Casual as an Anchor: Resolving Supervision Misalignment in Formality Transfer Dataset

Hyojeong Yu*
hyoj.yu@snu.ac.kr
Seoul National University
Seoul, Korea

Hyukhun Koh*
hyukhunkoh-ai@snu.ac.kr
Seoul National University
Seoul, Korea

Minsung Kim
kms0805@snu.ac.kr
Seoul National University
Seoul, Korea

Kyomin Jung†
kjung@snu.ac.kr
Seoul National University
Seoul, Korea

Abstract

Formality transfer is commonly framed as a symmetric bidirectional task between informal and formal registers. We argue that this framing conceals a supervision design flaw in existing benchmarks such as GYAFC: binary human rewrites encode relative stylistic shifts rather than absolute human notions of formality. Consequently, models learn to generate pseudo-formal outputs that satisfy benchmark labels while failing to produce genuinely formal language. We quantify this misalignment by re-evaluating benchmark "formal" labels under a human-aligned definition of formality, revealing substantial discrepancies that propagate to consistent informal→formal failures across model families. To address this issue, we reconceptualize formality transfer as a graded dimension rather than a binary attribute. To operationalize this view, we introduce a three-level spectrum—informal, casual, and formal—where casual serves as an explicit intermediate state that clarifies supervision signals. Based on this framework, we introduce 3LF, a dataset providing parallel supervision across all three levels. Training on 3LF substantially reduces informal→formal failures and improves alignment with human perception. For example, GPT-4.1-nano improves from 0.06 to 0.88 F1 in the informal→formal direction despite 3LF being significantly smaller than GYAFC. We further demonstrate that these gains cannot be reproduced through in-context learning alone and provide qualitative analyses of ambiguity-driven errors and meaning distortions. Overall, our findings demonstrate how supervision design shapes stylistic alignment and highlight the importance of alignment-aware benchmark construction in controllable text generation.

CCS Concepts

• **Computing methodologies** → **Natural language generation**; **Language resources**.

Keywords

formality transfer, dataset construction, theory-grounded supervision

ACM Reference Format:
Hyojeong Yu, Hyukhun Koh, Minsung Kim, and Kyomin Jung. 2026. Casual as an Anchor: Resolving Supervision Misalignment in Formality Transfer Dataset. In *3rd HEAL Workshop at CHI Conference on Human Factors in Computing Systems, Barcelona, Spain.* ACM, New York, NY, USA, 15 pages.

1 Introduction

Text style transfer, particularly formality transfer, is a central task in controllable text generation. It involves rewriting a sentence into a target stylistic register—such as converting informal text into a formal tone—while preserving its original meaning [5, 22, 26, 28]. Traditionally, formality transfer is treated as a binary transformation between informal and formal registers. However, our comprehensive evaluation reveals a persistent directional asymmetry: state-of-the-art models consistently underperform when converting informal text to formal language, while performing relatively well in the reverse direction [19, 21, 29, 32].

We show that this asymmetry arises from a misalignment between benchmark supervision and human perceptions of formality. Existing datasets such as GYAFC [28] rely on binary human rewrites that primarily emphasize surface-level corrections rather than producing genuinely formal expressions characterized by hedging, nominalization, or passive constructions [3, 4, 15]. As a result, this supervision misalignment collapses distinct stylistic intents into a single "formal" label, encouraging models to learn relative shifts instead of aligning to human-defined formality. This simplification obscures the fact that formality naturally forms a graded spectrum-informal, casual, and formal-where casual serves as an intermediate register that is grammatically clean but less rigid than formal text.

To address this misalignment, we introduce 3LF, a dataset explicitly designed to model formality as a three-level spectrum (informal–casual–formal), providing aligned sentence triples across all levels. By making the intermediate casual state explicit, 3LF

*Both authors contributed equally to this research.
†Corresponding Author

HEAL@CHI '26, Barcelona, Spain

offers a more interpretable and alignment-aware supervision signal grounded in clear linguistic criteria.

We evaluate multiple model families (GPT-4.1-nano, Flan-T5-Large, and DeepSeek-Distill-Qwen-1.5B) under controlled training settings, comparing 3LF with traditional binary supervision. Across all models, training on 3LF substantially improves informal→formal performance-a direction where binary-trained models consistently fail. These gains cannot be replicated through in-context learning alone, and are supported by qualitative analyses of meaning distortions and ambiguity-driven errors. These findings suggest that dataset supervision design plays a central role in shaping human-perceived stylistic quality, and therefore deserves greater research attention alongside advances in prompting and model scaling.

In summary, our contributions are threefold:

- We systematically analyze directional asymmetry in formality transfer and show that persistent informal→formal failures stem from structural misalignment in benchmark supervision.
- We introduce a theoretically grounded three-level formality spectrum (informal–casual–formal) and present 3LF, a carefully constructed dataset that provides explicit and interpretable supervision for each formality level.
- We demonstrate that alignment-aware training with 3LF significantly improves informal→formal transfer and yields outputs that better reflect human-defined formality.

2 Related Work

Formality style transfer traditionally involves converting text between formal and informal styles, often using datasets like GYAFC [28]. Recent critiques highlight that these benchmarks focus on superficial modifications, inadequately representing true linguistic formality. Lai et al. [20], Liu et al. [21] emphasize these shortcomings, noting that even advanced models frequently generate outputs with informal elements. Toshevska et al. [32] suggest enhancing models through knowledge graphs to achieve deeper linguistic transformations, though rigorous human evaluation remains needed. Further research [22, 23, 29] argues for more robust evaluations that include intermediate stylistic states, criticizing overly simplistic binary formality definitions. Recent proposals advocate for expert-guided intermediate feedback to improve transformations, especially from informal to formal styles, yet empirical validation is still necessary.

Overall, these studies emphasize the need for richer datasets and nuanced evaluation approaches to address limitations in existing benchmarks like GYAFC.

3 Revisiting Formality in Existing Benchmarks

To analyze the asymmetry in formality transfer, we revisit existing datasets and uncover annotation inconsistencies that introduce misaligned supervision signals and compromise their reliability as evaluation standards. While GYAFC [28] has become the standard benchmark for formality transfer evaluation, our systematic analysis reveals fundamental issues with its formality annotations that may explain the observed directional bias in model performance.

