Contrast Comes at a Cost: Perception-centered Gendered Style Transfer with Large Language Models

Anonymous ACL submission

Abstract

Gendered style transfer is the task of rewriting text toward a feminine or masculine style while preserving its core meaning. While prior work typically operationalizes the target style using author identity labels, these categories do not always align with gendered linguistic expression. Therefore, we propose a perception-centered framework that rewrites text to be more likely *perceived* as feminine or masculine and examines whether perceived gendered style can be systematically induced through text rewriting. Using several instruction-tuned LLMs, we generate paired rewrites of the same input text under four target style operationalizations: author identity, gendered voice, general style descriptions, and fine-grained guidelines. We evaluate these rewrites based on stylistic change, meaning preservation, and grammatical acceptability, and compare them with human-authored rewrites. Our results show that LLMs can produce distinct perceived gendered styles, though outcomes depend strongly on prompt formulation. Fine-grained guidelines produce the strongest stylistic contrast, while identity-based prompts better preserve meaning. 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 authors' gender identity labels.[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 transfer is typically operationalized through *author identity labels* (e.g., male vs. female authors), framing the task as converting between these categories (Kang et al., 2019; Sudhakar et al., 2019). However, transferring between male-authored and female-authored texts does not necessarily correspond to transferring between masculine and feminine expression, since gendered language can be perceived independently of the author's identity (Troiano et al., 2023; Nguyen et al., 2014). 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 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. Building on recent advances in generating controlled rewrites with Large Language Models (LLMs) (Lu et al., 2023;

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

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 framing* (e.g., writing as a woman or a man), *voice-* and *style-oriented formulations*, and a variant with *fine-grained stylistic guidelines*. We evaluate the resulting rewrites with a classifier trained on human annotations from Chen et al. (2025) and compare to rewrites from authors with different gender identities (Kang et al., 2019)—an operationalization commonly used in gendered style transfer. Our results show that LLMs can generate rewrite pairs that differ in perceived gendered style while largely preserving meaning (§ 4), but 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, 1990, 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.

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 intended 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.

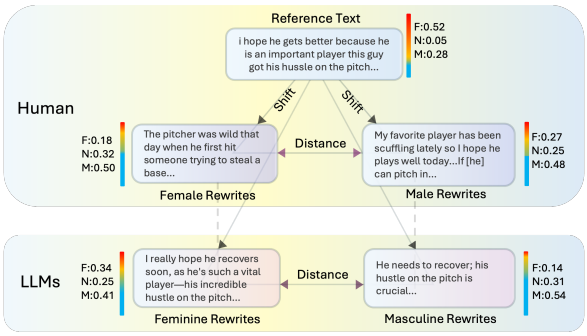

Figure 1: Methodology overview: a reference text (top) is rewritten into feminine and masculine variants by human authors (middle) or LLMs (bottom). A classifier assigns probability distributions over feminine (F), neutral (N), and masculine (M) styles, indicated by color bars. Style shift measures the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

3.1 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 ("denotation"). The Visual Story Telling (ViST) dataset (Huang et al., 2016), containing human-written five-sentence narratives paired with image sequences, is used as the *source texts* for keyword extraction. Because multiple annotators independently describe the same denotation, the dataset contains parallel stories that convey the same underlying situation while differing in linguistic realization.

We then sample 500 *source texts* from PASTEL as *reference texts*. To construct the *human-written* gender-contrasted pairs, we find all stories written for the same reference text and randomly select one *female-authored* and one *male-authored*.[2]

[2] If multiple rewrites are available per gender for the same story, we select one per gender using a fixed sampling rule to

3.2 Models

We use five instruction-tuned, open-source LLMs to generate gender-contrasted rewrite pairs. Our model set covers three widely used families (Llama, Gemma, and Qwen), and additionally includes DeepSeek as a large open model. In total, we evaluate DeepSeek-V3 (Liu et al., 2024a), google/Gemma-3-27B-it (Kamath et al., 2025), google/Gemma-2-27B-it (Riviere et al., 2024), Qwen/Qwen2.5-7B-Instruct (Yang et al., 2024), and meta-llama/Llama-3.1-8B-instruct (Grattafiori et al., 2024), spanning sizes from 7B to 27B.

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.

