ICLEF: In-Context Learning with Expert Feedback for Explainable Style Transfer

Arkadiy Saakyan[1] and **Smaranda Muresan**[1,2]
[1] Department of Computer Science, Columbia University
[2] Data Science Institute, Columbia University
a.saakyan@cs.columbia.edu, smara@columbia.edu

Abstract

While state-of-the-art large language models (LLMs) can excel at adapting text from one style to another, current work does not address the **explainability** of style transfer models. Recent work has explored generating textual explanations from larger teacher models and distilling them into smaller student models. One challenge with such approach is that LLM outputs may contain errors that require expertise to correct, but gathering and incorporating expert feedback is difficult due to cost and availability. To address this challenge, we propose ICLEF, a novel human-AI collaboration approach to model distillation that incorporates **scarce** expert human feedback by combining *in-context learning* and *model self-critique*. We show that our method leads to generation of high-quality synthetic explainable style transfer datasets for formality (E-GYAFC) and subjective bias (E-WNC). Via automatic and human evaluation, we show that specialized student models fine-tuned on our datasets outperform generalist teacher models on the explainable style transfer task in one-shot settings, and perform competitively compared to few-shot teacher models, highlighting the quality of the data and the role of expert feedback. In an extrinsic task of authorship attribution, we show that explanations generated by smaller models fine-tuned on E-GYAFC are more predictive of authorship than explanations generated by few-shot teacher models.

1 Introduction

Attribute style transfer is the task of transforming a given text along a particular style dimension, such as changing its formality, bias, or level of offensiveness (Lample et al., 2019; Sudhakar et al., 2019; Jin et al., 2022). Formality style transfer (e.g., informal→formal) could be useful in any writing assistance system, while neutralizing text that contains subjective bias would be an important tool for Wikipedia editors (Pryzant et al., 2020) or journalists (Rosenberg and Fischer, 2023).

Style transfer approaches have primarily focused on the text re-writing task (e.g., informal →formal, subjective bias → neutral) using various methods from supervised (Rao and Tetreault, 2018; Pryzant et al., 2020; Zhong et al., 2021) to unsupervised (Krishna et al., 2020) and zero-shot methods using LLMs (Reif et al., 2022) (see also Jin et al. (2022) for a survey on style transfer). However, to our knowledge, no effort has focused on providing *textual explanations* for the style transfer task. For example, when transforming an informal sentence "I would throw them out asap !" into a formal paraphrase "I would dispose of them promptly" it would be useful to provide an explanation of the informal attributes in the input sentence (e.g., textese ("asap"), colloquialism ("throw out")), and formal attributes for the paraphrase (e.g., lexical sophistication ("promptly" and "dispose"); lack of abbreviations ("I would")). Similarly, for neutralizing subjective bias in "Orbis latinus, integral site of romance language" → "Orbis latinus, comprehensive site of romance language", it would be useful to have an explanation about which word/phrase in the input is biased and why as well as the type of bias (e.g., Framing ("integral" implies a subjective evaluation on the site's importance)). The model's explanations could help the user better assess the correctness of the style transfer system, could be used as features in downstream tasks such as authorship attribution (Section 5), or could act as a defense against spurious correlations (Ludan et al., 2023; Camburu et al., 2018) and annotation artifacts (McCoy et al., 2019; Poliak et al., 2018). To enable explainability in style transfer models, we provide the following contributions:

- A new task of *explainable* style transfer for which, in addition to sentence rewriting, the model needs to generate textual explanations.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16141–16163

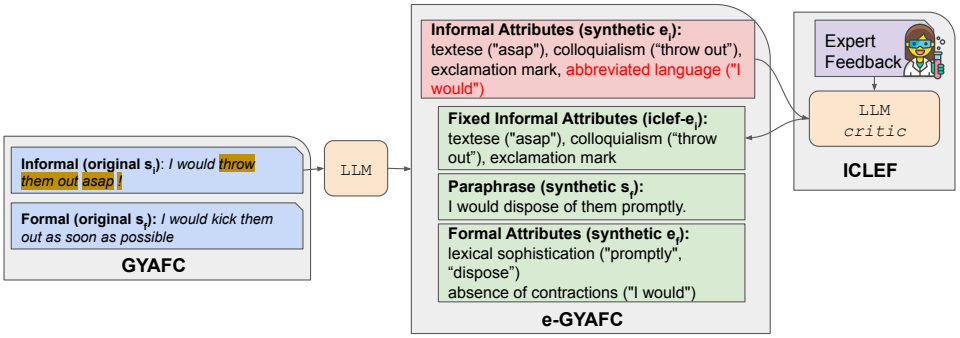

Figure 1: Generating E-GYAFC: formality style transfer dataset GYAFC (Rao and Tetreault, 2018) is augmented with semi-structured natural language explanations. The LLM generates the informal attributes of the input sentence, a formal paraphrase, and the formal attributes of the resulting sentence. Expert feedback is incorporated via in-context learning and self-critique to refine the initial generations.

- A novel human-AI collaboration framework, *In Context-Learning with Expert Feedback (ICLEF)* (see Figure 1, Figure 2, and §3.2). The approach combines model distillation for explanation generation (Ho et al., 2022; Magister et al., 2023) with self-critique ability of LLMs (Madaan et al., 2023; Bai et al., 2022b; Saunders et al., 2022; Scheurer et al., 2023, *inter alia*), where the critic, unlike in prior work, is instantiated with expert demonstrations.

- Using ICLEF, we create for the first time datasets for explainable style transfer by augmenting an existing formality style transfer dataset GYAFC (Rao and Tetreault, 2018) and the neutralizing subjective bias dataset WNC (Pryzant et al., 2020) with textual explanations (§3). We show that the datasets generated with the help of ICLEF, E-GYAFC and E-WNC, are of good quality via automatic and expert evaluation, and that ICLEF-fixed instances are preferred (§3.3).

- Experiments that show that student models outperform teacher models in one-shot setting and perform comparably even with few-shot teacher models in automatic and expert evaluation, confirming the utility and quality of the datasets (§4). Moreover, in an extrinsic evaluation, we show that explanations generated by student models fine-tuned on our data produce a better signal for the authorship attribution task than the explanations produced by few-shot teacher models (§5).

We release the data, models, and code to encourage further research on explainability, learning from scarce human feedback, and style transfer. [1]

2 Related Work

Knowledge distillation and human feedback Model or knowledge distillation is a process of fine-tuning a smaller student model to imitate the behaviour of a more competent teacher model (Beyer et al., 2022; Buciluundefined et al., 2006; Hinton et al., 2015). Knowledge distillation became a popular technique, allowing to generate datasets of similar quality to the crowd-sourced data (West et al., 2022), especially when combined with a model-in-the-loop approach (Wiegreffe et al., 2022; Bartolo et al., 2022; Chakrabarty et al., 2022). Recent work explores model distillation with natural language explanations (Wang et al., 2023a; Ho et al., 2022; Magister et al., 2023), showing that large language models are capable of generating acceptable enough reasoning steps for student models to learn. Approaches to incorporate human feedback such as RLHF (Stiennon et al., 2020) and DPO (Rafailov et al., 2023) require large amounts of crowdsourced data and have not been generally shown to be effective for expert preferences. Imitation learning from human feedback (ILF) (Scheurer et al., 2023) utilizes human feedback to improve model-generated instances, and then fine-tunes on that data. Unlike these works, we focus on incorporating *expert* feedback which is naturally scarce and expensive to collect. Unlike other self-critique approaches (Madaan et al., 2023; Bai et al., 2022b; Saunders et al., 2022), we condition the model on expert corrections to incorporate high-quality human feedback.

[1] github.com/asaakyan/explain-st

Textual Explanations Natural language explanations have been utilized for a variety of tasks Wiegreffe and Marasovic (2021), such as natural language inference (Camburu et al., 2018), commonsense (Rajani et al., 2019; Aggarwal et al., 2021), social norm entailment (CH-Wang et al., 2023). We focus on creating natural language explanations for the style transfer task, which has not been addressed before.

Style transfer Style transfer approaches range from instruction-based methods (Reif et al., 2022) and paraphrasing (Krishna et al., 2020), to approaches focused on learning in low-resource settings (Patel et al., 2022). Much of style transfer work focuses on style representations that decouple style and content (Wegmann et al., 2022; Wegmann and Nguyen, 2021), however most of these methods are not designed to be interpretable. Interpretable approaches rely on constructing interpretable embeddings, such as LIWC (Tausczik and Pennebaker, 2010) or LISA (Patel et al., 2023). Zhong et al. (2021) proposed identifying biased segments in conjunction with neutralizing biased text. Unlike these approaches, we propose to use natural language explanations to further enhance model interpretability.

3 Building Datasets for Explainable Style Transfer

We build two explainable style transfer datasets by first augmenting existing datasets with synthetic textual explanations generated by a teacher model (§3.1), and then improving the generated data using our In-Context Learning with Expert Feedback (ICLEF) framework (§3.2).