3.1 Linguistic Definition of Formality Spectrum

First, as pointed out by Yang and Carpuat [34], there is a concern about whether GYAFC truly contains formal states. Therefore, to rigorously define the different levels of formality, we refer to the theoretical, decontextualized definition of formal expressions [15]. Informal and formal expressions are determined by the frequency of non-deictic words present in the given sentence. Non-deictic words include nouns, adjectives, prepositions, and articles. More use of non-deictic words results in a more passive, formal tone. In contrast, a higher proportion of deictic words such as pronouns, verbs, adjectives, and interjections, results in a more direct, informal tone. However, it is mentioned that formality lies on a continuum, and that all linguistic expressions will be situated somewhere in between the two extremes. We define this ambiguous in-between style as "casual" tone.

To rigorously define the levels of formality, we draw on the theoretical and decontextualized definition of formal expressions proposed by Heylighen [15]. Based on such definitions, we characterize the sequences along a formality spectrum—Informal, Casual, and Formal—by identifying representative linguistic features such as:

- **Informal:** Characterized by the presence of slang, netspeak, interjections, emojis, non-standard spelling, and grammatical errors. This tone resembles spontaneous conversation in online settings.
  "LOL that was sooo weird. idk what just happened but omg O_O"
- **Casual:** Uses contractions, abbreviations, and direct address (e.g., "you", "hey") but avoids overtly informal elements such as emojis or slang. It is relaxed yet grammatically clean.
  "Hey, I'm not sure what happened, but it was quite weird."
- **Formal:** Employs hedging phrases (e.g., "it appears that", "may suggest"), nominalization, and passive constructions. This tone emphasizes objectivity and detachment.
  "It appears that an unexpected event occurred, the nature of which remains unclear."

To translate these linguistic principles into a consistent annotation framework, we design a rule-based decision tree, illustrated in Figure 1.

3.2 Binary Formality Labels as Relative Preferences: Evidence from GYAFC

Building on our definition of formality as a semantic continuum, we systematically examine the limitations of traditional binary-oriented benchmarks and classifiers. Specifically, we conduct a three-stage evaluation to assess the quality of GYAFC's "formal" labels. First, we apply a traditional formality classifier [9] (See details of classifier construction in Appendix A) to all sequences labeled as "formal" in GYAFC (approximately 100,000 sequences), achieving a classification accuracy of 91.1%. However, this high score stems from classifiers trained on traditional datasets. To perform a more stringent evaluation, we employ theoretically grounded criteria to verify whether the sequences are truly formal. Due to resource limitations, we sample 1,000 examples and re-evaluate them using an LLM-based classifier, specifically GPT-4o (See Appendix Table 5 for details), according to the definition in Section 3.1. The results

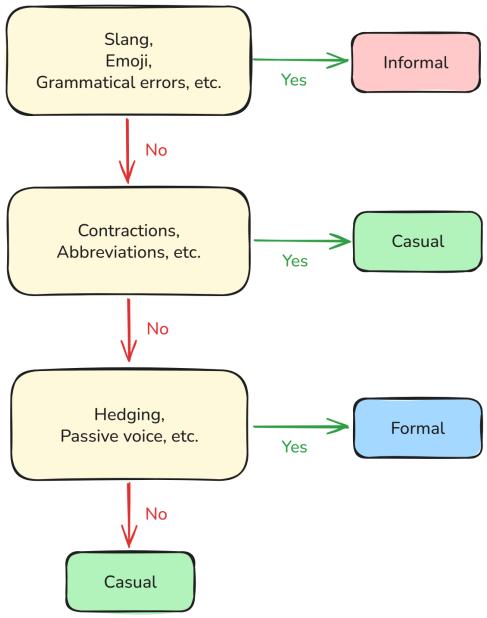

Figure 1: Rule-based decision tree for classifying sentence style.

reveal a striking discrepancy: only 98 sequences (9.8%) meet strict formality criteria, while 902 sequences (90.2%) are reclassified as informal or casual. This massive discrepancy—from 91.1% to 9.8% formal classification—reveals a fundamental flaw in the human annotation design of the benchmark, where labels fail to align with theoretically grounded notions of formality.

A detailed investigation by three human annotators further confirms that the dataset frequently conflates casual or semi-formal text with genuinely formal language, often depending on the domain. The annotation process was reliable, with consistently high inter-annotator agreement (Fleiss' $\kappa > 0.7$ across all tasks), and disagreements were resolved by majority vote. For instance, the sentence *"It is very similar to the age-old question, 'What if my blue is your red?'"* contains the term age-old, which might be considered formal in a general sense, but in professional domains such as OpenReview, it may not qualify as formal. Similarly, *"Blue is my favorite color."* expresses a personal opinion and would be considered informal in the context of an official statement.

3.3 Binary Annotation Induces Minimal-Edit Bias

This discrepancy can be traced to GYAFC's annotation setup, where annotators rewrite informal sentences into formal versions. However, human evaluation of formal references in GYAFC reports an average score of 0.38 on a -3 to 3 scale [28], indicating that these targets occupy a limited region of the stylistic spectrum. Rather than representing genuinely formal language, the rewrites often reflect relative shifts away from informality, operationalizing formality through the removal of surface markers rather than the incorporation of register-specific features. In broader text style transfer research, the lack of standardized evaluation procedures and inconsistent human evaluation practices have been shown to impede reliable style judgments, with automated metrics often poorly aligned with human perception of stylistic quality [24].

This relativistic framing—where formality is defined in relation to the input rather than through a theoretically grounded criterion [34]—undermines the reliability of downstream evaluation and encourages supervision misalignment. Models trained under such benchmarks tend to produce pseudo-formal outputs that satisfy skewed criteria without achieving true formality. These observations motivate the need for an explicit stylistic anchor that defines formality beyond relative rewrites and provides stable reference points across intermediate states—an insight that directly informs the design of our three-level dataset.