To investigate how different operationalizations of *gendered style* in the prompt affect the resulting rewrites, we test four prompting variants corresponding to the conceptual distinctions explained in §2: gender identity, gender stance/voice, and gender expression/performance (see full prompts in Table 3 in §A). 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. Finally, in *gendered style with fine-grained instructions*, 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.

All LLMs are prompted with an instruction template (see §A.1 in §A) 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 above). Across all prompting conditions, we ask for the same hard constraints: (i) preserve meaning and factual content exactly (entities, numbers, dates, locations, claims, intent, sentiment strength, tense, and modality); (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 constrains about naturalness and output length. Models are instructed to output two lines in a fixed format, with prefixes indicating the target condition, for example, for style prompts:

```
Feminine: ...
Masculine: ...
```

3.4 Evaluation

We evaluate human 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 Appendix A.4.

Perceived Gendered Style To assess gendered style contrast in rewrites we train a perceived gendered style classifier on the dataset of Chen et al. (2025). It contains 510 texts, each rated by 10 crowdworkers on a five-point feminine–masculine scale, and each annotation includes a self-reported confidence score. We collapse the original five-point scale into three categories: F (very/somewhat feminine), N (neutral), and M (very/somewhat masculine) and train models to predict a three-way probability distribution for a text t:

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

We experimented with four classifiers: a random forest (Breiman, 2001), a feature-based Multilayer Perception, and BERT-based classifiers (Devlin et al., 2018), using 5-fold cross-validation. We evaluate models both against majority-vote labels and the rated style distribution using Jensen-Shannon (JS) distance. Among these classifiers, the BERT-based model trained with per-annotation confidence weighting achieved the best overall performance (Table 4 and Table 5 in §A.2). We therefore use this confidence-weighted BERT classifier for all rewrites in our evaluation.

Style Distance (Pairwise Contrast) We compute style distance between two rewrites using the

ensure a single pair per story.

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 for masculine shift $\Delta_M(t', t) = p(M \mid t') - p(M \mid t)$. 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 Core meaning of a text can be preserved without direct word overlap (e.g., via paraphrases and synonyms). To capture meaning preservation beyond surface 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 a pair of texts (t'_F, t'_M), we report a *bidirectional entailment* score defined as the minimum of the two directional entailment probabilities. We use the minimum to ensure that high scores reflect strong semantic consistency in *both* directions (i.e., neither text introduces information not supported by the other):

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

We use this metric as our primary indicator of content preservation: higher scores indicate stronger semantic consistency, i.e., more meaning preserved within the pair and relative to the reference. For comparison, we also compute embedding-based cosine similarity using Sentence-BERT (Abdul-Mageed et al., 2019).

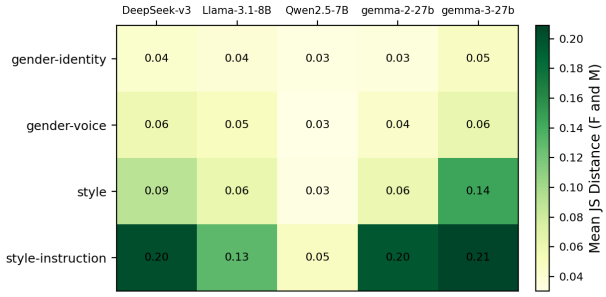

Figure 2: Style Distance between perceived feminine and masculine style. Each cell shows the Jensen-Shannon (JS) distance between the predicted style distributions of t'_F and t'_M for a given condition. Darker green indicates greater stylistic contrast. The average style distance for human rewrite pairs 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?* To answer this question, we analyze perceived gendered stylistic variation produced by both human and LLM-generated rewrites. We quantify the intensity of this variation both *within* each rewrite pair and *relative* to the reference text. A successful gender-contrasted rewrite 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 results of the style distance for all tested models. A corresponding result for human rewrite pairs is .12. We can observe that the largest average pairwise distance is observed under the fine-grained style instruction-based setting, especially for Gemma-3 (mean JS = .21) and DeepSeek-V3 and Gemma-2 (mean JS

