier, but thankfully didn't quite hit it. To my relief, there was no one inside, though the car looked quite damaged. A few other caring people were also nearby, checking if the owners needed assistance. I took some pictures before leaving the scene shortly after.	**PGS:** [**0.399**, 0.310, 0.291] **MP:** 0.438 **MS:** 0.772 **Grammar:** 0.925
Masculine	After work, I encountered an overturned car near a cement blockade. It narrowly missed the barrier. No one was in the vehicle, but it was heavily damaged. Others were present to offer help. I documented it with photos and then left.	**PGS:** [0.221, **0.428**, 0.351] **MP:** 0.687 **MS:** 0.772 **Grammar:** 0.972
Style		
Feminine	I had just finished work when I encountered a flipped-over car. It was positioned so near the concrete barrier, though it hadn't actually collided with it. Thankfully, no one was inside the vehicle, but it appeared quite damaged. A few other individuals were also present, checking if the owners required assistance. I departed the area not long after capturing some photos.	**PGS:** [0.314, **0.362**, 0.324] **MP:** 0.748 **MS:** 0.973 **Grammar:** 0.824
Masculine	I'd just left work when I saw a car that had flipped over. It was really close to the cement barrier but didn't hit it. Nobody was in the car, luckily, though it looked pretty banged up. Other people were around too, seeing if the owners needed help. I left shortly after taking pictures.	**PGS:** [**0.364**, 0.347, 0.289] **MP:** 0.932 **MS:** 0.973 **Grammar:** 0.966

Table 20: DeepSeek-V3 example illustrating the contrast–fidelity trade-off. PGS= $[p(F), p(M), p(N)]$; MP = meaning preservation (vs. reference); MS = meaning similarity (within pair). Color highlights mark lexical changes: olive frequent paraphrases; teal content from reference paraphrased in that rewrite; blue new phrases not present in the reference.

		Meaning Similarity: Fem ⇔ Masc				Meaning Preservation							
		Base		Instr.		Masc → Ref Base		Masc → Ref Instr.		Fem → Ref Base		Fem → Ref Instr.	
Model	Approach	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Human	Identity	0.052	0.182			0.060	0.136			0.053	0.127		
Deepseek-v3	Identity	0.810	0.339	**0.835**	0.297	0.833	0.308	**0.839**	0.290	0.806	0.327	0.807	0.312
	Voice	**0.831**	0.304	0.821	0.314	0.835	0.295	0.822	0.306	0.793	0.321	0.793	0.319
	Style	0.811	0.319	0.570	0.418	**0.847**	0.283	0.767	0.339	0.803	0.319	0.611	0.400
Gemma3	Identity	0.799	0.323	0.770	0.340	0.801	0.319	0.785	0.321	0.763	0.334	0.744	0.345
	Voice	0.777	0.335	0.758	0.348	0.772	0.329	0.757	0.341	0.738	0.351	0.710	0.363
	Style	0.654	0.396	0.558	0.422	0.794	0.317	0.759	0.339	0.662	0.376	0.581	0.399
Gemma2	Identity	**0.826**	0.318	**0.832**	0.294	0.783	0.333	0.777	0.334	**0.819**	0.311	0.789	0.326
	Voice	0.822	0.319	0.804	0.318	0.776	0.335	0.776	0.330	0.789	0.329	0.767	0.334
	Style	0.796	0.328	0.544	0.423	0.765	0.333	0.675	0.379	0.761	0.336	0.566	0.409
Llama3.1	Identity	0.755	0.359	0.582	0.416	0.682	0.372	0.505	0.402	0.698	0.368	0.524	0.392
	Voice	0.644	0.398	0.492	0.419	0.559	0.408	0.436	0.410	0.639	0.383	0.496	0.404
	Style	0.567	0.416	0.464	0.416	0.537	0.412	0.692	0.374	0.577	0.401	0.396	0.391
Qwen2.5	Identity	0.819	0.333	**0.836**	0.309	0.803	0.323	0.825	0.307	0.795	0.338	0.789	0.336
	Voice	0.819	0.333	0.827	0.323	0.817	0.313	0.795	0.328	0.811	0.324	0.768	0.344
	Style	0.733	0.402	0.791	0.338	0.768	0.355	0.800	0.324	0.772	0.349	0.770	0.345

Table 21: Meaning similarity and meaning preservation by model, approach, and instruction condition.

		Base		Instr.	
Model	Approach	Avg_{fem}	Avg_{masc}	Avg_{fem}	Avg_{masc}
Human	Identity	0.787	0.809		
Deepseek-v3	Identity	0.906	0.889	0.914	0.895
	Voice	0.904	0.880	0.921	0.883
	Style	0.906	0.873	0.920	0.864
Gemma3	Identity	0.917	0.891	0.915	0.895
	Voice	0.910	0.888	0.920	0.885
	Style	0.909	0.881	0.916	0.891
Gemma2	Identity	0.917	0.917	0.924	0.912
	Voice	0.932	0.923	0.926	0.916
	Style	0.924	0.914	0.931	0.915
Llama3.1	Identity	0.933	0.937	0.935	0.930
	Voice	0.930	0.923	0.932	0.925
	Style	0.937	0.923	0.922	0.906
Qwen2.5	Identity	0.896	0.897	0.880	0.889
	Voice	0.887	0.885	0.887	0.882
	Style	0.869	0.875	0.915	0.909

Table 22: Average grammar acceptability of rewrites in each gendered style. The reference texts have an average grammar acceptability of .724.