4 3LF Dataset Construction

To resolve the conflation of casual or semi-formal text with genuinely formal language, we require a new dataset to deal with the directional asymmetry in performance across formality transfer. In particular, we treat the casual register as a human-interpretable anchor between informal and formal styles, enabling the construction of unambiguous alignment targets. To this end, we rewrite casual sentences into both formal and informal variants, forming rigorously aligned style triples. The following section details our methodology for creating the 3-Level Formality(**3LF**) dataset.

4.1 Design Principle: Casual as a Stylistic Anchor

When constructing a dataset with a larger stylistic gap, direct transformation between informal and formal registers poses a significant challenge, as models often struggle to perform such large stylistic shifts while preserving the original meaning. Prior work has similarly noted that effectively adjusting stylistic attributes without compromising core content remains a central difficulty in the field [17, 22]. Also, informal sentences frequently exhibit syntactic incompleteness, omitting essential components such as subjects or predicates (e.g., "unless it's with the wrong man."), resulting in underspecified semantic content. Formalizing such expressions requires reconstructing implicit structure and supplying missing contextual and structural information, rather than merely adjusting surface style. As a result, informal→formal transfer is not simply the reverse of formal→informal rewriting; the two directions differ in their structural and informational demands. Treating them as symmetric tasks therefore obscures a fundamental asymmetry embedded in the transformation itself. In contrast, the casual register occupies an intermediate position along a graded dimension of formality. We hypothesize that such a casual anchor reduces the requirement for semantic inference over underspecified informal inputs, thereby disentangling content completion from stylistic transformation.

Starting from the casual anchor, a formal variant can be generated by selectively amplifying formal linguistic features—such as hedging, nominalization, and passivization—while suppressing informal elements. Conversely, an informal variant can be obtained by removing formal features and intensifying informal characteristics. This anchored formulation reduces the risk of meaning distortion and enables more reliable construction of stylistically distinct yet semantically aligned sentence variants, effectively decomposing a

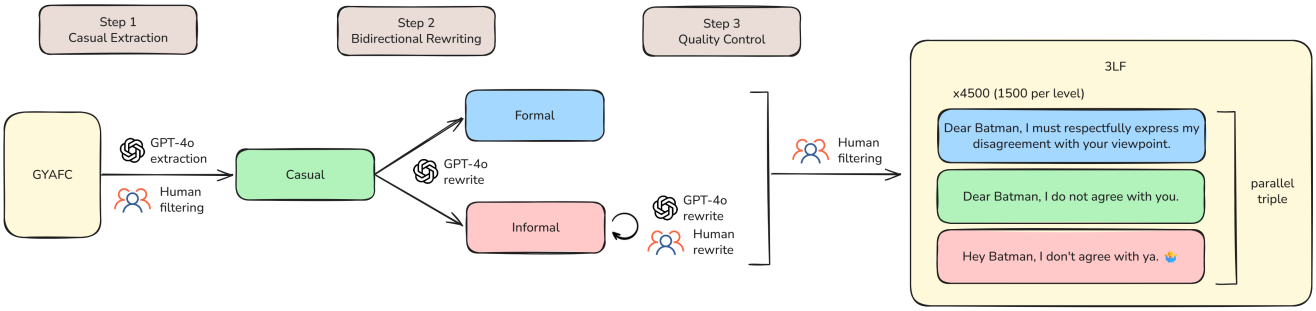

Figure 2: Construction pipeline of 3LF. Casual sentences are first identified from GYAFC using LLM-assisted filtering under fixed linguistic criteria. Each sentence is then rewritten into formal and informal variants within a human-in-the-loop pipeline, where human revision is applied at every stage to ensure alignment consistency and annotation quality.

large and ambiguous shift into two smaller, more tractable transitions.

4.2 Construction Pipeline

We utilize GPT-4o for dataset construction within a human-in-the-loop pipeline designed to ensure annotation quality. Human revision is applied at every stage of LLM-assisted rewriting, and all annotators involved possess advanced proficiency in English. (See Appendix A) Annotators were provided with our formality decision tree illustrated in Figure 1 to guide all revision processes. The entire construction pipeline is shown in Figure 2.

Training Set We use the training split of GYAFC, a standard benchmark for formality style transfer. Based on our formality decision tree, we use GPT-4o as an automatic judge to extract sentences exhibiting casual tone from the original corpus. Our prompt follows the evaluation template introduced in Koh et al. [16]. (See Appendix D.2) Then, we instruct the LLM to rewrite each extracted casual sentence into both formal and informal variants, thereby constructing parallel and explicitly aligned data. Human annotators verify whether each sentence conforms to the categories defined in Figure 1. For the informal rewrites, we apply multiple rounds of targeted revision using curated prompts, along with human verification to ensure sufficiently strong informal signals. In total, we craft 4500 sequences, 1500 samples for formal, casual, informal respectively.

To verify the effectiveness of the 3LF dataset construction process, we also create a NAIVE-3LF set for comparison. We sample 4,500 informal sentences from the GYAFC corpus and rewrite them directly into formal style using GPT-4o, without employing the intermediate casual state construction approach.

Test set For evaluation, we construct a test set spanning two formality levels: informal and formal. The test data are drawn from two sources: GYAFC, derived from Yahoo Answers, and the Pavlick dataset [26], which covers four domains-news, email, answers, and blogs-and is skewed toward more formal language. We sample 200 informal instances from the GYAFC test split and 200 formal instances from the Pavlick test split. Because the Pavlick dataset provides formality scores ranging from -3 to +3, we retain only examples with positive average scores to ensure a high degree of formality. Following human evaluation of the sampled examples, the final test set comprises 400 sentences, evenly distributed with 200 examples. All annotations were conducted following the procedure illustrated in Figure 1.

4.3 Dataset Quality and Integrity

We further assess potential data leakage by measuring lexical overlap between each training dataset (GYAFC, NAIVE, and 3LF) and the test set. Specifically, we report n-gram overlap statistics (from 1-gram to 5-gram) to quantify surface-level similarity between training and test instances.