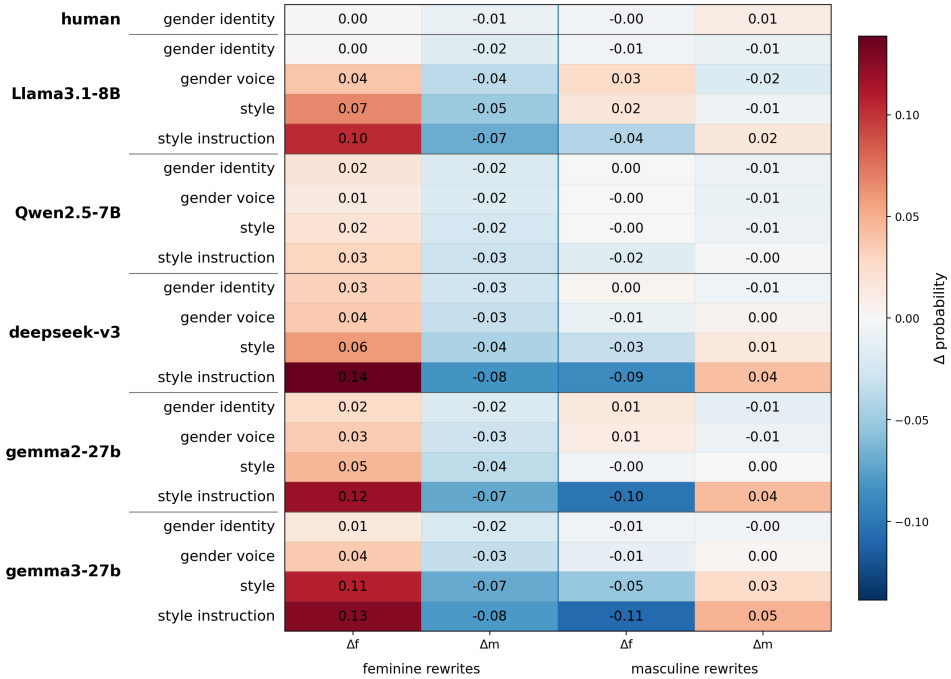

Figure 3: Style Shift relative to reference text. Each cell reports $\Delta_s = p(s \mid t') - p(s \mid t)$ for style class $s \in \{F, M\}$. Red denotes a positive shift (higher probability than the reference), and blue denotes a negative shift (lower probability than the reference). Thus, shifting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

= .20), followed by Llama-3. Human rewrites are comparable to Llama-3.1 and higher than all Qwen-based approaches, but lower than the style instruction-based outputs of the other models. This suggests that although the human pairs differ in author identity (female- vs. male-authored), they do not exhibit as strong a contrast in perceived gendered style as pairs produced under fine-grained style instructions. This is to say, prompting models with explicit linguistic style guidance yields a more contrastive perceived gendered style than identity-based variation alone.

Style Shift Figure 3 shows the results of the style shift of rewrites relative to the reference text. Across models, the *instruction* prompting setting generally produces a larger average shift toward the feminine style than other prompting approaches. For feminine style rewrites, the largest average shift is observed from DeepSeek-V3 under *style instruction*, which move further away from the reference texts toward a more strongly feminine style (PD = .14). Gemma-3 shows the second-largest shift (PD = .13), followed by Gemma-2 (PD = .12) and Llama-3.1 (PD = .10). In contrast, human female rewrites do not exhibit a significant shift relative to the reference style. A similar lack of significant shift is observed for Qwen models in the *gender identity, voice* and *style* conditions.

For masculine rewrites, the average shift from the reference texts toward a masculine style is moderate and less extreme than the corresponding shift observed for feminine rewrites. Still, the *style instruction* prompting approach outperforms the other approaches. *Style instruction* Gemma-3 exhibits the largest shift toward masculine style (PD = .05), while *style instruction* DeepSeek-V3 is slightly lower, with a shift only .01 smaller than Gemma-3. Interestingly, for Llama-3.1 under the *gender identity, voice* and *style* conditions, masculine rewrites shift on average more toward the feminine style than toward the masculine style.

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 condition yields the most contrasted gendered rewrite pairs for most models except those based on Qwen. At the same time, the *style-instructions* condition also produces 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 iden-*

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 1: 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.

tity prompting produces the least contrasted pairs, with lower style contrast than both *gender voice* and *gender style* prompting.

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 the meaning similarity within rewrite pairs (feminine vs. masculine, see details of the NLI-based metric in §3.4). Across all prompting approaches, the *gender voice* setting yields the highest meaning similarity across models within each model family, followed by the *style* condition and then the *style instruction* condition. Specifically, rewrite pairs generated by DeepSeek-V3 under the *gender voice* approach achieve the highest meaning similarity among all approaches (mean = .831), followed by Gemma-2 with *gender identity* (mean = .826) and Qwen2.5 with *gender identity&voice* (mean = .819). As a reference baseline, human rewrite pairs show the lowest meaning similarity (mean = .052).