3.1 Augmenting Style Transfer Datasets with Synthetic Textual Explanations

Formality style transfer The GYAFC (Rao and Tetreault, 2018) formality style transfer dataset contains parallel formal and informal sentences. The informal sentences are collected from Yahoo Answers, and formal paraphrases were crowdsourced using Amazon Mechanical Turk (AMT). We use ChatGPT-3.5 to generate explanations and formulate the following multi-step generation task: given an informal sentence from GYAFC s_i, generate a structured explanation of its informal attributes e_i, then generate a formal paraphrase s_f based on these attributes, then the formal attributes of the resulting paraphrase e_f. Generating both e_i and e_f allows us to train models in both directions (formal $\rightarrow$ informal and informal $\rightarrow$ formal). We use a semi-structured format for the explanations. Specifically, we ask the model to generate a list of attributes followed by an excerpt from the sentence as the evidence: attribute ("evidence"), see examples in Figure 1. These explanations have a consistent format, making it easier to verify and automatically evaluate.

Subjective bias style transfer We focus on the task of neutralizing subjective biased language introduced by Pryzant et al. (2020) to make sentences follow the Wikipeida Neutral Point of View Policy.[2] We start with the Wikipedia Neutrality Corpus (WNC) (Pryzant et al., 2020), a parallel corpus of 180,000 sentence pairs originating from Wikipedia edits of subjective biased language. The goal is to generate an explanation (e_b) for the type of bias present in the biased sentence (s_b), following the scheme proposed by Pryzant et al. (2020) and Recasens et al. (2013): Framing, Epistemological, and Demographic (see definitions in Appendix I). It is estimated by Pryzant et al. (2020) that a small percentage of cases, the instances in the WNC contain noise. We add an additional "No Bias" label for these cases to reduce hallucinated biases for neutral sentences. In this case, we ask the model to output "This sentence does not contain bias" as the explanation (see second WNC example in Table 1). The explanation is structured as Type of Bias ("evidence" reasoning). Then, the teacher model generates an unbiased paraphrase (s_n). See Figure 2 for an overview. At the time we developed this dataset, ChatGPT-4 became available, so we use this more powerful model as a teacher, especially since generating explanations for this task might require more reasoning capabilities. We do not generate explanations for the neutrality of the paraphrase as we are not exploring the neutral to biased paraphrase direction due to ethical concerns.

3.2 In-Context Learning from Expert Feedback (ICLEF)

ChatGPT generations might contain errors (e.g., the generated style attribute "abbreviated language" with the evidence "I would" in Figure 1). To improve the quality of the data, we turn to expert feedback, since previous work has identified that crowdworkers on platforms such as Amazon Mechanical Turk could be unreliable for open-ended generation

[2] Wikipedia.org

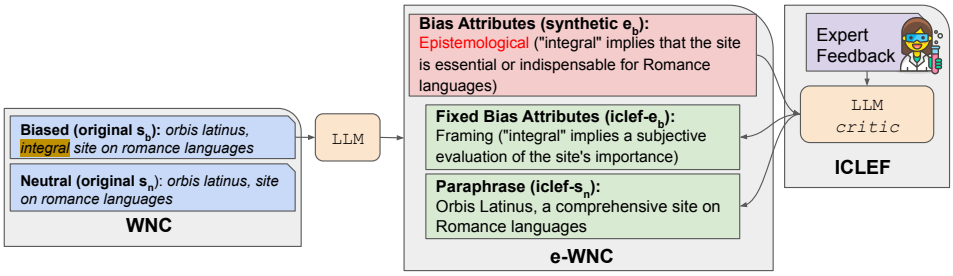

Figure 2: Generating E-WNC: WNC (Pryzant et al., 2020) is augmented with natural language explanations. The LLM generates the bias attributes of the input sentence and an unbiased paraphrase. Expert feedback is incorporated via in-context learning and self-critique to refine the initial generations.

tasks (Karpinska et al., 2021), and might even rely on ChatGPT to provide their answers (Veselovsky et al., 2023). The crux of our approach is to combine in-context learning and self-critique abilities of LLMs by instantiating the LLM-critic model with few-shot expert feedback demonstrations.

E-GYAFC For the formality style transfer task, we hire an expert annotator with a Masters degree in linguistics on Upwork[3]. Our annotation protocol (see Appendix J) provides a non-exhaustive reference to formality and informality attributes and asks the annotator to provide feedback on which attributes in e_i, e_f are incorrect if any, among other information. We provide 50 random samples for annotation. The annotation process took each expert only 2-3 hours. We find that the rate of critical errors observed in formality explanations is significantly lower (($\approx 8\%$ for formality explanations vs $\approx 56\%$ for informality explanations), so we only focus on applying the LLM-critic to incorrect informality attributes (e_i). To do so, we instantiate an LLM-Critic model by prompting another LLM (ChatGPT-3.5) with 35 expert human feedback corrections in-context (we find that this number leads to satisfactory correctness of $\approx 87\%$, see Appendix C for how performance changes given the amount of feedback) and ask it to act as an annotator on the new instances to identify incorrect attributes in them (see prompt in Table 12 in Appendix E). To mitigate the risk of generating new incorrect attributes, we only query the model to identify and remove incorrect attributes in e_i, and if there are any, provide a fixed *iclef*-e_i where they are removed. We refer to the resulting model as LLM-Critic (see how *abbreviated langauge* is removed in Figure 1).

In this way, we fix $\approx 30\%$ of the generated data (2853 instances) – all instances for which the critic has predicted that an improvement is needed. The resulting data (which we refer to as E-GYAFC) contains 9,960 original s_i, synthetic s_f instances with corresponding *iclef*-e_i, synthetic e_f attribute explanations. We randomly split the data into 8,000 training instances and 1,960 held-out test instances.

E-WNC For the neutralizing subjective bias task, we hire an expert annotator with a PhD in linguistics on Upwork. We provide 50 random samples ensuring equal representation for each type of bias. To counteract potential annotation biases, we decided to use two annotators, one of the authors acting as the additional annotator, and then randomly sample the instances in equal proportions. The annotation protocol asks to provide a corrected explanation instead of e_b if the synthetic explanation contains an incorrect type of bias or wrong justification. We randomly sample 35 distinct instances of the two annotators' feedback, finding that this number leads to satisfactory correctness of $\approx 93\%$ (see Appendix C). We provide the expert critiques in-context in a similar way to E-GYAFC (see bottom prompt in Appendix E, Table 12). Since new bias attributes might have been introduced, we regenerate the paraphrase s_n given the new explanation (see corrected type of bias and a new paraphrase in Figure 2). We fix 8% of synthetic explanations in this manner (the lower rate of errors is explained by the higher quality of initial generations due to the use of a more powerful ChatGPT-4 model). The resulting dataset (E-WNC) contains 3,000 original WNC s_b biased sentences, *iclef*-e_b bias explanations, as well as the corresponding *iclef*-s_n neutralized sentences. We randomly split the data into 2,500 training instances and 500 held-out test instances.

[3]Upwork.com

Informal (s_i)	**Gen. expl.** (synthetic e_i)	**ICLEF expl.** (*iclef*-e_i)
hopefully you aren't too old or you are screwed.	informal greeting ("hopefully"), slang ("screwed"), contraction ("aren't")	slang ("screwed"), contraction ("aren't")
more info, we are both in our very late twenties.	[...], omission of prepositions ("in our very late twenties")	abbreviation ("info"), colloquialism ("very late twenties")
Biased (s_b)	**Gen. expl.** (synthetic e_b)	**ICLEF expl.** (*iclef*-e_b)
[...] a play on the title of the popular mtv series, "unplugged".	Epistemological ("popular" implies that the MTV series is universally well-liked)	Framing ("popular" is a subjective term that implies the MTV series is widely liked)
[..] kendal, cbe (born 25 september 1946) is an english actress known in the united kingdom [...].	Demographic ("actress" implies that the person is female and could perpetuate gender stereotypes or assumptions)	This sentence does not contain bias.
claims for the existence of paranormal psychic abilities such as clairvoyance are highly controversial.	This sentence does not contain bias.	Epistemological ("highly controversial" implies that the existence of paranormal psychic abilities is widely disputed)

Table 1: Qualitative comparison of dataset instances before and after application of ICLEF.

3.3 Dataset quality

Automatic evaluation of paraphrase quality We estimate paraphrase quality automatically using Mutual Implication Score (MIS) (Babakov et al., 2022) and Formality Score (see §4.2 for metrics details) between our formal paraphrases and the ones in GYAFC. We find that our paraphrases are of comparable quality with an MIS of 81.30 vs. 83.08 for GYAFC, yet we achieve a higher formality score of 98.43 vs. 89.39 for GYAFC (see Table 2). For example, for the GYAFC example in Figure 1, the formal paraphrase contains *kick them out*).

Similarly, for the E-WNC dataset we report bias score from an off-the-shelf classifier (see §4.2) along with MIS. The MIS scores are 79.32 for original paraphrases vs. 85.58 for our paraphrases, indicating higher semantic similarity. The neutrality scores are 69.34 vs. 72.64, indicating higher neutrality of our paraphrases (see Table 2).

	e-GYAFC		**e-WNC**	
	MIS	**Formality**	**MIS**	**Neutrality**
Orig. para.	**83.08**	89.39	79.32	69.34
Cand. para.	81.30	**98.43**	**85.58**	**72.64**

Table 2: Synthetic paraphrases (generated via model distillation for E-GYAFC and E-WNC) exhibit higher quality overall in automatic evaluation compared to original paraphrases (from GYAFC and WNC, respectively).