Dataset	1-gram	2-gram	3-gram	4-gram	5-gram
GYAFC	0.771	0.478	0.207	0.061	0.016
NAIVE	0.380	0.164	0.024	0.003	0.000
3LF	0.401	0.193	0.036	0.005	0.001

Table 1: N-gram Overlap Statistics

Dataset		Characters	Words
GYAFC	Formal	51.34	10.30
	Informal	55.87	10.97
3LF	Formal	80.07	13.79
	Casual	53.19	10.43
	Informal	49.19	9.94

Table 2: Sentence-Level Statistics on GYAFC and 3LF

As shown in Table 1, lexical overlap between the training and test sets is limited at higher n-gram orders: while GYAFC exhibits a relatively high 1-gram overlap (0.771), overlap drops sharply for longer sequences (5-gram = 0.016). Both NAIVE and 3LF demonstrate substantially lower overlap across all n-gram orders, with near-zero overlap beyond trigrams.

These results indicate that the test set is not trivially recoverable from the training data and that performance gains observed in our experiments cannot be attributed to surface-level lexical memorization. Overall, 3LF maintains low lexical redundancy with the evaluation set while preserving sufficient linguistic diversity for robust training.

Additionally, sentence-level statistics (Table 2) show that 3LF exhibits clear stratification across levels, with formal sentences

substantially longer and lexically richer, consistent with established linguistic characterizations of formal registers [4, 15].

5 Experiments

To assess the effectiveness of our 3LF dataset as supervision for formality-aware generation, we conduct controlled experiments on a generative formality transfer task. Specifically, we conduct controlled experiments across diverse model families to examine whether training with 3LF leads to more reliable and well-grounded formality transformations. This section details the experimental setup and evaluation protocol.

5.1 Experimental Setup

For training, we consider three settings: (i) a dataset incorporating the introduced casual state (**3LF**), (ii) a dataset without it (**NAIVE**), and (iii) the baseline dataset (**GYAFC**).

To investigate the impact of dataset quality on generation performance, we fine-tune three models: GPT-4.1-nano, Flan-T5-Large, and DeepSeek-Distill-Qwen-1.5B—across all three datasets and compare their results. We used customized prompts for each model, as shown in Appendix D.3.

Next, we evaluate generation quality using a combination of automatic and human-centered metrics. For formality assessment, we compute precision, recall, and F1 scores for each rewriting direction (informal-to-formal and formal-to-informal), along with overall accuracy across both directions.

To further assess the quality of generated outputs, we measure fluency and meaning preservation, focusing only on sentences that exhibit the correct target formality. Fluency is evaluated automatically using GPT-4o, which assigns an integer score between 0 and 5 based on sentence naturalness and grammar. Our evaluation prompt is based on recent LLM-based style transfer evaluation templates [23–25] with temperature 0. Following Koh et al. [16], we further incorporate an explicit definition-based rubric for formality, ensuring that the evaluator applies theoretically grounded criteria rather than relying on implicit stylistic preferences (See Appendix D.3). For meaning preservation, three annotators independently determine whether the original and rewritten sentences convey the same meaning. Final labels are decided by majority vote. Due to the low fluency of outputs generated by DeepSeek-1.5B, we restrict annotation to generations from GPT-4.1-nano and T5-large, which consistently demonstrate higher fluency.

5.2 Formality Transfer Results

As shown in Table 3, a consistent pattern emerges across all three models: fine-tuning on 3LF yields substantial improvements in formality accuracy, with the most pronounced gains in the informal-to-formal (I→F) direction. For instance, GPT-4.1-nano achieves an I→F F1 score of 0.88 when trained on 3LF, compared to 0.06 on GYAFC and 0.83 on NAIVE—an absolute improvement of +0.82 over the GYAFC baseline. This sharp contrast indicates that the primary weakness of existing benchmarks lies in the insufficient supervision for genuine formal generation. By anchoring stylistic transformations to a shared intermediate register, 3LF introduces a cognitively grounded reference point along the formality continuum, reducing polarity ambiguity and producing outputs that more closely align with human-perceived notions of formality.

Importantly, the improvements are not confined to a single direction. In the formal-to-informal (F→I) setting, GPT-4.1-nano and DeepSeek-1.5B also show measurable F1 gains over the GYAFC baseline. This suggests that the bidirectional alignment induced by 3LF enhances stylistic controllability more broadly, stabilizing both mappings rather than disproportionately benefiting one.

Additionally, all models exhibit increased overall accuracy when trained on 3LF compared to NAIVE, validating the effectiveness of our data generation pipeline. Notably, these improvements also hold when compared against GYAFC, despite the smaller scale of the 3LF dataset. This result underscores that alignment quality and stylistic clarity can outweigh sheer data quantity in training effective style transfer models.

5.3 Generation Quality

We apply two metrics-fluency and meaning preservation-to evaluate generation quality. For fluency, GPT-4.1-nano and T5-large consistently achieve high scores above 3.5 regardless of the training dataset, indicating their robustness to data variation. In contrast, DeepSeek-1.5B exceeds this threshold only when trained on our 3LF dataset, while its fluency score drops below 2.0 when trained on GYAFC. These results suggest that training data quality has a substantial impact on the model's ability to produce fluent outputs.

We further evaluate meaning preservation for GPT-4.1-nano and T5-large. GPT-4.1-nano maintains consistently high scores across all datasets. The model trained on 3LF achieves the highest preservation score, indicating that the anchored alignment structure of our dataset helps the model preserve semantic content while performing non-trivial stylistic transformations.

T5-large achieves an even higher meaning preservation score of 0.8750 on the GYAFC dataset. However, a qualitative inspection of the generated outputs reveals that this high score is largely driven by the model's tendency to copy or minimally modify the input sentence rather than performing substantive rewriting. Consequently, the model exhibits almost no effective formality transfer, with overall accuracy dropping to as low as 0.29. In contrast, when trained on 3LF, the model achieves a substantially higher accuracy of 0.46, while maintaining competitive preservation scores, indicating more meaningful stylistic transformation rather than superficial retention.

Overall, these results indicate that effective formality transfer requires supervision that encourages non-trivial stylistic transformation while preserving semantic content. Datasets that provide explicit and well-aligned stylistic supervision—such as 3LF—enable models to achieve this balance, yielding fluent outputs with faithful meaning preservation and reliable formality control.