We corroborate these findings with the similarity measure based on Sentence-BERT. The overall trend is the same as for the NLI-based metric: the *gender identity* prompting condition yields the highest meaning similarity and preservation among prompting approaches apart from Qwen models. We therefore report these results in Table 6 in §A.3.

Meaning Preservation Table 1 (four right columns) shows the results of the meaning preservation of rewrites relative to the reference texts. For each rewrite style, masculine rewrites generated by DeepSeek-V3 under the *style* condition preserve the most meaning relative to the reference texts (mean = .847), followed by DeepSeek-V3 under the *gender voice* condition (mean = .835). As a reference baseline, the human masculine rewrites show the lowest meaning preservation relative to the reference texts.

For feminine rewrites, Gemma-2 under the *gender identity* condition preserves the most meaning relative to the reference texts (mean = .819), followed by Qwen2.5 under the *gender voice* condi-

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 2: 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.

tion (mean = .811). Human female rewrites again show the lowest meaning preservation relative to the reference texts (mean = .053).

Grammar Acceptance Table 2 shows the average grammar acceptance rewrites in each gendered style across human and LLMs rewrites. Overall, the average grammatical acceptability of LLM-generated rewrites in both styles is higher than that of human rewrites. Llama3.1 under the *style* condition achieves the highest scores (fem_mean = .937; mas_mean = .923), exceeding even the average acceptability of the reference texts. As a baseline, human rewrites in both gendered styles have the lowest average grammatical acceptability, but they remain the closest to the reference texts (avg_fem = .787; avg_masc = .809).

These results indicate that LLM-generated rewrites (both feminine and masculine) are, on average, more grammatically acceptable than human rewrites and also more acceptable than the original reference texts. Unlike style contrast and meaning preservation, we do not observe a single prompting approach that consistently dominates across models. Although human rewrites receive lower acceptability scores than LLM outputs, they remain closest to the reference texts.

To answer **RQ2**, we find that LLM-generated rewrites consistently achieve higher meaning similarity within pairs and stronger meaning preservation relative to the reference than human rewrites. Within each model, the *gender identity* prompting condition yields the highest pairwise meaning similarity, whereas the *style instruction* condition tends to produce the lowest within-pair similarity (with Qwen2.5 as an exception). When comparing rewrites to the reference text, DeepSeek-V3 produces the most meaning-preserving *masculine* rewrites, while Gemma-2 achieves the strongest meaning preservation for *feminine* rewrites.

6 Discussion and Conclusion

In this work, we investigated whether LLMs can perform perception-centered gendered style transfer and how different prompt formulations affect the resulting stylistic contrast. We found that LLMs can indeed generate gender-contrasted pairs, but the results also reveal a key trade-off. The instruction-based prompting condition produces the strongest stylistic contrast, yet it tends to reduce meaning similarity within the pair, suggesting that stronger contrast is sometimes achieved alongside increased paraphrasing or content drift. In contrast, the gender-identity prompting condition yields the weakest stylistic contrast but the highest meaning similarity, indicating that minimal linguistic variation is introduced when gender is framed primarily as identity rather than as style. Qualitative examples (see Appendix A.4) suggest that this trade-off might 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 more on surface-level paraphrasing, 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.

7 Limitations

Our work has several limitations. First, our evaluation is primarily automatic: we operationalize perceived gendered style with a classifier trained on human judgments, we rely on NLI/embedding similarity for meaning preservation and a CoLA-based acceptability model for naturalness. While these metrics enable scalable comparison across models and prompting conditions, they reflect human evaluations of perceived contrast, meaning, and fluency only indirectly and potentially inaccurately. Moreover, as the supervised models applied here are necessarily conditioned on their training data, they may exhibit dataset-specific cues and imperfect calibration. We therefore interpret classifier scores primarily as a comparative signal across prompting conditions rather than as a definitive measure of gendered style, and leave broader robustness checks and direct human evaluation to future work.

Second, our comparisons are conducted on a set of 500 reference texts from only one source data set. A larger and more diverse evaluation set would strengthen conclusions about generality across domains, topics, and writing styles.

