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

Arkadiy Saakyan [1], Smaranda Muresan [1]

[1]Columbia University

Motivation and Contributions

- New task of explainable style transfer: in addition to sentence rewriting, generate textual explanations of what attributes were changed.
- Novel human-AI collaboration framework, In Context-Learning with Expert Feedback (ICLEF) (see Figure 1, Figure 2). Combines model distillation for explanation generation [2, 4] with self-critique ability of LLMs [3, 1, 7, 8], 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 [6] and the neutralizing subjective bias dataset WNC [5] with textual explanations.
- 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 (see Tables 2, 1, and Table 1 for examples).
- Experiments that show that student models outperform teacher models in one-shot setting and perform comparably even with few-shot teacher models in automatic (see Figure 3) and expert evaluation (see Figure 4).
- We show that explanations generated by student models fine-tuned on our data produce a better signal for the authorship attribution task (see Figure 5). We also show that informal paraphrase from our model results in most drastic performance reduction of AI-generated text detectors (see Figure 6).

Overview: e-GYAFC generation

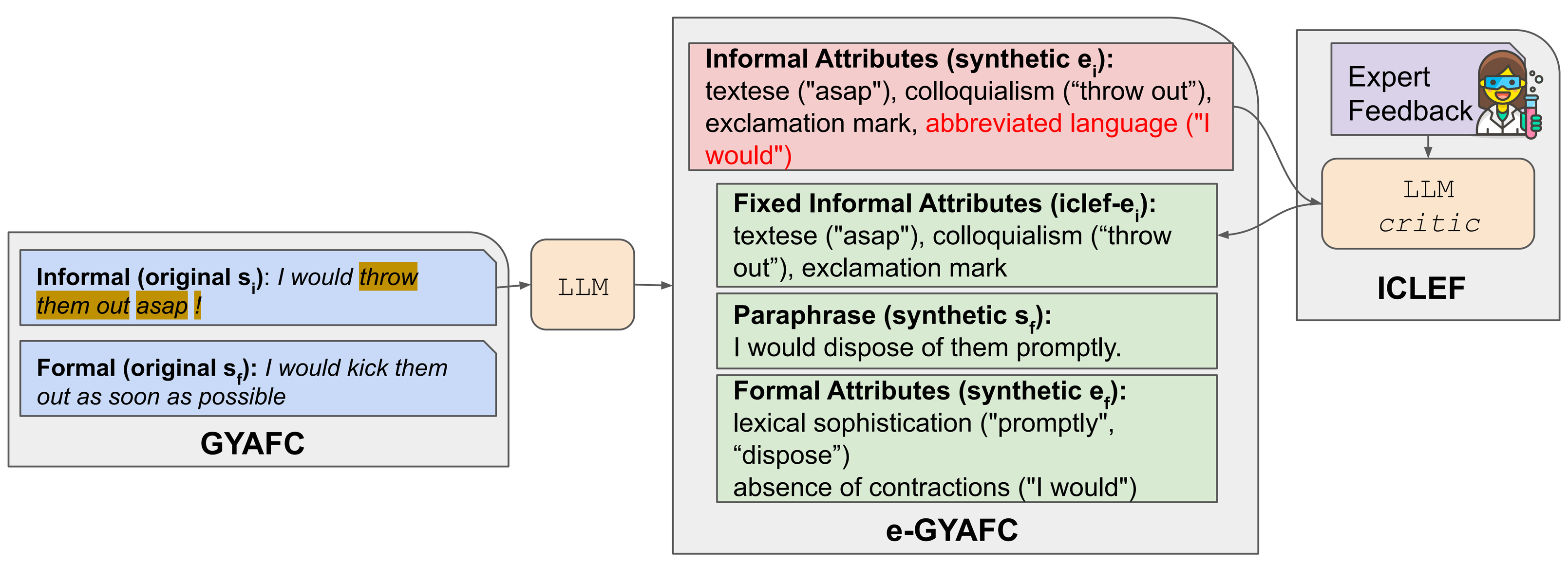

Figure 1. Generating e-GYAFC: formality style transfer dataset GYAFC [6] is augmented with semi-structured natural language explanations. The LLM generates the informal attributes of the input sentence, a formal paraphrase, and the formal attributes of the resulting sentence. Expert feedback is incorporated via in-context learning and self-critique to refine the initial generations.