5.4 Qualitative Analysis

Coreference / Discourse-level distortion. We further conduct qualitative analysis on generated outputs. For T5-large, we observe discourse-level shifts in deixis and coreference that subtly alter interpretation (e.g., "No way im 5'4 and he's 6'2" → "I am 5'4 and he is 6'2"). For GPT-4.1-nano, there are several types of meaning

Model	Dataset	F→I			I→F			Acc.	Fluency	Meaning Preservation
		P	R	F1	P	R	F1			
GPT-4.1-nano	GYAFC	0.49	0.93	0.64	0.33	0.04	0.06	0.4825	3.5759	0.8100
	NAIVE	0.81	0.88	0.84	0.86	0.80	0.83	0.8350	3.9162	0.8225
	3LF	0.83	0.98	0.90	0.98	0.81	0.88	**0.8950**	**4.2212**	**0.8525**
T5-large	GYAFC	0.37	0.57	0.45	0.02	0.01	0.01	0.2925	4.3333	**0.8750**
	NAIVE	0.40	0.56	0.47	0.27	0.16	0.20	0.3600	4.5069	0.6775
	3LF	0.46	0.36	0.41	0.47	0.56	0.51	**0.4650**	**4.5806**	0.6900
DeepSeek-1.5B	GYAFC	0.53	0.99	0.69	0.92	0.12	0.20	0.5525	1.5385	-
	NAIVE	0.64	0.76	0.70	0.71	0.57	0.63	0.6675	2.1386	-
	3LF	0.72	1.00	0.84	1.00	0.61	0.76	**0.8075**	**3.9164**	-

Table 3: Bidirectional style transfer results. Accuracy is averaged over both directions. F→I: formal→informal; I→F: informal→formal.

distortion. These include entity shifts (e.g., "good god how old are you" → "One may wonder about the age of the individual in question"), numerical inaccuracies, and subjective bias injection. These errors reflect failures to preserve speaker stance and referential structure, highlighting that stylistic rewriting interacts with pragmatic meaning rather than merely surface-level editing.

Underspecification in Informal Language. A second class of errors stems from the inherent underspecification of informal language. Informal expressions frequently omit subjects, contextual grounding, or explicit propositional structure, making their semantic intent ambiguous. When rewriting such inputs into formal style, models must implicitly infer missing information, which introduces aleatoric uncertainty. As a result, the transformation task becomes entangled with content completion rather than purely stylistic modulation. This explains why informal-to-formal rewriting is particularly error-prone and motivates our use of a casual anchor to reduce semantic ambiguity prior to formal transformation. Representative examples for each error cases are provided in Appendix E.

5.5 Comparison with In-Context Learning

Model	Accuracy	F1 (F→I)	F1 (I→F)
Zero-Shot	0.52	0.16	0.66
ICL-GYAFC	0.54	0.67	0.23
ICL-3LF	0.76	0.80	0.69
FT-GYAFC	0.48	0.64	0.06
FT-NAIVE	0.83	0.84	0.83
FT-3LF	0.89	0.90	0.88

Table 4: Comparison Across Supervision Strategies

To verify that the observed improvements stem from 3LF supervision rather than the inherent capability of the base model, we compare fine-tuning (FT) with in-context learning (ICL) on both 3LF and GYAFC. We additionally report zero-shot performance as a baseline. To assess the model's default stylistic prior, we remove any artificial role framing (e.g., "You are not an AI assistant...") and evaluate GPT-4.1-nano under a four-shot prompting setup. The four-shot demonstrations are sampled from each dataset according to an informal-targeted selection criterion.

As shown in Table 4, ICL-3LF exhibits the same failure patterns observed in earlier experiments, including directional asymmetry and meaning distortions in informal-to-formal rewriting. While ICL partially mitigates these issues, fine-tuning on 3LF yields substantially more consistent improvements, reducing informal-to-formal failures and producing outputs that better align with the intended formality register. In contrast, fine-tuning on GYAFC performs even worse than ICL-GYAFC, which itself is comparable to zero-shot performance. We further observe that the zero-shot model struggles more with generating informal sentences than formal ones, a pattern we also encountered during dataset construction. Notably, performance on the I→F direction declines with increased exposure to GYAFC supervision, suggesting that supervision derived from a narrowly defined binary benchmark may bias the model away from its original stylistic prior. These results further support our claim that relative binary supervision may provide insufficient signal for genuine formal generation.

6 Conclusion

This study challenges the long-standing binary perspective of formality transfer and demonstrates that effective style transformation requires modeling formality as a graded dimension. By introducing the 3LF dataset—explicitly incorporating an intermediate casual state—we present a structured framework that addresses the persistent informal→formal collapse observed in prior benchmarks such as GYAFC. Beyond the specific case of formality transfer, our findings carry broader design implications. The way stylistic categories are operationalized during dataset construction fundamentally shapes model behavior and perceived competence. While much work has explored advances in model architecture, prompting strategies, and scale, comparatively less attention has been devoted to examining how supervision encodes linguistic constructs. Our results suggest that dataset design is not a peripheral engineering choice, but a central factor influencing alignment with human-perceived language categories. Accordingly, supervision pipelines should be treated as core research artifacts requiring theoretical grounding and empirical validation. We hope this work contributes to the development of more nuanced, faithful, and controllable text generation systems that better reflect linguistic reality and align more closely with human perception.

Acknowledgments

This work was supported by Institute of Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) [No.RS-2023- 00229780, Development of Artificial Intelligence Technology for Process-focused Evaluation(Student's Learning Diagnosis)]. K. Jung is with ASRI, Seoul National University, Korea.

References

[1] Madhusudhan Aithal and Chenhao Tan. 2021. On Positivity Bias in Negative Reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 294–304. doi:10.18653/v1/2021.acl-short.39

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[3] Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

[4] Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.

[5] Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, Bonjour, Salve! XFORMAL: A Benchmark for Multilingual Formality Style Transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 3199–3216. doi:10.18653/v1/2021.naacl-main.256

[6] Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[8] Jonathan Culpeper. 2011. *Impoliteness: Using Language to Cause Offence*. Cambridge University Press.