Human evaluation For E-GYAFC we hire 3 expert annotators, 2 of which performed the annotation, as well as an independent expert annotator with a masters degree in linguistics. We ask their preferences on 100 randomly sampled instances with respect to the explanations (synthetic e_i vs. *iclef*-e_i) and paraphrases (original s_f in GYAFC vs. synthetic s_f in E-GYAFC). In addition, we ask for acceptability judgments (whether the paraphrase or explanation are correct and complete) for the preferred paraphrase and separately for e_f. We report preference or equal rates and acceptability rates.[4] Overall, we found that our dataset instances are considered acceptable, with the average acceptability rate for e_i, s_f, e_f being 87%, 77%, 98% respectively (row 1 in Table 3). Synthetic paraphrases are generally preferred to the ones in the GYAFC corpus (average 77%), and *iclef*-explanations are preferred or are equal in quality with original generations on average in 90% of cases (row 2 in Table 3). We computed the pairwise accuracy between annotator responses for all categories of E-GYAFC evaluation, and found that it averages at 81% across all categories. We provide more details in Appendix A. Table 1 shows qualitative examples of successful edits with ICLEF. Figure 3 shows the top 10 most frequent informality attributes.

	e-GYAFC			**e-WNC**	
	$\mathbf{e}_i$	$\mathbf{s}_f$	$\mathbf{e}_f$	$\mathbf{e}_b$	$\mathbf{s}_n$
Acceptability	87%	77%	98%	73%	74%
Preference	90%	77%	-	78%	77%

Table 3: Acceptability and Preference Rates (between synthetic explanation vs. *iclef* explanation, and synthetic paraphrase vs. original paraphrase form the dataset) for E-GYAFC and E-WNC.

For E-WNC, we hire 2 annotators, one of whom performed the ICLEF annotation. To meaningfully evaluate the preference of *iclef*-explanations compared to the synthetic ones, we ask to provide feed-

[4]We compute preference or equal preference among acceptable instances. For acceptability, we compute dispreferred instances as unacceptable.

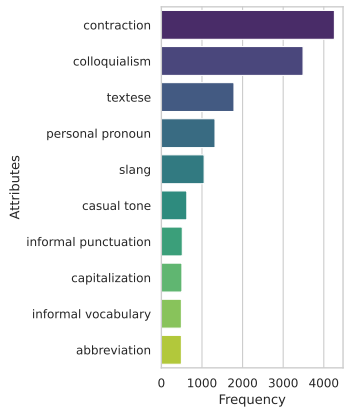

Figure 3: Top 10 informal attributes. See top 50 (in)formality attributes in Appendix Figure 5, 6).

back on 50 instances where the explanation has been updated. The average preference for *iclef*-e_b compared to synthetic e_b is 78%. Average acceptability rate of *iclef*-e_b is 73%. The average preference for synthetic s_n compared to the neutral sentence in the E-WNC corpus is 76%. Average acceptability rate of synthetic s_n is 74%. The pairwise accuracy between the annotators is 77%. More details can be found in Appendix B.

4 Evaluation of Student and Teacher Models on the Explainable Style Transfer Task

We focus on evaluating the performance of select large language models on the *explainable* style transfer task. We do not evaluate post-hoc rationale systems (generating attributes given the paraphrase pair), since such pipeline models are less likely to reflect the underlying reasons for the model prediction, while models that jointly predict and rationalize exhibit desirable properties for faithful explanations (Wiegreffe et al., 2021). For explainable formality style transfer, we test the generation of e_f, s_i, e_i given s_f (Formal→Informal) and e_i, s_f, e_f given s_i (Informal→Formal) on a held-out test set from E-GYAFC. We evaluate how closely the model generated e_i, e_f match E-GYAFC explanations, and we evaluate the semantic closeness and paraphrase quality for s_i, s_f with reference-free metrics. Similarly, we evaluate e_b, s_n for the neutralizing subjective bias task using a test set from E-WNC. We report F1 scores for bias classification in s_n.

4.1 Models

We test two smaller student models fine-tuned on our datasets. We fine-tune LLaMA-7B (Touvron et al., 2023) and Alpaca-7B (Taori et al., 2023) models on our data converted to the Alpaca instruction format. For formality style transfer, we fine-tune in both Formal→Informal and Informal→Formal directions separately (→), as well as in both directions in a multi-task fashion (↔). For subjective bias transfer, we only fine-tune in one direction. In addition, we test the teacher models (ChatGPT-3.5=GPT-3.5 and ChatGPT-4=GPT-4) in few-show setting, which is an ambitious baseline: first, they were used for data generation which biases reference-based metrics, second, they are prompted with improved instances of the data. Since the teacher models are closed models, we also test a representative open instruction-tuned model larger than the student, Vicuna-13B (Chiang et al., 2023) (Vic in tables), in few-shot setting. See detailed description of the models, hyperparameters, prompts, and additional experiments in Appendix F.

4.2 Automatic Evaluation

We use the following metrics:

- BLEU (Papineni et al., 2002): We measure the amount of exactly matched formal and informal attributes and evidences between the generated structured explanation and reference explanation in E-GYAFC and E-WNC.
- Mutual Implication Score (MIS) (Babakov et al., 2022) is a symmetric measure of text semantic similarity based on a RoBERTa (Liu et al., 2019) model fine-tuned for natural language inference and paraphrase detection used in prior work (e.g., Patel et al., 2022).
- Style Accuracy: For E-GYAFC, we use Formality/Informality Score[5]: RoBERTa (Liu et al., 2019) fine-tuned to predict whether sentences are formal or informal using GYAFC and Online Formality Corpus (OFC) (Pavlick and Tetreault, 2016). It achieves up to 0.98 ROC AUC. For E-WNC, we use Bias Score[6]: DistilBERT model (Sanh et al., 2020) fine-tuned for bias classification on the BABE media bias dataset annotated by experts (Spinde

[5] huggingface.co/s-nlp/roberta-base-formality-ranker
[6] huggingface.co/social-media-fairness/classifier-bias-sg

Model	Size	Formal → Informal				Informal → Formal				**Average**
		Form.Attrs. BLEU	MIS	Informality	Inform.Attrs. BLEU	Inform.Attrs. BLEU	MIS	Formality	Form.Attrs. BLEU	
Vic_1	13B	23.16	83.24	35.26	10.97	27.31	61.22	**98.70**	9.88	43.72
Vic_5	13B	24.16	85.18	33.58	13.09	27.95	73.18	98.20	15.45	46.35
Vic_{10}	13B	19.78	82.00	49.17	12.35	30.97	73.26	97.86	17.14	47.82
GPT-3.5_1	?	28.88	90.12	39.65	12.80	27.65	81.98	97.68	10.36	48.64
GPT-3.5_5	?	33.98	**90.37**	38.23	16.14	36.03	85.37	97.54	16.30	51.74
GPT-3.5_{10}	?	36.78	89.57	49.81	18.51	37.36	**85.62**	97.75	20.31	54.46
$LLaMA_{\rightarrow}$	7B	39.64	85.31	61.33	19.86	38.02	81.80	97.77	25.10	56.10
$Alpaca_{\leftrightarrow}$	7B	**40.42**	81.76	**66.71**	**21.11**	**40.34**	79.43	98.57	**25.75**	**56.76**

Table 4: Performance of instruction-tuned and fine-tuned models on the explainable formality style transfer task. Best bolded, best non-fine-tuned underlined.

et al., 2021), on which it obtaines F1 score of up to 79.

Given that the bias type detection task can be viewed as a classification task (only 3 labels are present), we report F1 scores for bias type classification in the explanation. We also report the average across all metrics.

Results Table 4 shows model performance on the explainable formality style transfer task. While the Vicuna model does well in terms of style transfer (as evidenced by high MIS and Formality scores), it lacks in explanation quality (overall low BLEU scores). Student models perform better than the one-shot teacher model and competitively in 10-shot scenario judging by the Average score. Table 5 shows model performance on the explainable subjective bias style transfer task. Similarly, student models outperform the teacher model in one-shot setting. They also outperform few-shot models weaker than the teacher model (Vicuna and GPT-3.5).

As for style transfer performance without considering the explanations, we see a slight decrease in the Informal→Formal (-0.93% avg. MIS and Formality compared to one-shot teacher) and Bias→Unbias (-2.13% avg. MIS and Neutrality) tasks. This is expected as consistent with prior work such as e-SNLI (Camburu et al., 2018).[7] We see a sizeable increase in performance for Formal→Informal (+12.60%) direction. The informality scores are much better for the student models, perhaps due to the tendency to generate more formal speech both by teacher models and the other instruct models.

[7]"...while sacrificing a bit of performance, we get a better trust that when EXPLAINTHENPREDICT predicts a correct label, it does so for the right reasons."