Overview: e-WNC generation

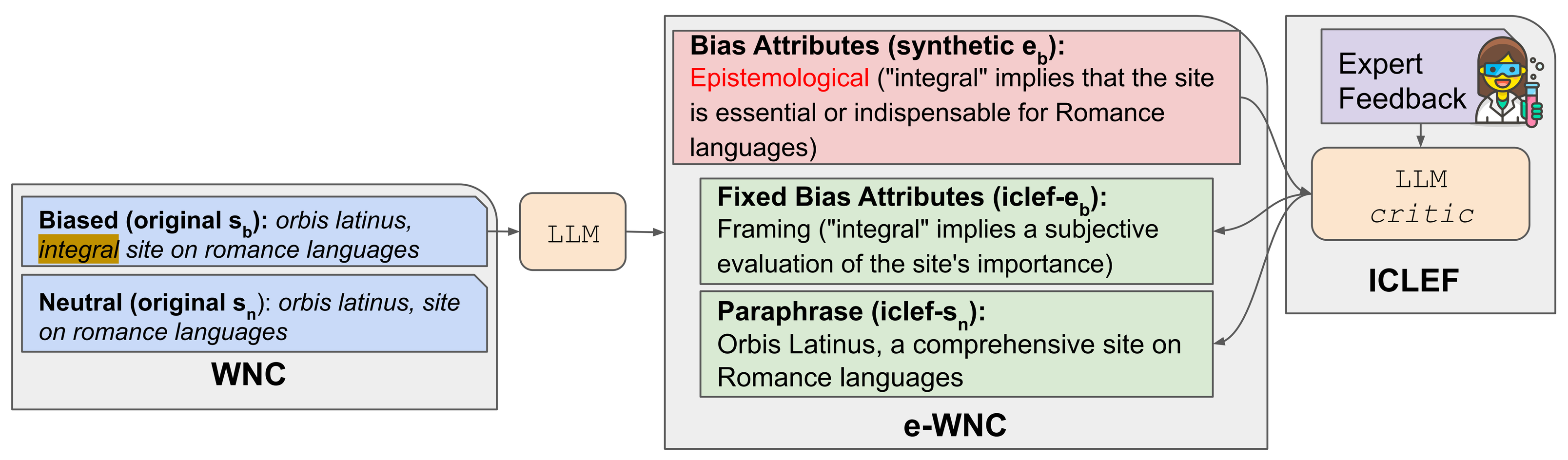

Figure 2. Generating e-WNC: WNC [5] is augmented with natural language explanations. The LLM generates the bias attributes of the input sentence and an unbiased paraphrase. Expert feedback is incorporated via in-context learning and self-critique to refine the initial generations.

Comparison: Before and After ICLEF

Informal (s_i)	Gen. expl. (synthetic e_i)	ICLEF expl. (icelf-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")
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)

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

Dataset Quality

- Automatic evaluation

	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

	e-GYAFC			e-WNC	
	e_i	s_f	e_f	e_b	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.