[9] Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2023. Detecting Text Formality: A Study of Text Classification Approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, Ruslan Mitkov and Galia Angelova (Eds.). INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 274–284. https://aclanthology.org/2023.ranlp-1.31/

[10] Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024. Text-Tuple-Table: Towards Information Integration in Text-to-Table Generation via Global Tuple Extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9300–9322. doi:10.18653/v1/2024.emnlp-main.523

[11] Pedro Calais Guerra, Wagner Meira, and Claire Cardie. 2014. Sentiment analysis on evolving social streams: how self-report imbalances can help. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (New York, New York, USA) *(WSDM '14)*. Association for Computing Machinery, New York, NY, USA, 443–452. doi:10.1145/2556195.2556261

[12] Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022. Balancing out Bias: Achieving Fairness Through Balanced Training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 11335–11350. doi:10.18653/v1/2022.emnlp-main.779

[13] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing* 40, 1 (2023).

[14] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing* 40, 1 (2023), 75–87. doi:10.1016/j.ijresmar.2022.05.005

[15] Francis Heylighen. 1970. Formality of Language: definition, measurement and behavioral determinants. (02 1970).

[16] Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. 2024. Can LLMs Recognize Toxicity? A Structured Investigation Framework and Toxicity Metric. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 6092–6114. doi:10.18653/v1/2024.findings-emnlp.353

[17] Chaona Kong, Jianyi Liu, Yifan Tang, and Ru Zhang. 2025. Neuron Activation Modulation for Text Style Transfer: Guiding Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 7735–7747. doi:10.18653/v1/2025.findings-acl.403

[18] 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, Article 110 (Nov. 2024), 38 pages. doi:10.1145/3678179

[19] Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-Specific Neurons for Steering LLMs in Text Style Transfer. *ArXiv* abs/2410.00593 (2024). https://api.semanticscholar.org/CorpusID:273023003

[20] Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-Specific Neurons for Steering LLMs in Text Style Transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 13427–13443. doi:10.18653/v1/2024.emnlp-main.745

[21] Pusheng Liu, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024. Step-by-Step: Controlling Arbitrary Style in Text with Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 15285–15295. https://aclanthology.org/2024.lrec-main.1328/

[22] Sourabrata Mukherjee and Ondrej Dušek. 2024. Text Style Transfer: An Introductory Overview. arXiv:2407.14822 [cs.CL] https://arxiv.org/abs/2407.14822

[23] Sourabrata Mukherjee, Atul Kr. Ojha, John Philip McCrae, and Ondrej Dusek. 2025. Evaluating Text Style Transfer Evaluation: Are There Any Reliable Metrics?. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, Abteen Ebrahimi, Samar Haider, Emmy Liu, Sammar Haider, Maria Leonor Pacheco, and Shira Wein (Eds.). Association for Computational Linguistics, Albuquerque, USA, 418–434. doi:10.18653/v1/2025.naacl-srw.41

[24] Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2023. A Call for Standardization and Validation of Text Style Transfer Evaluation. arXiv:2306.00539 [cs.LG] https://arxiv.org/abs/2306.00539

[25] 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*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 21550–21564. doi:10.18653/v1/2025.findings-emnlp.1175

[26] Ellie Pavlick and Joel Tetreault. 2016. An Empirical Analysis of Formality in Online Communication. *Transactions of the Association for Computational Linguistics* 4 (2016), 61–74. doi:10.1162/tacl_a_00083

[27] Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: a case study on the Enron corpus. 86–95.

[28] Sudha Rao and Joel Tetreault. 2018. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality 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)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 129–140. doi:10.18653/v1/N18-1012

[29] Arkadiy Saakyan and Smaranda Muresan. 2024. ICLEF: In-Context Learning with Expert Feedback for Explainable Style Transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 16141–16163. doi:10.18653/v1/2024.acl-long.854

[30] Chen Shani, Dan Jurafsky, Yann LeCun, and Ravid Shwartz-Ziv. 2025. From Tokens to Thoughts: How LLMs and Humans Trade Compression for Meaning. arXiv:2505.17117 [cs.CL] https://arxiv.org/abs/2505.17117

[31] Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward Informal Language Processing: Knowledge of Slang in Large 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)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1683–1701. doi:10.18653/v1/

2024.naacl-long.94

[32] Martina Toshevska, Slobodan Kalajdziski, and Sonja Gievska. 2025. Style Knowledge Graph: Augmenting Text Style Transfer with Knowledge Graphs. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, Genet Asefa Gesese, Harald Sack, Heiko Paulheim, Albert Merono-Penuela, and Lihu Chen (Eds.). International Committee on Computational Linguistics, Abu Dhabi, UAE, 123–135. https://aclanthology.org/2025.genaik-1.13/

[33] Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo A. Huberman. 2005. E-Mail as Spectroscopy: Automated Discovery of Community Structure within Organizations. *The Information Society* 21, 2 (2005), 143–153. arXiv:https://doi.org/10.1080/01972240590925348 doi:10.1080/01972240590925348

[34] Xinchen Yang and Marine Carpuat. 2025. Steering Large Language Models with Register Analysis for Arbitrary Style Transfer. arXiv:2505.00679 [cs.CL] https://arxiv.org/abs/2505.00679

A Evaluator Choices

Model	Formal	Informal	Total
Monolingual	0.952	0.903	0.912
Multilingual	1.000	0.000	0.175
LLM Evaluation	1.000	0.976	0.980

Table 5: Performance comparison across formal and informal domains

Before exploring the inherent issues in traditional benchmark, we begin by evaluating classifiers on a widely used formality dataset, specifically selecting sequences with high annotator agreement (scores exceeding 2.5 or below -2.5) [26], to select an appropriate formality evaluator. We compare existing formality classifiers such as a monolingual classifier [18] and a multilingual classifier [9]. In addition, given recent advancements, LLM-based evaluations have emerged as promising alternatives for tasks requiring semantic understanding and application of nuanced definitions [16]. We therefore also experiment with an LLM-based evaluator (GPT-4o) with temperature 0, using carefully curated prompts grounded in the theoretical, decontextualized definition of formality proposed by Heylighen [15] (See Appendix D.1). As shown in Table 5, the LLM-based evaluation achieves the highest overall performance. Unless otherwise specified, subsequent experiments in this paper adopt the LLM-based formality classifier. In the following human evaluation, all annotations were performed by a group of professional annotators with advanced proficiency in English. The annotators have extensive experience in English-language NLP tasks, including contributions to multiple academic papers, and possess strong linguistic backgrounds. All annotators have demonstrated high English fluency through prior professional and academic work.