Model	F1	Attrs. BLEU	MIS	Neutr. score	**Avg**
Vic_1	17.49	2.88	74.61	68.33	48.61
Vic_5	16.28	9.58	81.22	73.12	54.64
Vic_{10}	23.08	9.35	84.66	74.90	56.30
GPT-3.5_1	34.83	8.15	83.18	75.07	55.47
GPT-3.5_5	28.78	13.94	81.71	73.11	56.25
GPT-3.5_{10}	37.83	17.25	82.92	73.18	57.78
GPT-4_1	36.87	12.33	82.38	**75.72**	56.81
GPT-4_5	41.24	14.91	82.36	75.48	57.58
GPT-4_{10}	39.82	17.07	**83.48**	74.87	58.47
$Alpaca_{\rightarrow}$	65.03	24.46	82.57	71.63	59.55
$LLaMA_{\rightarrow}$	**67.25**	**25.81**	**83.48**	71.33	**60.21**

Table 5: Performance on Biased→Unbiased explainable style transfer. GPT-3.5 = ChatGPT-3.5, GPT-4 = ChatGPT-4, Vic = Vicuna-13B. Number of shots in underscripts. Best score bolded.

4.3 Expert Human Evaluation: Preference Judgments

We evaluate how well models' outputs align with expert preferences. For `e-GYAFC`, we sample 100 instances from the test set and the corresponding model outputs from Vic_1, the teacher model GPT-3.5 in one-shot setting (GPT-3.5_1), and the best student fine-tuned model. We ask for their preferences in terms of correctness and completeness of explanations, as well as in terms of paraphrase semantic preservation. We find that student model ($Alpaca_{\leftrightarrow}$) generations are preferred to teacher model GPT-3.5_1 and $Vicuna_1$ 53% of the time by the expert linguist (GPT-3.5_1 is preferred 42%), indicating the student model is more aligned with expert preferences. See a qualitative example in Table 6. We also evaluate how well do smaller fine-tuned models perform compared to the teacher model in a few-shot setting. We sample 40 outputs and ask for preferences between generations from Vicuna and GPT-3.5 in the 10-shot setting (GPT-

3.5_{10}) as well as the student fine-tuned model. The fine-tuned model performs competitively with 49% preference for GPT-3.5_{10} and 43% preference for the small student fine-tuned model.

Similarly, we hire an expert annotator and provide 50 test set instances from E-WNC along with model outputs from GPT-3.5_{10} (weaker than the teacher model), 10-shot Vicuna, and the best student fine-tuned model (LLaMA$_{\rightarrow}$). We find that LLaMA$_{\rightarrow}$ is preferred at a rate of 56%, followed by GPT-3.5_{10} (24%). See qualitative examples in Table 7. We also measure preferences to one-shot teacher model (GPT-4_1), finding comparable preference of 42% for LLaMA$_{\rightarrow}$ and 40% for GPT-4_1.

Formal (Input)
If it is feasible, allow love to prevail.
Vicuna$_1$
`-Formal Attributes:` passive voice ("it is feasible", "allow love to prevail"), use of verb "to be" ("is feasible", "prevail") `-Informal Paraphrase:` It's okay to let love rule. `-Informal Attributes:` active voice ("it's okay", "let love rule"), informal verb use ("rule")
GPT$_1$
`-Formal Attributes:` conditional clause ("if it is feasible"), elevated vocabulary ("allow love to prevail") `-Informal Paraphrase:` If it's possible, just go with love. `-Informal Attributes:` simple language ("if it's possible"), imperative form ("just go"), contraction ("it's")
Alpaca$_{\leftrightarrow}$
`-Formal Attributes:` precise language ("feasible"), use of conditional clause ("if it is feasible"), lexical sophistication ("prevail") `-Informal Paraphrase:` let love win if it's possible. `-Informal Attributes:` contraction ("it's"), casual tone ("let love win")

Table 6: Comparison between generations from a one-shot instruction-tuned model (Vicuna, ChatGPT), and our best small student fine-tuned model for explainable formality style transfer.

Biased (Input)
some cacti produce **beautiful** flowers, which like spines and branches arise from areoles. `-Bias Attributes (Gold):` Framing ("beautiful" adds subjectivity and implies a positive evaluation of the flowers)
Vicuna$_{10}$ (Incorrect)
`-Bias Attributes:` Framing (using "some" suggests that not all cacti produce beautiful flowers) `-Neutralized Paraphrase:` Certain cacti produce beautiful flowers, which like spines and branches arise from areoles.
GPT-3.5$_{10}$ (Incorrect)
`-Bias Attributes:` Epistemological (using the word "beautiful" implies that the flowers produced by cacti are universally considered attractive) `-Neutralized Paraphrase:` Some cacti produce flowers, which, like spines and branches, arise from areoles.
LLaMA$_{\rightarrow}$ (Correct)
`-Bias Attributes:` Framing ("beautiful" suggests a subjective evaluation of the flowers) `-Neutralized Paraphrase:` Some cacti produce flowers, which like spines and branches arise from areoles.

Table 7: Comparison between generations from a few-shot instruction-tuned models (10-shot Vicuna, ChatGPT-3.5), and our best small student fine-tuned model for explainable bias style transfer.

5 Extrinsic Evaluation of Formality Style Transfer Explanations

We use Authorship Verification (Martindale and McKenzie, 1995; Coulthard, 2004; Neal et al., 2017) as an extrinsic task utilizing PAN 2022 (Bevendorff et al., 2022) data. This is a binary classification task of deciding if two texts belong to the same author or not. The two input texts are two raw documents (e.g., paragraphs from blog posts), one of which is written by author A, and another one which is either written by the same author A or by a different author B. We then derive a representation (features to be used for classification) of these input texts using our explainable style transfer model. We run Alpaca$_{\text{IF}\rightarrow\text{F}}$ on each text (at most 15 sentences per author are considered) and extract explanations containing informality attributes and evidence (see Table 8). In a preliminary evaluation of the usefulness of these features, we compute the similarity between authors by measuring the percentage of overlapping attributes. Note that the evidence fragments corresponding to the attributes are not used in this preliminary experiment. We then use the percentage of overlapping attributes as a classification score, for example, if author A uses colloquialism and textese, and author B uses colloquialism and abbreviation, their similarity score is the number of common attributes divided by the number of unique attributes between the authors, or $\frac{1}{3}$ (0.33). Here, only 1 attribute is common (colloquialism) whereas the total identified attributes is 3 (colloquialism, abbreviation, textese). Following the computation described above, we take the similarity score as a confidence for a binary prediction task. If the similarity is high, there is a high chance the authors are the same (prediction = 1), and if not they are likely not the same (prediction = 0). The

underlying assumption is that it is more likely that the same author would use some of the informality features they used previously (not every sentence in PAN is completely informal, but informality is a very broad category, so when authors do use some informal attributes they can provide a signal for authorship). We compute ROC AUC between the confidence scores (author similarity score) and ground truth predictions from the PAN dataset (1 if authors are the same and 0 if not). We compare explanations from Vicuna$_{10}$, GPT-3.5$_{10}$ and Alpaca$_{\text{IF}\rightarrow\text{F}}$ by their predictive signal for this task. Explanations by Alpaca$_{\text{IF}\rightarrow\text{F}}$ achieve an AUC of 56.4, whereas explanations from the Vicuna and GPT-3.5 models achieve a score of 50.0 and 47.0 respectively. This indicates a potential application of the explanations generated by the student model fine-tuned on our dataset (E-GYAFC) to be used as interpretable authorship features that can be explored in future work.

Attribute	Evidence
Colloquialism	"assumed they all started off low!?", "typing it out"
Textese	"xx"
Informal Vocabulary	"give you a call", "arrange something"
Informal Tone	"hoping to borrow a couple of charging leads"

Table 8: Informality features for authorship identification: on the left, informality attributes identified by our model, on the right, textual evidence provided by it.

6 Conclusion

We propose a framework to augment two style transfer datasets with semi-structured textual explanations. To improve quality of model distillation and incorporate expert feedback, we propose ICLEF (In-Context Learning from Expert Feedback), a novel human-AI collaboration framework leveraging both in-context learning and self-critique abilities of LLMs. We evaluate smaller student models fine-tuned on the resulting datasets compared to large teacher models and conduct expert human evaluation. We also extrinsically evaluate the explanations for the formality style transfer on the downstream task of authorship attribution.

7 Limitations

The GYAFC dataset does not contain all types of informal and formal language, namely, they mostly focus on interpersonal relationships (the subset used for this paper) and entertainment. Future work could consider extending our approach to other style transfer datasets, including ones more encompassing of formality.

While our methods are intended to produce faithful explanations, there can still be instances when a model does not rely on the attributes in order to complete the paraphrase. We also observed that hallucinations can still be present in our fine-tuned models' explanations and hope that future work will try to address these issues. We also note that our approach does not replace expert annotation as it heavily relies on LLMs that may still hallucinate. It is only meant to be applicable in scenarios where expert feedback is expensive and/or difficult to gather.

One limitation of our method is that we used a relatively small number of experts to conduct our study. However, we believe that this setting mirrors real-life conditions where experts are usually scarcely available. We hope our approach provides a more general framework for incorporating expert feedback that can be adjusted to experts' needs (e.g., a forensic linguist may require a different style transfer explanation than a literary critic).

Fine-tuning and running inference on large models requires expensive computational resources. However, we hope that our study presents a convincing argument that fine-tuning a smaller model once may be more efficient and accurate than running a large general-purpose model with elaborate long-context prompts.

