COLUMBIA UNIVERSITY

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

Arkadiy Saakyan
Columbia University

Smaranda Muresan
Columbia University

Attribute style transfer

Current:

I would throw them out asap ! → I would dispose of them promptly.

Formality

Dear Sir or Madam, May I introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer
Sudha Rao, Joel Tetreault

John McCain exposed as an unprincipled politician

↓

John McCain described as an unprincipled politician

Bias

Automatically Neutralizing Subjective Bias in Text
Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, Diyi Yang

What we want:

I would throw them out asap ! → I would dispose of them promptly.

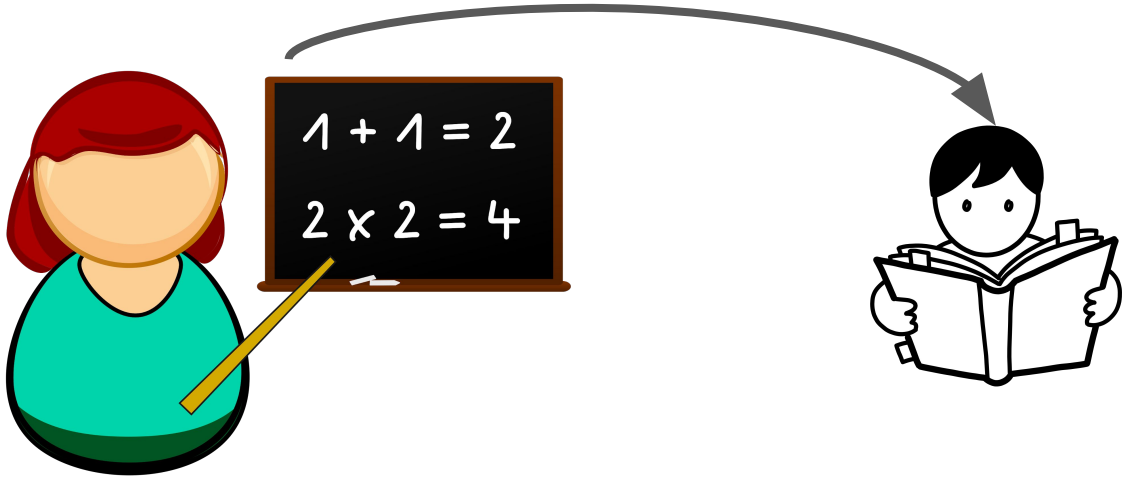
Informal Attributes:
textese ("asap"),
colloquialism (“throw out”),
exclamation mark

John McCain exposed as an unprincipled politician → John McCain described as an unprincipled politician

Bias Attributes:
Epistemological (“Exposed” is a factive verb that presupposes the truth of its complement that McCain is unprincipled)

Model distillation framework

- Teacher model generates explanations
- Generations from teacher model are used to fine-tune smaller models



Teacher model (Large LM)

Student model (Smaller LM)

Key Challenges

1. Necessity to rely on large models
2. Explanations may not be accurate
3. Scarcity of expert feedback

Proposed Solution:

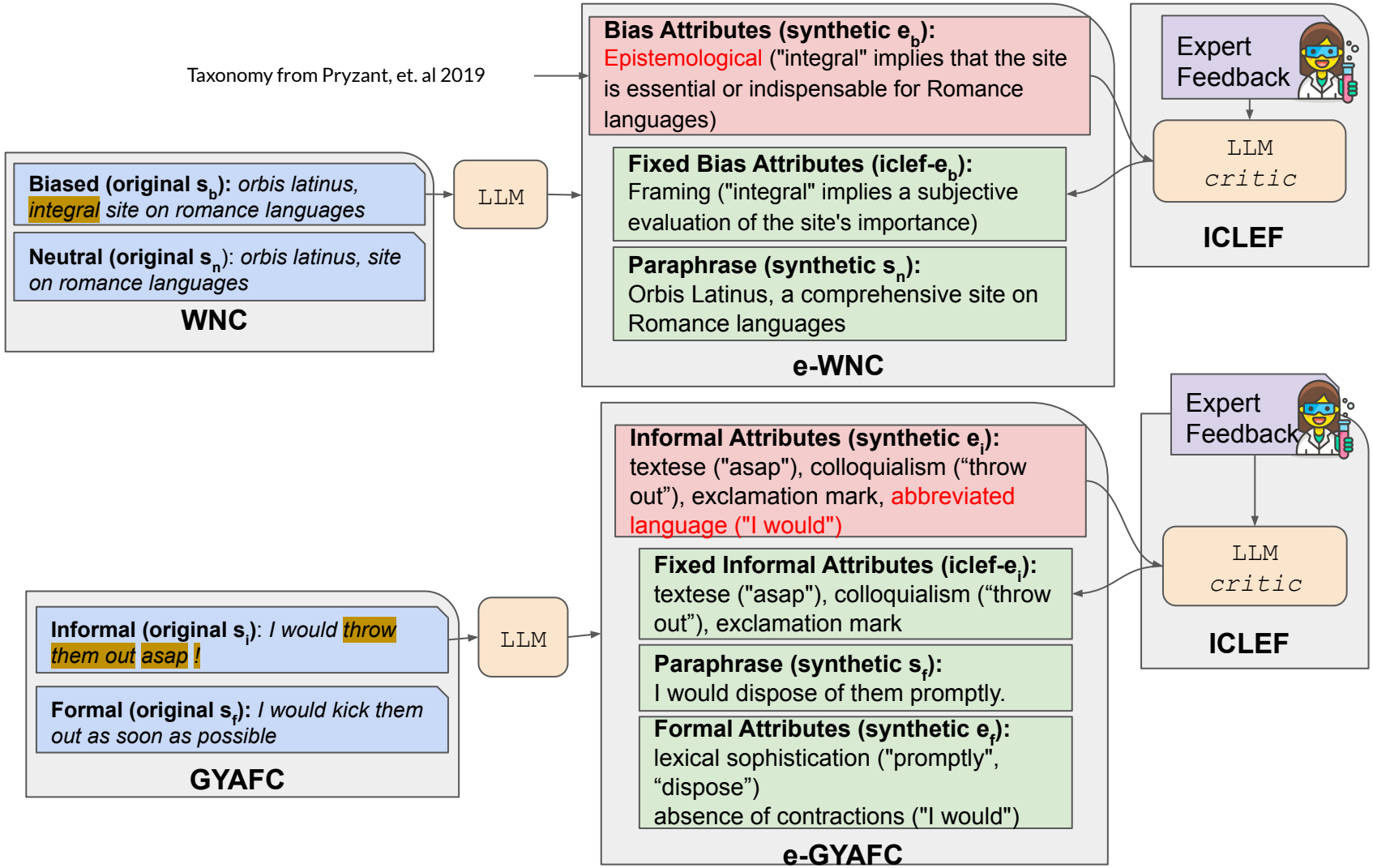
- Human-AI collaboration approach to *model distillation* (**ICLEF)** to create such datasets:
  - 1) distill natural language explanations from large LMs
  - 2) leverage in-context learning and self-critique abilities of LLMs
- Advantage: we can use this dataset to fine-tune ***smaller models for explainable style transfer***