Model Evaluation

- Automatic Evaluation

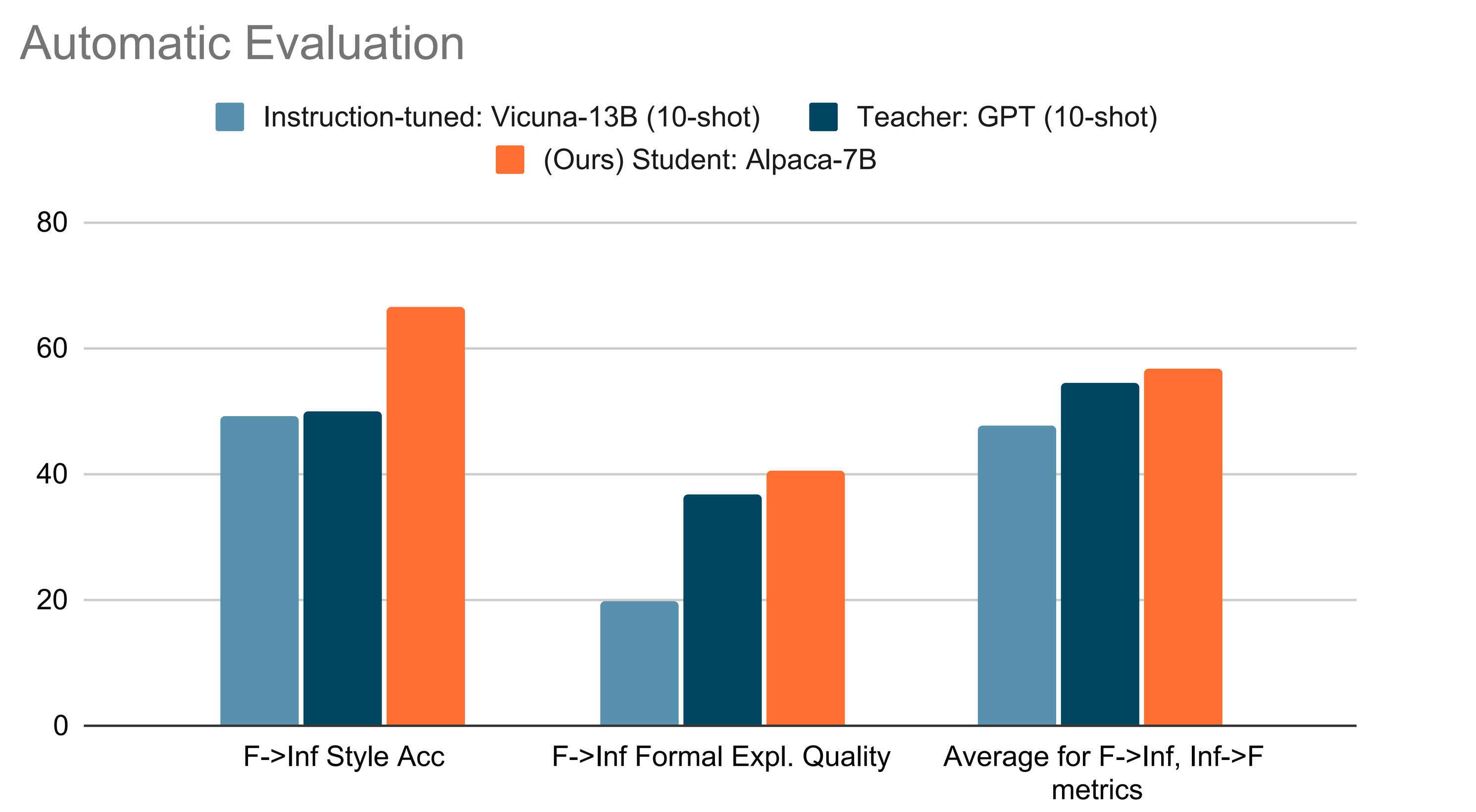

Figure 3. Performance of instruction-tuned and fine-tuned models on the explainable formality style transfer task.