8 Ethics Statement

The GYAFC corpus was created using the Yahoo Answers corpus: L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0. This Yahoo Answers corpus can be requested free of charge for research purposes. Access to our GYAFC dataset will require users to first gain access to this Yahoo Answers corpus. Authors obtained permission to access the dataset.

Our datasets do not include any protected data to the best of our knowledge. All annotators are fairly compensated for their work in accordance with their asking rate (typically over 20 USD per hour).

Our bias style transfer model is only intended for use in a human-in-the-loop fashion and not by itself to adjudicate bias in text. We hope that the explanation generation capacity of our model will

improve upon existing bias classifiers that typically do not provide textual explanations. In Appendix G, we show how style transfer can be used to evade AI-text detectors. Similarly to Krishna et al. (2023), we reiterate that this is not to provide a way to attack such systems, but to bring awareness to the community that current detectors are easy to evade. Moreover, we bring to attention the issue of detecting text on which style transfer paraphrase has been applied. We hope that future work develops systems capable of defending against such attacks, perhaps utilizing explanations generated by our system.

GYAFC and WNC may potentially contain offensive data as they are crowdowdsourced, however, on samples that we saw we did not find alarming ethical issues.

Acknowledgements

We thank the annotators for their work and providing detailed feedback. We also would like to thank the reviewers for productive and engaging discussions. This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Nikolay Babakov, David Dale, Varvara Logacheva, and Alexander Panchenko. 2022. A large-scale computational study of content preservation measures for text style transfer and paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 300–321, Dublin, Ireland. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 2022. Models in the loop: Aiding crowdworkers with generative annotation assistants. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3754–3767, Seattle, United States. Association for Computational Linguistics.

Edward Beeching, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

Janek Bevendorff, Berta Chulvi, Elisabetta Fersini, Annina Heini, Mike Kestemont, Krzysztof Kredens, Maximilian Mayerl, Reynier Ortega-Bueno, Piotr Pęzik, Martin Potthast, et al. 2022. Overview of pan 2022: Authorship verification, profiling irony and stereotype spreaders, and style change detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference*

of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, pages 382–394. Springer.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. 2022. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10925–10934.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541, New York, NY, USA. Association for Computing Machinery.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sahil Chaudhary. 2023. Code alpaca: An instruction-following llama model for code generation. `https://github.com/sahil280114/codealpaca`.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Malcolm Coulthard. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied linguistics*, 25(4):431–447.

Databricks. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. Blog post.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *ArXiv*, abs/2212.10071.

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.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

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.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Josh Magnus Ludan, Yixuan Meng, Tai Nguyen, Saurabh Shah, Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2023. Explanation-based finetuning makes models more robust to spurious cues.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason.

Colin Martindale and Dean McKenzie. 1995. On the utility of content analysis in author attribution: "the federalist". *Computers and the Humanities*, 29(4):259–270.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6).

OpenAI. 2023. New ai classifier for indicating ai-written text.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2022. Low-resource authorship style transfer with in-context learning.

Ajay Patel, Delip Rao, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting llms.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

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)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. 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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Scott Rosenberg and Sara Fischer. 2023. Newsrooms try ai to check for bias and error. Accessed: 2024-02-15.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

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.

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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-06-20.

Edward Tian. 2023. Ai content detector and writing captcha for chat gpt, openai, bard, education.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks.

PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2023a. PINTO: Faithful language reasoning using prompt-generated rationales. In *The Eleventh International Conference on Learning Representations*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023b. How far can camels go? exploring the state of instruction tuning on open resources.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Anna Wegmann and Dong Nguyen. 2021. Does it capture STEL? a modular, similarity-based linguistic style evaluation framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7109–7130, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-kirkpatrick. 2022. Paraphrastic representations at scale. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 379–388, Abu Dhabi, UAE. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Yang Zhong, Jingfeng Yang, Wei Xu, and Diyi Yang. 2021. WIKIBIAS: Detecting multi-span subjective biases in language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1799–1814, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Human evaluation details: E-GYAFC

We hire three annotators for preference evaluation: A1 (bachelors degree in linguistics), A2 (bachelors degree in linguistics and a masters degree in education), A3 (bachelor and master degrees in linguistics) on Upwork. Due to high prevalence of the positive class, there is a high chance of random agreement, hence we provide a more granular look into the expert annotations than the inter-rater agreement in Table 9. Pairwise accuracy between annotator responses for all categories of E-GYAFC evaluation, and found that it averages at 81% across all categories. Annotator A2 expressed concerns that the paraphrases may sound unnatural due to excessive formality (we believe it is due to the context in which the informal expression would be uttered) and that explanations sometimes miss punctuation errors (which, while important, is not critical for model-generated explanations).

	e_i pref.	e_i accept.	s_f pref.	s_f accept.	e_f accept.
A1	91%	95%	64%	64%	98%
A2	87%	84%	76%	75%	96%
A3	91%	83%	91%	93%	100%

Table 9: Expert evaluation of E-GYAFC dataset quality. We report percentage of time each item was preferred, as well as acceptability judgements.

B Human evaluation details: E-WNC

For E-WNC, we hire two expert annotators A1 and A2. A1 is a professional with over ten years experience in translation and interpretation, data annotation, linguistics and publishing. A2 is a PhD in linguistics with background in psycholinguistics and neurolinguistics and experience in writing, proofreading and editing academic texts

	e_b pref.	e_b accept.	s_n pref.	s_n accept.
A1	82%	84%	84%	86%
A2	74%	62%	70%	62%

Table 10: Expert evaluation of E-WNC dataset quality. We report percentage of time each item was preferred, as well as acceptability judgements.

C How does ICLEF performance change depending on the amount of feedback provided?

We perform a small-scale study to explore how performance of the self-critique component of ICLEF changes depending on the number of in-context examples provided. We sample 15 instances of synthetic explanations (before applying ICLEF) and evaluate generated critiques for correctness when providing 1, 10, or 35 instances of expert feedback few-shot. We evaluate the critiques for correctness, i.e. they have to both identify the incorrect attribute and not introduce any incorrect attributes. Figure 4 shows how by increasing the amount of feedback to 35-shot, the correctness is raised to $\approx$ 87% for E-GYAFC and $\approx$93% for E-WNC, which can be deemed satisfactory.

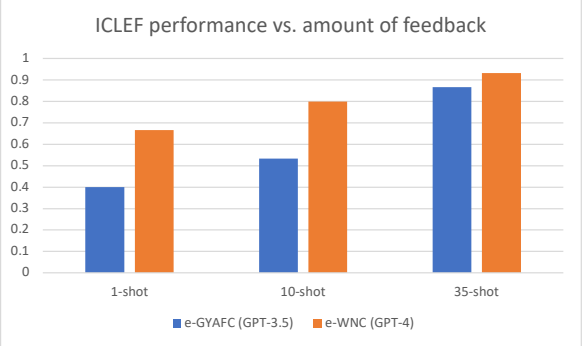

Figure 4: ICLEF performance increases with amount of feedback, reaching satisfactory accuracy at around 35 shots.

D Additional estimations for trustworthiness of explanations

We conduct expert evaluation of our data in Section 3.3. To further confirm trustworthiness of explanations, we check that every evidence fragment actually occurs in the text. For every data instance, we compute what percentage of evidences was present in the sentence. We report average across the respective datasets and explanations in Table 11 summarizes the results, where it is evident that the explanations have high trustworthiness. We note that some explanations discuss absence of certain words in which case they should not be present in the sentence in the first place. Moreover, our structured explanation format would make it very easy for the end-user to verify the correctness of the explanation and alert them of potentially incorrect paraphrase.

	e-GYAFC		e-WNC
	$\mathbf{e}_i$	$\mathbf{e}_f$	$\mathbf{e}_u$
Trustworthiness	95.0%	92.3%	98.9%

Table 11: Explanation Trustworthiness Rates Across Datasets: percentage of evidences actually present in the sentence.

E Prompts

You are an extremely attentive and critical annotator with background in stylometry and linguistics. You will be provided with an informal sentence. You will also be provided with an explanation of its informality attributes. Decide whether the explanation is incorrect, and if so, describe what attributes were listed incorrectly. EXAMPLES: <...>
You are an extremely attentive and critical annotator with background in ethics, journalism, critical thinking and bias identification. You will be provided with a possibly biased sentence. You will also be provided with an explanation of its bias attributes. <...> If the explanation is incorrect, reply with a correction. Focus on three main types of bias in a sentence: <...> EXAMPLES: <...>

Table 12: Prompts for LLM-critic models. Top is used for formality style transfer, bottom is used for subjective bias style transfer.