B Formality Transfer with Sentiment

Domain	Cases	Positive Modulation	Negative Modulation	Total Rate
General	28	18	10	14.0%
Review	35	18	17	17.5%
Social/Chat	49	12	37	24.5%
Professional	43	7	36	21.5%
Real-time	40	24	16	20.0%

Table 6: Sentiment Modulation across domain-specific generations

B.1 Sentiment Softening Across Domains

Throughout our experiments, we observe that formality transfer often results in a softening of the original sentences' emotional tone.

This phenomenon warrants deeper investigation and discussion. Therefore, we analyze patterns of sentiment shifts in four different domains: review style, casual social media and chat-based, professional communication, and real-time and retrospective response. **Review-style** domains—including Amazon product reviews, Yelp restaurant critiques, and film evaluations—display informality through highly subjective, evaluative, and personalized language, typically accompanied by explicit sentiment polarity [13]. Conversely, **casual social media and chat-based** platforms, such as Reddit discussions, Twitter interactions, and online news commentary, express informality via colloquialisms, contractions, emojis, and paralinguistic markers, reflecting spontaneous and immediate emotional responses [31]. In contrast, **professional communication** contexts—encompassing peer reviews on OpenReview, LinkedIn discourse, formal email correspondence, and enterprise Slack exchanges—commonly balance relational rapport with linguistic formality, characterized by lexical precision, syntactic clarity, and measured affective expression [27, 33]. Finally, **real-time and retrospective response** domains, exemplified by live-stream commentary, event-specific discourse, and reflective social media posts, frequently employ intensified emotional expressions, compressed syntactic structures, and temporal anchoring strategies, typically realized through present-tense narration and emotive discourse markers [10, 11].

B.2 Empirical Analysis of Sentiment Shift

For this analysis, we again use GPT-4o, assigning sentiment labels using the siebert/sentiment-roberta-large-english classifier [14]. We also use domain-specific prompts as in Appendix D.4. As shown in Table 6, notable cases of sentiment softening occur consistently across all domains during formality transformation. The systematic sentiment modulation we observe in Table 6 is not a side effect but a core mechanism explaining both the asymmetry problem and our framework's success.

Formal language typically emphasizes politeness, social deference, and interpersonal distance over emotional expression [6, 26], deliberately avoiding emotionally charged lexical items and overt subjective expressions, resulting in predominantly neutral or mildly positive sentiment patterns [1]. Conversely, informal language frequently employs colloquialisms, contractions, slang, and syntactic looseness—linguistic devices that inherently convey emotional stance and subjective opinions [4, 8], making informal utterances naturally conducive to sentiment-laden content.

Due to their likelihood-based training objectives, LLMs inherently favor frequent co-occurrence observed in such real-world training data [2, 7], and Shani et al. [30] further shows that LLMs emulate abstract cognitive patterns from human language use, leading to subtle but systematic shifts in affective expression during style transfer [12].

While sentiment modulation is therefore a natural linguistic phenomenon, certain practical scenarios require careful consideration to prevent unintended communicative effects. In professional and formal contexts especially, it may be essential for users to preserve specific sentiment orientations despite stylistic transformations, particularly when communicative intent critically depends on maintaining the original emotional stance.

C Detailed Analysis on 3LF

C.1 Sentence-Level Statistics

We report sentence length and word distribution statistics across formality levels in Table 2. In GYAFC, formal sentences average 51.34 characters, compared to 55.87 for informal ones. In 3LF, the distributions are clearly stratified: informal (49.19 chars / 9.94 words),

casual (53.19 / 10.43), and formal (80.07 / 13.79). These results confirm that formal outputs in 3LF are substantially longer and lexically richer than their informal and casual counterparts, aligning with established linguistic observations that formal registers favor nominalization, hedging, and syntactic elaboration. [4, 15]

C.2 Fluency Evaluation

For fluency metrics, we conducted additional experiments using the CoLA classifier (iproskurina/tda-roberta-large-en-cola). We measured the number of sentences classified as acceptable in each dataset: GYAFC yielded 342 out of 400, NAIVE 312 out of 400, and 3LF 341 out of 400. These results suggest that, although most outputs are broadly acceptable, CoLA is not sufficiently sensitive to capture fine-grained distinctions in fluency across datasets. This contrasts with our GPT-4o–based fluency evaluations, which provide clearer separation and align more closely with human judgments.

D Prompt Examples

D.1 Evaluation Prompt

Evaluation Prompt

<SYS>
You are the formality style transfer agent. Here are the definitions of Formal, Casual, Informal sequence respectively.
- Formal sequence : employs hedging phrases (e.g., "it appears that", "may suggest"), nominalization, and passive constructions. This tone emphasizes objectivity and detachment.
- Casual sequence : uses contractions, abbreviations, and direct address (e.g., "you", "hey") but avoids overtly informal elements such as emojis or slang. It is relaxed yet grammatically clean.
- Informal sequence : is characterized by the presence of slang, netspeak, interjections, emojis, non-standard spelling, and grammatical errors. This tone resembles spontaneous conversation in online settings.
Evaluate the Target sequence, and nothing else.
</SYS>
<INST>
Evaluate the Target sequence. If the sequence is formal, answer 1, and nothing else. else the sequence is informal, answer 0, and nothing else.
Target sequence :
The answer is
</INST>

D.2 3LF Prompt

Formality Classification Prompt

You are not an AI assistant. You are a specialized formality classification machine that can only output three integer labels: 0, 1, or 2. Label the following sentence on formality based on the presence of specific linguistic features. Strictly follow the following labeling rules:
Label 0: - Assign Label 0 if the sentence contains **any** of the following: slang, netspeak, interjections, emojis, non-standard spellings, or grammatical errors.
Label 1: - If none of the above features are present, assign Label 1 if the sentence includes **any** of: contractions, abbreviations, or direct address.
Label 2: - If none of the above apply, and the sentence contains **any** of the following: hedging phrases, nominalizations, or passive voice — assign Label 2.
Note: - Label 1 should also be assigned to all the other sentences that do not contain strong stylistic features listed above.
Take a deep breath and think step by step.