- Human Evaluation

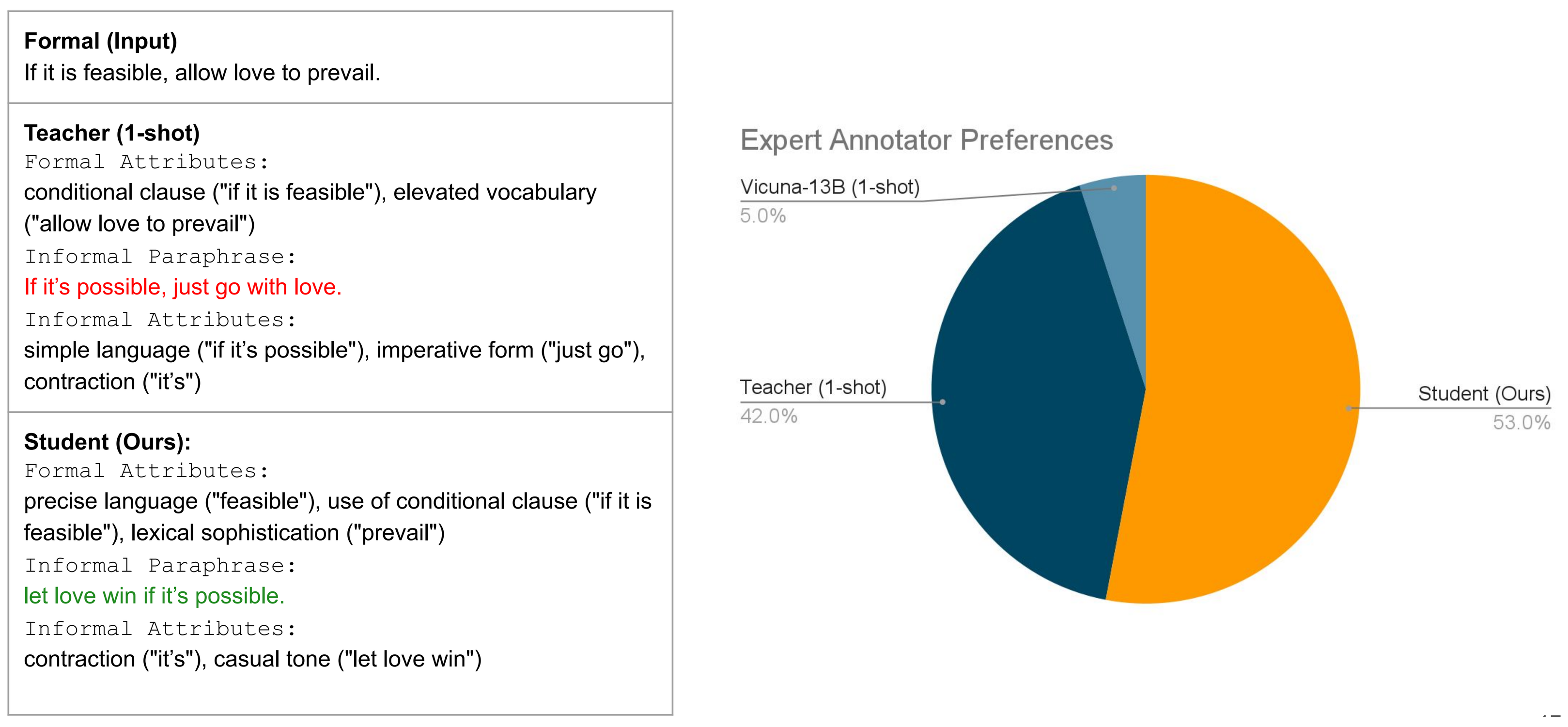

Figure 4. Comparison between generations from a one-shot instruction-tuned model (Vicuna, ChatGPT), and our best small student fine-tuned model for explainable formality style transfer.

Extrinsic Evaluation

- Comparing informality explanations on their predictive value for authorship verification task.
  - **Task:** decide if two texts belong to the same author
  - **Approach:**
    - Apply explainable style transfer model to extract informality attributes
    - Use % of overlapping attributes as a score

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

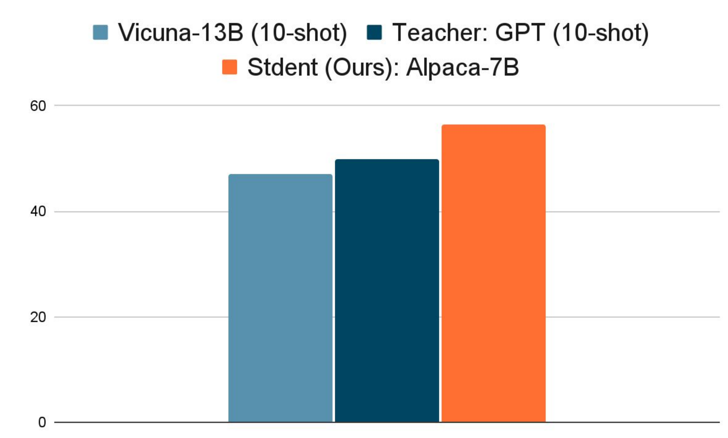

Figure 5. Using informality features for authorship detection.

- Informality paraphrase reduces efficacy of AI-generated text detectors.
  - How well can informal paraphrase be detected as AI-generated?
    - GPT-F: Formal (AI-generated)
    - GPT-Inf: ChatGPT informal paraphrase
    - Our model

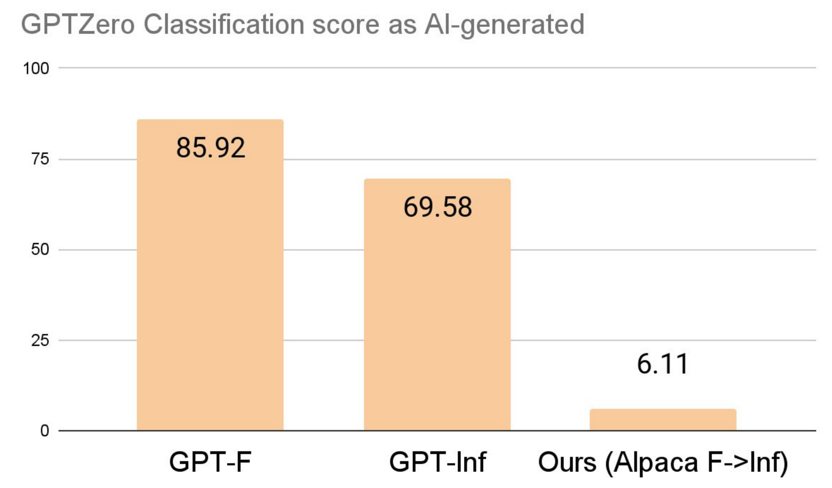

Figure 6. Comparison between original ChatGPT generation, ChatGPT informal paraphrase and informal paraphrase by our model.

References

[1] 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. Constitutional ai: Harmlessness from ai feedback, 2022.

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

[3] 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. Self-refine: Iterative refinement with self-feedback, 2023.

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

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

[6] Sudha Rao and Joel Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, 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, June 2018. Association for Computational Linguistics.

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

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