ChatGPT Explanation Generation Prompt We provide an instruction as a system prompt ("You are an expert forensic linguist...") and 6 examples of the task in the OpenAI ChatML format.[8]

`Instruction: You are an expert forensic linguist. Your task is to identify infromal attributes in a setnece, modify them to create a formal sentence, and then output the attributes of the generated formal sentence. Use the following format: attribute (excertp from text in quotation marks). Make sure to provide a complete list of informal and formal atttributes. Focus on what has changed between formal and informal sentences. Informal writing tends to be more casual, personal, and conversational than formal writing. Here are some common features of informal writing: Contractions: Informal writing often uses contractions, such as "I'm," "can't," "won't," and "they've," which are generally avoided in formal writing. ...`

`Examples: Informal greetings and sign-offs: Informal writing often uses casual greetings, such as "Hi" or "Hey," and sign-offs like "Cheers" or "Take care." Informal: if ur under 18 u have a BIG PROBLEM. Informal Candidates: '18', 'BIG', 'PROBLEM.', 'if', 'u', 'ur Attributes of Informal Style: textese ("ur", "u"), capitalization ("BIG PROBLEM"), colloquialism ("BIG PROBLEM")...`

Prompt for ChatGPT ICLEF generation `You are an extremely attentive and critical annotator with background in stylometry and linguistics. You will be provided with an informal sentence. You will also be provided with an explanation of its informality attributes. Decide whether the explanation is incorrect, and if so, describe what attributes were listed incorrectly.`

`EXAMPLES: Informal Sentence: Look, If you really like this person, just tell her. Informal Attributes: colloquialism ("just tell her"), contraction ("If you"), simple sentence structure. Attributes Listed Incorrectly: contraction ("If you" is not a contraction)...`

F Model details, additional models and experiments

Below we provide some additional experiments and models used for them as well as experimentation details.

Instruction-tuned Models All instruction-tuned models are provided with the same one-shot prompt (modulo special token requirements) and generation parameters.

- MPT-7B-Instruct: built by finetuning MPT-7B (Team, 2023) on a dataset derived from the Databricks Dolly-15k (Databricks, 2023) and the Anthropic Helpful and Harmless (Bai et al., 2022a) datasets.

- Alpaca-7B (Taori et al., 2023) a model fine-tuned from the LLaMA-7B model on 52K instruction-following demonstrations generated with the Self-Instruct framework (Wang et al., 2023c).

[8] github.com/openai/openai-python/blob/main/chatml.md

- Vicuna-13B (Chiang et al., 2023): an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT[9]. It places first in the Huggingface Open LLM Leaderboard (Beeching et al., 2023) based on human and GPT-4 evaluation as of writing this paper.

- Falcon-40B (Almazrouei et al., 2023) causal decoder-only model trained on 1,000B tokens of RefinedWeb (Penedo et al., 2023) enhanced with curated corpora.

- Tülu-65B (Wang et al., 2023b) a 65B LLaMa model finetuned on a mixture of instruction datasets (FLAN V2 (Longpre et al., 2023), CoT (Wei et al., 2022), Dolly (Databricks, 2023), Open Assistant 1 (Köpf et al., 2023), GPT4-Alpaca (Peng et al., 2023), Code-Alpaca (Chaudhary, 2023), and ShareGPT).

ChatGPT For ChatGPT-3.5 we use `gpt-3.5-turbo-1106`. For ChatGPT-4 we use `gpt-4`.

Fine-tuned models We fine-tune below models on E-GYAFC. $\rightarrow$ indicates fine-tuning two models in each direction, and $\leftrightarrow$ indicates fine-tuning on combined data in both directions.

- FLAN-T5-XL$_{\leftrightarrow}$ (Chung et al., 2022) approximately 3B parameter instruction-tuned model based on the T5 architecture (Raffel et al., 2020).

- LLaMA-7B$_{\rightarrow}$ (Touvron et al., 2023) model by Meta trained on 1 trillion tokens.

- Alpaca-7B$_{\rightarrow,\leftrightarrow}$ (Taori et al., 2023) a model fine-tuned from the LLaMA-7B model on 52K instruction-following demonstrations. In addition, we test Alpaca-7B$_{noexpl}$ as the model fine-tuned for Formal to Informal style transfer with no explanations provided in the fine-tuning data or in the output.

Fine-tuning hyperparameters We fine-tune all models using the script provided in the Stanford Alpaca repository.[10] We use exact same hyperparameters, except for batch size which we adjust to 1 due to memory constraints. We fine-tune our models on 4 A100 NVIDIA 40GB GPUs. We train our models for 3 epochs with learning rate 2e-05 with cosine rate scheduler and warmup ratio of 0.03. We did not perform hyperparameter search. We report results form single runs with random seeds preserved due to computational constraints.

Inference parameters We use the same hyperparameters for generation across all models, that is temperature=0.7, top p=0.9, max new tokens = 256. We use the huggingface library.

One-shot instruction is provided below:
`Identify informal attributes in a given sentence, modify them to create a formal sentence, and then output the attributes of the generated formal sentence.`
`For example:`
`Informal: how can you tell if a girl likes you or not?`
`Informal Attributes: direct question form ("how can you tell"), informal language ("girl", "likes you") Formal: What are some indications that a woman may be interested in you? Formal Attributes: indirect question form ("what are some indications"), lexical sophistication ("woman", "interested in you")`
`For the following sentence, identify informal attributes in a given sentence, modify them to create a formal sentence, and then output the attributes of the generated formal sentence.`

For Tulu, we add the <asistant> and <user> tokens as advised by model developers.

Packages We used transformers (Wolf et al., 2020) for language model inference and NLTK package (Bird and Loper, 2004) for sentence tokenization.

Additional experiments with instruct and fine-tuned models We additionally perform experiments with MPT-7B, Falcon-40B, Tulu, and non-fine-tuned Alpaca. We also fine-tune an Alpaca model with no explanations. See Table 13. The model fine-tuned without explanations (Alpaca$_{noexpl}$) achieves comparable performance, indicating that generating explanations does not significantly hurt performance on the standard style transfer task.

Fluency of style transfer Fluency could be used as an additional metric to measure the quality of the style transfer. However, fluency measures negatively correlate with informality, even though the semantic content stays the same. Because of that, we

[9] sharegpt.com
[10] Alpaca GitHub

Model	Size	Formal → Informal				Informal → Formal				**Average**
		Form.Attrs. BLEU	MIS	Informality	Inform.Attrs. BLEU	Inform.Attrs. BLEU	MIS	Formality	Form.Attrs. BLEU	
MPT	7B	24.59	51.84	9.82	2.10	23.26	46.26	58.40	0.86	27.14
Alpaca	7B	17.07	80.74	32.53	6.51	23.67	73.69	86.82	8.27	41.16
Falcon	40B	8.38	28.12	13.43	1.23	20.80	38.01	62.69	7.13	22.47
Tülu	65B	24.90	19.60	7.12	0.02	27.76	26.69	27.34	0.28	16.71
FLAN-T5$_{\rightarrow}$	3B	0.00	8.54	0.01	0.00	0.00	9.82	0.91	0.00	2.41
Alpaca$_{\rightarrow}$	7B	39.98	84.70	61.99	19.22	40.56	81.69	97.96	24.71	56.35
Alpaca$_{noexpl}$	7B	-	85.34	54.75	-	-	83.20	91.10	-	-

Table 13: Performance of additional instruction-tuned and fine-tuned models on the explainable formality style transfer task.

preferred to use semantic similarity of the output to the input (MIS) as the primary measure across all tasks. Below is the result of an experiment we conducted with perplexity (lower is better), following Suzgun et al. (2022) exactly to compute the fluency metric.

Model	PPL Inf. → For.	PPL For. → Inf.	PPL Bias → Unbiased
gold	33.32	68.96	31.26
gpt1	31.24	21.39	28.21
gpt5	34.83	21.10	27.59
gpt10	32.55	23.82	27.62
best model	30.21	38.61	31.97

Table 14: Fluency evaluation results. GPT refers to the teacher model, best model refers to Alpaca for formality transfer and LLaMA for bias transfer.

As can be seen from Table 14, for formal and unbiased paraphrase the perplexity is comparable with the teacher in 1-shot and few-shot settings as well as the gold reference, whereas for informality the gold data has the worst perplexity, making this metric unfit when we want to transfer from formal to informal style. We want to add that we performed a human evaluation in Section 4.3 that includes the evaluation of the transferred sentences.

Performance of PromptRerank Baseline We considered including a PromptRerank baseline (Suzgun et al., 2022) as it is one of the previous SOTA approaches. On our observation the models did not perform competitively possibly because the models described in the paper are smaller than the contemporary LLMs. Future work may explore adjustments on this baseline to adapt it to the new models. Performance is depicted in Table 15.

BLEU and BERTScore between transferred text and gold references We opted for reference-free metrics since they would be less biased toward producing output that most closely matches the reference. However, we now ran the experiment with BLEU and BERTScore between transferred text (output) and gold reference from our e-GYAFC and e-WNC datasets. BERTScore is based on contextualized-embeddings and usually preferred to using BLEU-4 (which is based on 4-gram overlap) for evaluation of text generation and paraphrases. Results are shown below in Table 16.

We see that using BERTScore, the student model is comparable to the teacher model and both are semantically close to the reference. BLUE scores are lower due to more stricter n-gram overlap requirements but that is across all models and scores are again comparable between smaller student models and the larger teacher models.