3LF Generation Prompt

You are not an AI assistant. You are a specialized formality transfer machine that can only output rewritten sentences. Rewrite the following casual sentence to an informal sentence based on the presence of specific linguistic features:
Informal: - A sentence is informal if it contains **any** of the following: slang, netspeak, interjections, emojis, non-standard spellings, or grammatical errors.
Casual: - A sentence is casual if none of the above features are present, and includes **any** of: contractions, abbreviations, or direct address.
Formal: - A sentence is formal if none of the above apply, and it contains **any** of the following: hedging phrases, nominalizations, or passive voice.
Note: - All the other sentences that do not contain strong stylistic features listed above are also considered casual.
Only answer with the rewritten sentence. Take a deep breath and think step by step.

D.3 Generation Prompt

GPT-4.1-nano

You are not an AI assistant. You are a specialized sentence rewriting model. Rewrite the given sentence to formal, based on the presence of specific linguistic features:
Informal: sentences containing slangs, netspeak, interjection, emoji, non-standard spelling, grammatical errors
Formal: sentences containing hedging phrase, nominalization, passive voice.
Only answer with the rewritten sentence. Take a deep breath and think step by step.

T5-large

Definition:
- Informal: A sentence is informal if it contains slangs, netspeak, interjections, emojis, non-standard spelling, or grammatical errors.
- Formal: A sentence is formal if it contains hedging phrases, nominalizations, or passive voice.
Instruction:
Rewrite the following sentence to formal.
Sentence: {sentence}

Deepseek 1.5B

You are not an AI assistant. You are a specialized sentence rewriting model. Rewrite the given sentence to formal, based on the presence of specific linguistic features:
Informal: sentences containing slangs, netspeak, interjection, emoji, non-standard spelling, grammatical errors
Formal: sentences containing hedging phrase, nominalization, passive voice.
Only answer with the rewritten sentence.
Sentence:

Fluency Evaluation Prompt

You are a specialized fluency scoring machine that can only output integer scores.
Score the following context on fluency on an scale from 0 to 5, where 5 is of the highest fluency.
Only focus on whether the sentence sounds natural and coherent.
Do not penalize an informal sentence for informal style features such as slang or contractions, and do not penalize a formal sentence for formal style features such as hedging phrases or passive voice.
Take a deep breath and think step by step.
Answer only with the integer value.

D.4 Analysis Prompt

System Prompt

<SYS> You are the formality style transfer agent. Here are the definitions of Formal, Casual, Informal sentence respectively.
- Formal sentence : employs hedging phrases (e.g., "it appears that", "may suggest"), nominalization, and passive constructions. This tone emphasizes objectivity and detachment.
- Casual sentence : uses contractions, abbreviations, and direct address (e.g., "you", "hey") but avoids overtly informal elements such as emojis or slang. It is relaxed yet grammatically clean.
- Informal sentence : is characterized by the presence of slang, netspeak, interjections, emojis, non-standard spelling, and grammatical errors. This tone resembles spontaneous conversation in online settings.
Make the formal sentence into informal sentence based on given instruction, and nothing else. **</SYS>**

General Informal Transfer Prompt

Make the formal sentence into informal sentence.
Informal Sentence : SEQUENCE
Formal Sentence :

Review Domain Prompt

Convert the following formal sentence into an informal review-style expression. Transform descriptive language into personal, evaluative expressions with clear positive/negative sentiment. Use review-typical phrases and authentic personal reactions.
Formal sentence: SEQUENCE
Convert to informal review style:

Conversational and Social Commentary Domain Prompt

Convert the following formal sentence into casual social media/chat style language. Use contractions, slang, emoticons, and reactive expressions that show immediate emotional response. Include satirical or meme-like elements where appropriate.
Formal sentence: SEQUENCE
Convert to social media/chat style:

Professional Communication Domain Prompt

Convert the following formal sentence into professional but approachable informal language. Maintain respect while reducing formality, use warm relationship-building language, and keep sentiment constructive and collaborative.
Formal sentence: SEQUENCE
Convert to professional informal style:

Real-time & Retrospective Expression

Convert the following formal sentence into immediate, real-time reaction language. Use strong emotional expressions, short punchy sentences, reactive words, and present-tense immediacy.
Formal sentence: SEQUENCE
Convert to real-time reaction style:

E Error Case Examples

E.1 GPT-4.1-nano

#	Category	Example
1	Character Shift	consensual → consantal
2	Entity Shift	good god how old are you → One may wonder about the age of the individual in question.
3	Number Shift	0.50 → by 0.52%
4	Addition (Entity-based)	"I Have Nothing" by Jennifer Hudson → The song "I Have Nothing" by Jennifer Hudson is performed in the context of a musical expression. For me it's definitely Jessica Alba and Angelina Jolie... → Jessica Alba and Angelina Jolie are the two actresses I consider to be the most appealing.
5	Bias Injection	In May, in his role as peace envoy, Blair met the education minister of the United Arab Emirates. → In May, shady Blair, just messing around as a peace envoy, met up with the UAE's education minister.
6	Unresolved Aleatoric Uncertainty	id have to say... Will Ferrell. → In my opinion, the actor Will Ferrell would be most representative. or, what about blue and green? → Alternatively, what might be considered utilizing the colors blue and green?

Table 7: Examples of typical generation errors of GPT-4.1-nano categorized by type.

E.2 T5-large

#	Category	Example
1	Deletion	have an equal opportunity to bid for capacity → bid for capacity
2	Key Shift	will close in September 2011 → in the near future No way im 5‘4 and he‘s 6‘2 → I am 5’4 and he is 6’2 The sources I have listed below have all the information → i have listed all the sources
3	Truncation	"Money that went to the armed forces that could have been or should have been spent on health and education, social services, was basically squandered. In any case the time is right now for democracy, for the people of of Guinea to get the elections they were hoping for," he added. → Money that went to the armed forces that could have been or should have been spent on health and education, social services, was basically squandered. In any case, the time.
4	Copy & Paste	It was trying so hard to be the next great American horror film. → It was trying so hard to be the next great American horror film.

Table 8: Examples of typical generation errors of T5 categorized by type.