Third, our experiments focus on generating and evaluating *contrastive* feminine vs. masculine rewrites. Although the style classifier includes a neutral class, we do not explicitly generate neutral rewrites or analyze any gendered style on the spectrum between feminine and masculine, which limits conclusions about more fine-grained or non-binary expressions of gendered style.

Fourth, we do not conduct a detailed feature/topic-level analysis of what linguistic/topical cues each prompting operationalization elicits. In particular, we do not analyze which features drive the classifier's predictions, which cues LLMs use to increase perceived contrast, or how these cues differ from those in human rewrites. We observe a trade-off between perceived style contrast and meaning preservation, but we do not fully characterize its causes or implications. It remains unclear whether this trade-off reflects limitations of prompting and generation, limitations of our automatic metrics, or a more fundamental property of how gendered style is realized in text. In particular, we do not establish whether high-contrast perceived gendered style can be achieved while maintaining consistently high semantic fidelity, or what kinds of linguistic changes would be sufficient to do so.

Finally, we keep decoding and generation settings fixed across all experiments (e.g., temperature, top-p, and max tokens) to ensure comparability across models and prompting conditions. As a result, we do not analyze the sensitivity of perceived contrast, meaning preservation, or grammatical acceptability to inference-time parameters (e.g., temperature) or other generation controls. Future work could systematically vary decoding settings to quantify how robust the observed trade-offs are and whether particular configurations improve contrast without increasing content drift.

8 Ethics & Potential Risks

Although our work does not conceptualize gender as an author identity-based construct and instead focuses on *perceived* gendered style, our operationalization remains limited. In particular, although the underlying data for training perceived gendered style classifier includes a neutral category, our rewriting and analysis focus on contrastive feminine vs. masculine outputs and do not explicitly incorporate neutral as a target style. This framing risks reifying a binary axis of gender expression and may overlook non-binary or gender-neutral realizations.

Our intent of studying gendered style transfer with LLMs is twofold. First, we assess the potential of LLM prompting to generate pseudo-parallel, gendered-contrasted rewrite pairs at scale. Second, we examine whether the resulting LLM-generated rewrites align with or diverge from human rewrites under the same inputs and evaluation criteria. Importantly, our contribution is not only the comparison itself: the generated rewrite pairs—especially those that achieve strong perceived contrast—constitute an artifact that can support further analysis of subtle *style biases* in LLM behavior. For example, such outputs can be used to probe whether language perceived as more feminine or more masculine systematically affects downstream behavior in socially impactful tasks (e.g., advice-giving).

In the meantime, even a perception-centered framing of "feminine" and "masculine" style carries ethical risks. These categories can reinforce stereotypes and normative expectations that may become entangled with identity, particularly when models rely on culturally salient cues (e.g., associating femininity with emotionality or deference and masculinity with directness). We therefore

emphasize that our labels operationalize perceived gendered style within a specific dataset and context rather than defining gender itself, and we view expanded targets (including neutral and non-binary styles), and human evaluation for stereotypical content as important directions for future work.

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.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

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

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

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.

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 Doğruöz, 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

A.3 Cosine Similarity

We compute embedding-based cosine similarity using Sentence-BERT[3]. 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 $\cos(t, t'_M)$.

A.4 Example from LLMs rewrites

This section analyzes the differences in rewrites across two prompting approaches *style* and *style instruction* from Table 7. With an example from DeepSeek-V3, Table 7 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.5 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.

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

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: <rewrite>` • `MAN: <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	Loss	Acc.	F1	F1†
Random forest	1.12 ± 0.015	0.56 ± 0.046	0.42 ± 0.044	0.58 ± 0.09
MLP (key_extracted_features)	1.06 ± 0.016	0.55 ± 0.032	0.46 ± 0.027	0.59 ± 0.13
BERT-base-uncased	1.04 ± 0.011	0.61 ± 0.054	0.49 ± 0.041	0.60 ± 0.20
BERT-base-uncased (conf.-weighted)	1.031 ± 0.013	0.62 ± 0.059	0.50 ± 0.062	0.61 ± 0.19

Table 4: Classifier comparison for three-way perceived gendered style prediction (F/M/N) with 5-fold cross-validation (mean$\pm$std). Uniform baseline loss is $\log 3 \approx 1.10$. Acc. is measured against majority-vote labels. †Macro F1 computed on items with human prob $\geq 70\%$ ($N = 166$).

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	**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 6: Average cosine similarity

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 7: 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.