G Discussion on applications of models fine-tuned for explainable style transfer

Explainable formal → informal style transfer is an interpretable adversarial attack on AI-generated text detection methods, including retrieval Krishna et al. (2023) established that paraphrasing easily evades the detection of AI-generated text, and proposed a retrieval-based defense. However, we hypothesize that retrieval-based metrics will degrade as similarity between generations becomes more ambiguous, as is the case for formality style transfer. For example, an adversarial agent might generate a post containing misinformation in typical “formal" language generated by a language model like ChatGPT. This text is relatively detectable by current classifiers and 100% detectable by retrieval-based methods. However, the agent might apply a style transfer model to lower the formality of the message. Alarmingly, not only this accomplishes the goal of spreading the AI-generated message more effectively as the result looks more like user-generated text, but, as

Model	MIS (Formal to Informal)	Informality	MIS (Informal to Formal)	Formality
Alpaca$_{\leftrightarrow}$	81.76	66.71	79.43	98.57
promptRerank	63.33	33.11	62.70	20.00

Table 15: Model performance on style transfer tasks.

Model	Formal → Informal		Informal → Formal		Bias → Unbiased	
	BLEU	BERTScore	BLEU	BERTScore	BLEU	BERTScore
gpt1	15.01	78.40	28.10	84.29	60.34	91.30
gpt5	15.36	78.74	28.71	84.48	64.85	91.78
gpt10	14.23	78.14	29.13	84.62	65.98	92.13
best model	15.33	77.57	20.75	80.78	70.49	93.23

Table 16: Performance comparison of models on BLEU and BERTScore metrics with the gold reference. GPT refers to the teacher model, best model refers to Alpaca$_{\leftrightarrow}$ for formality transfer and LLaMA$_{\rightarrow}$ for bias to unbiased transfer.

we show, it also decreases the chances of being detected as AI-generated by current methods.

We test this in the following setting: we use an online dataset of political tweets [11], and sample 30 of them. We ask ChatGPT to generate a political commentary post on the topic of the tweet (GPT-F), as well as an informal paraphrase of the said post (GPT-Inf). We manually annotate the resulting summaries and select those that look like they could be legitimate political messages posted on social media and have valid paraphrases. We then use our Alpaca$_{\text{F}\rightarrow\text{IF}}$ model to generate an informal paraphrase of the GPT-F posts sentence-by-sentence. We also verify that these paraphrases are semantically valid and close to the original GPT-Formal post and select 24 high-quality generations. We choose a relatively small sample since we want to verify the paraphrase was still close to the original sentence manually to ensure semantic control for the experiment.

We report detection scores[12] from 4 methods surveyed by Krishna et al. (2023): GPTZero (Tian, 2023), OpenAI classifier (OpenAI, 2023), DetectGPT (Mitchell et al., 2023), and their proposed retrieval methods based on BM25 (Robertson and Zaragoza, 2009) or P-SP (Wieting et al., 2022) retrievers. As can be seen in Table 17, the formal-to-informal transfer model significantly decreases detection scores of all AI-generated text detection methods, including the retrieval-based one (despite the fact that the retrieval corpus is significantly smaller than it would be in real-world). Interestingly, for the BM25 retrieval method, the ChatGPT paraphrases are slightly harder to detect than Alpaca$_{\text{F}\rightarrow\text{IF}}$, whereas it is easier for all other methods. Since we used ChatGPT to generate the original posts, we could not use the watermarking methods (Kirchenbauer et al., 2023), but this can be explored in future work.

This result highlights the need to investigate new methods of detecting style transferred AI-generated text. As formality style transfer remains an effective attack, informality features produced by our model could help improve such classifiers. We leave this investigation for future work.

Models	GPTZero	OpenAI	GPTDetect	BM25	P-SP
GPT-F	85.92	70.64	104.88	100	100
GPT-Inf	69.58	54.24	65.42	**48.15**	74.99
F → IF	**6.11**	**44.86**	**54.92**	58.68	**74.08**

Table 17: Performance of various AI-generated text detectors on informal paraphrases from our model. Even retrieval methods perform poorly in this setting.

H E-EGYAFC Statistics

We provide the distribution of 50 most frequent informal and formal attributes in E-GYAFC in Figures 5, 6.

I E-WNC Statistics

We provide a proportion of classes in E-WNC in Table 18.

[11] kaggle.com

[12] Since we do not have the human baseline text, we do not report the performance at 1% FPR, but for our study it is sufficient to show the scores decrease in absolute terms.

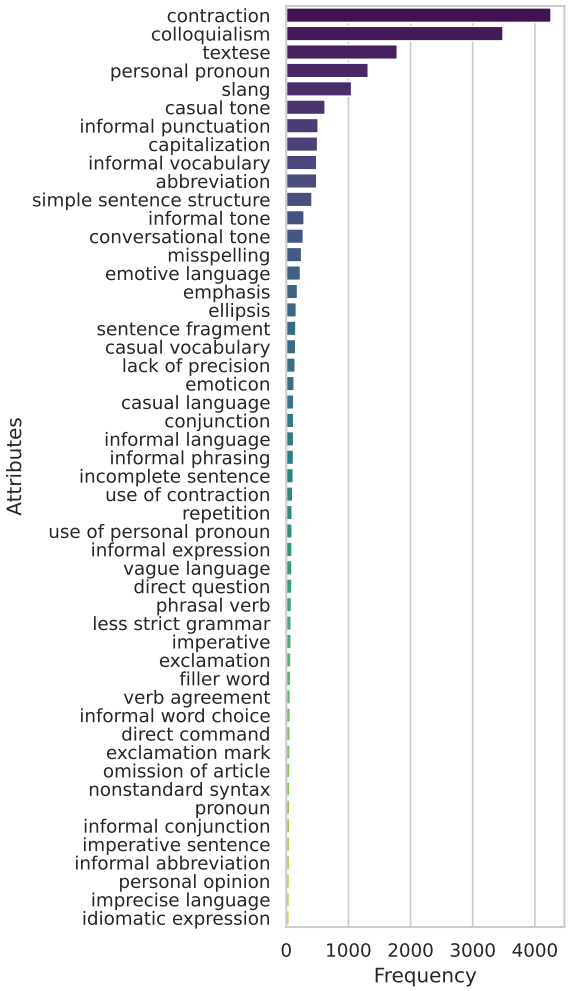

Figure 5: Distribution of 50 most frequent informal attributes in the E-GYAFC dataset.

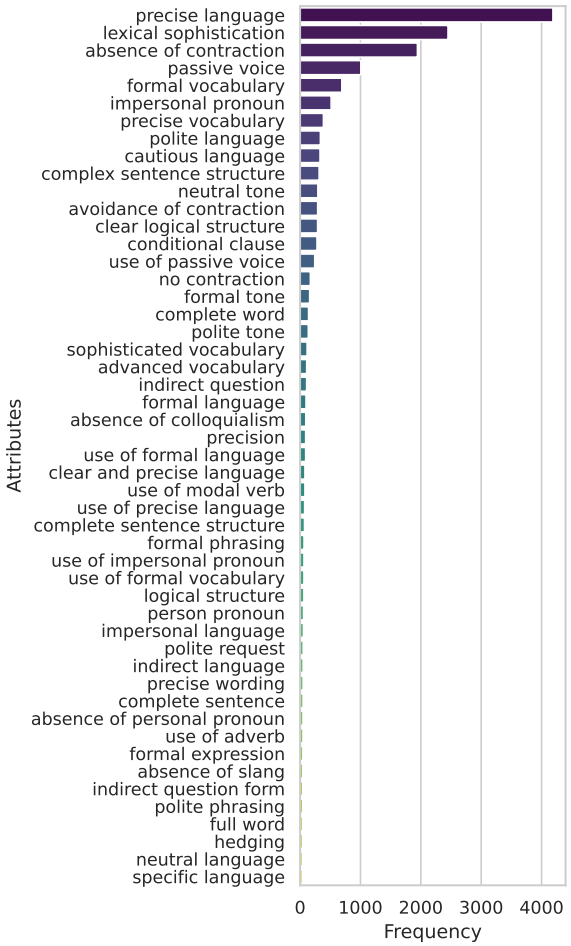

Figure 6: Distribution of 50 most frequent formal attributes in the E-GYAFC dataset.

J Annotation protocols

The screenshots for explanation interfaces are provided below in Figures 7, 8, 9. Similar annotation interfaces were used for the bias task.

K Annotator demographics

Annotators are part of a diverse demographic, geographically present in North America, Europe, and Southeast Asia (as reported by Upwork). All annotators indicated at least fluent to native English skill.

L AI assistants

The authors used AI assistants such as Co-Pilot and ChatGPT for writing code.

Category	(%)
Demographic	3.70
Epistemological	22.87
Framing	67.53
No Bias	5.90

Table 18: Proportion of classes in E-WNC

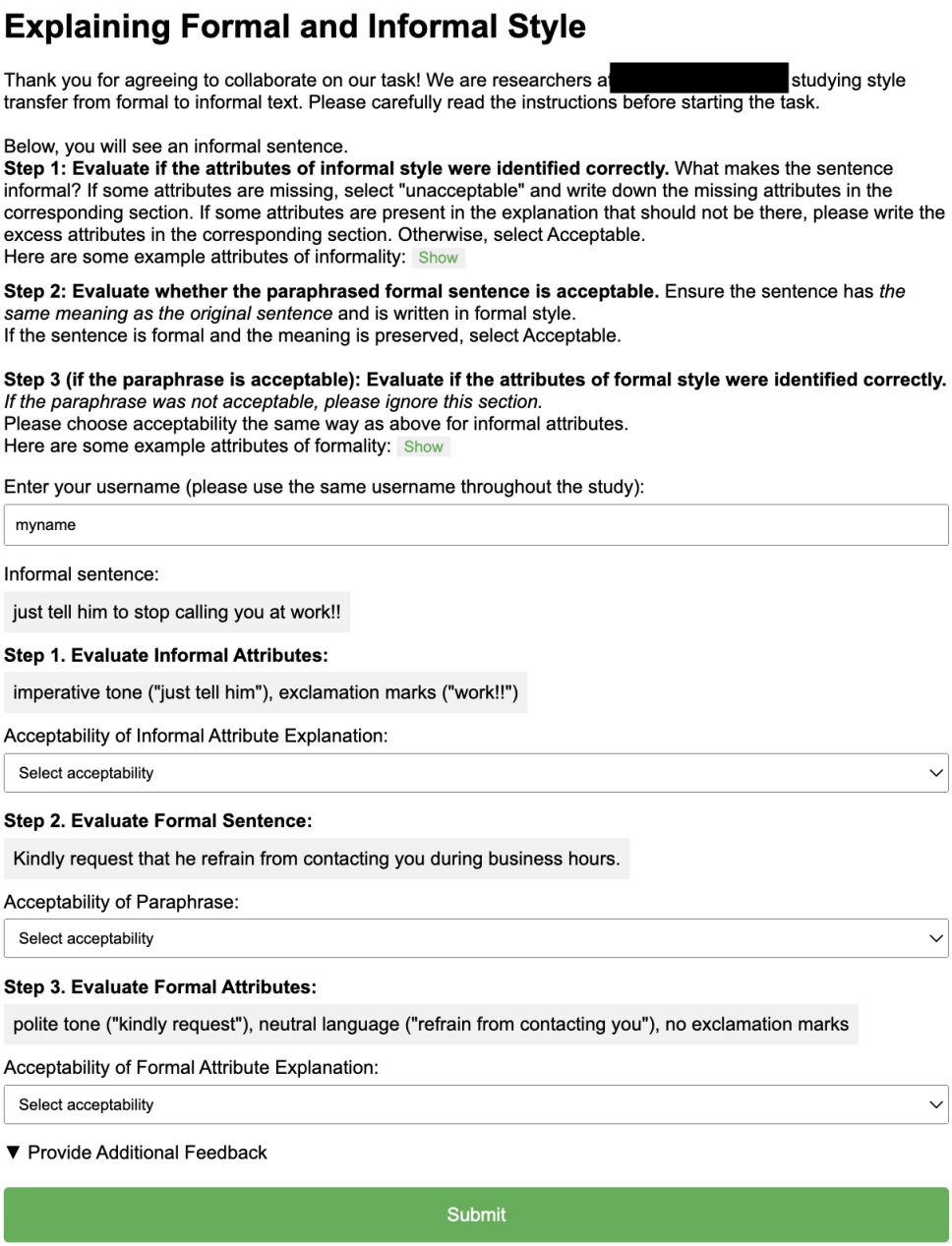

Figure 7: Annotation to gather feedback for ICLEF.

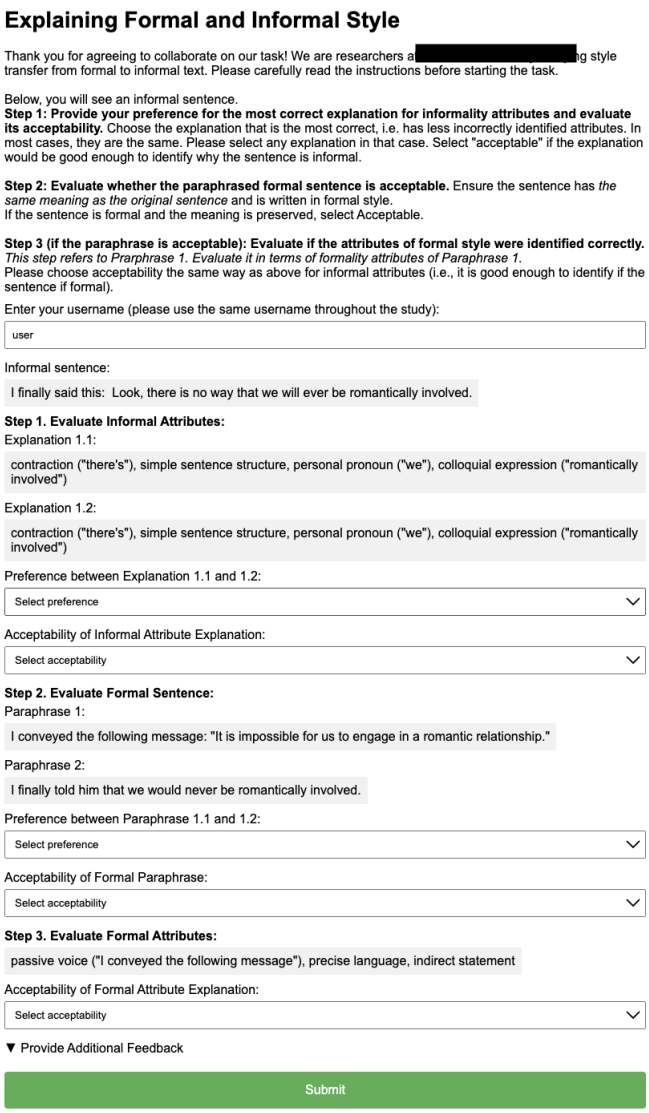

Figure 8: Annotation for eGYAFC data acceptability and preferences.

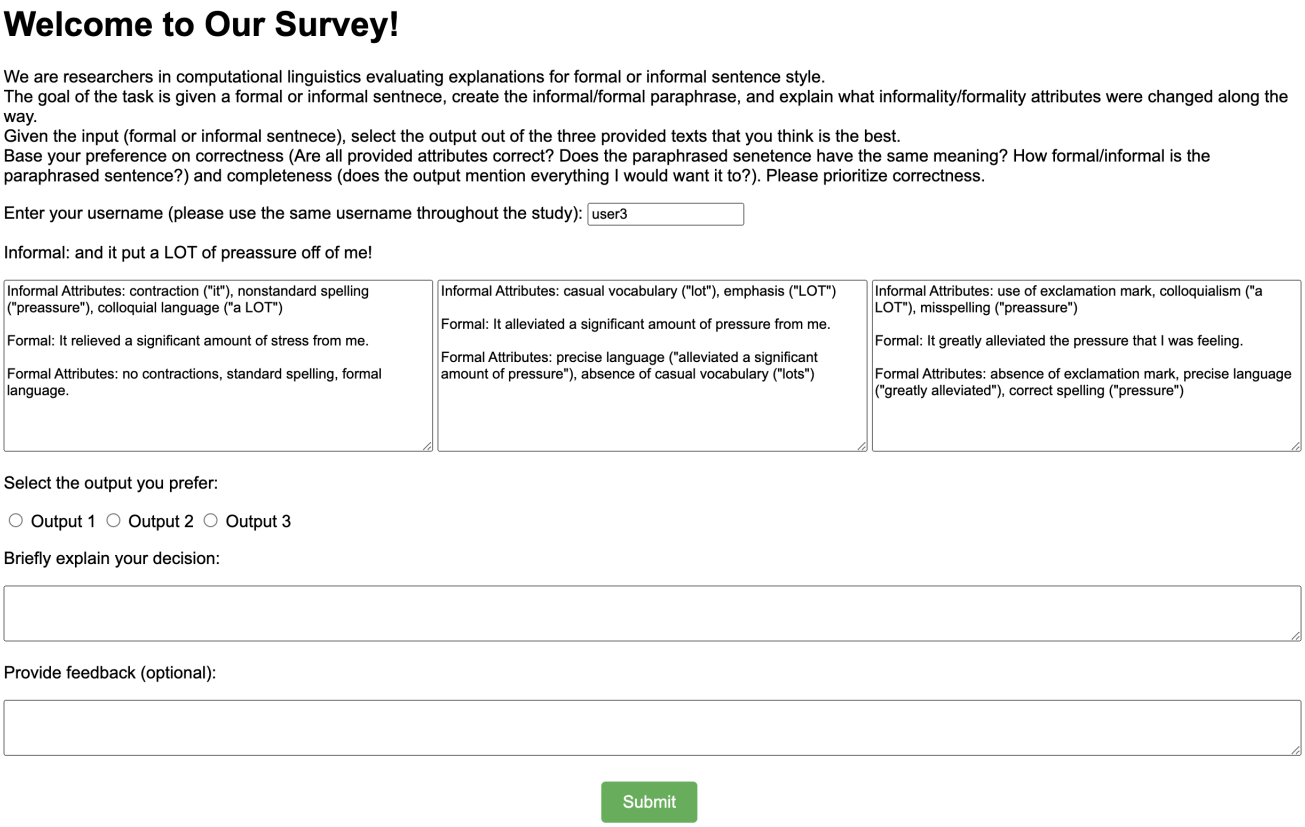

Figure 9: Annotation for model preferences.