Explainability

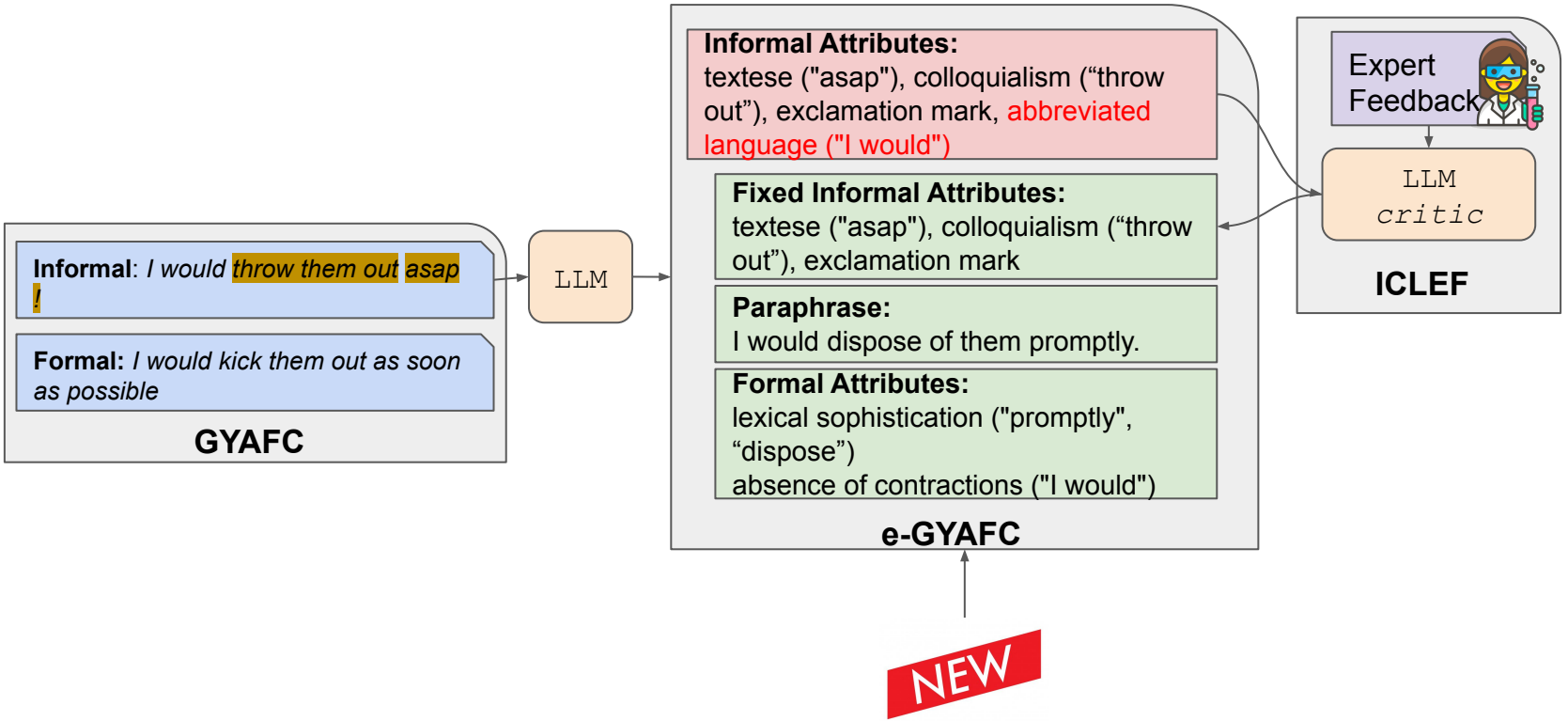
- Generating and Distilling natural language explanations:
  - PINTO: Faithful Language Reasoning Using Prompt-Generated Rationales (Wang et al., 2023)
  - Large Language Models Are Reasoning Teachers (Ho et al., 2022)
  - Teaching Small Language Models to Reason (Magister et al., 2023)
- Self-critique
  - Self-Refine: Iterative Refinement with Self-Feedback (Madaan et al., 2023)
  - Constitutional AI: Harmlessness from AI Feedback (Bai et al., 2022)
  - Self-critiquing models for assisting human evaluators (Saunders et al., 2022)
  - Training Language Models with Language Feedback at Scale (Scheurer et al., 2023)

Learning from human feedback

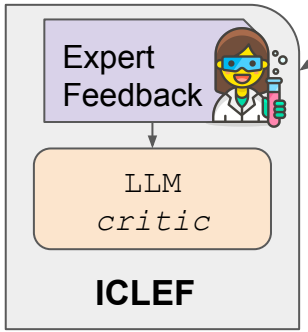
- Generated explanations might still have mistakes. How to leverage feedback to improve quality?
  - RLHF (Ouyang, 2022)
  - Chain-of-Hindsight (Liu et al.,2023),
  - Sequence Likelihood Calibration (Zhao et al., 2023),
  - Direct Preference Optimization (Rafailov et al., 2023)
- All work with large preference corpora from crowdworkers, but what to do with scarce expert feedback?


Taxonomy from Pryzant, et. al 2019
Bias Attributes (synthetic e_b):
Epistemological ("integral" implies that the site is essential or indispensable for Romance languages)
Biased (original s_b): orbis latinus, integral site on romance languages
Neutral (original s_n): orbis latinus, site on romance languages
WNC
LLM
Fixed Bias Attributes (iclef-e_b):
Framing ("integral" implies a subjective evaluation of the site's importance)
Paraphrase (synthetic s_n):
Orbis Latinus, a comprehensive site on Romance languages
e-WNC
Expert Feedback
LLM critic
ICLEF
Informal Attributes (synthetic e_i):
textese ("asap"), colloquialism (“throw out”), exclamation mark, abbreviated language ("I would")
Informal (original s_i): I would throw them out asap !
Formal (original s_f): I would kick them out as soon as possible
GYAFC
LLM
Fixed Informal Attributes (iclef-e_i):
textese ("asap"), colloquialism (“throw out”), exclamation mark
Paraphrase (synthetic s_f):
I would dispose of them promptly.
Formal Attributes (synthetic e_f):
lexical sophistication ("promptly", “dispose”)
absence of contractions ("I would")
e-GYAFC
Expert Feedback
LLM critic
ICLEF

Example for formality style transfer



Incorporating expert feedback



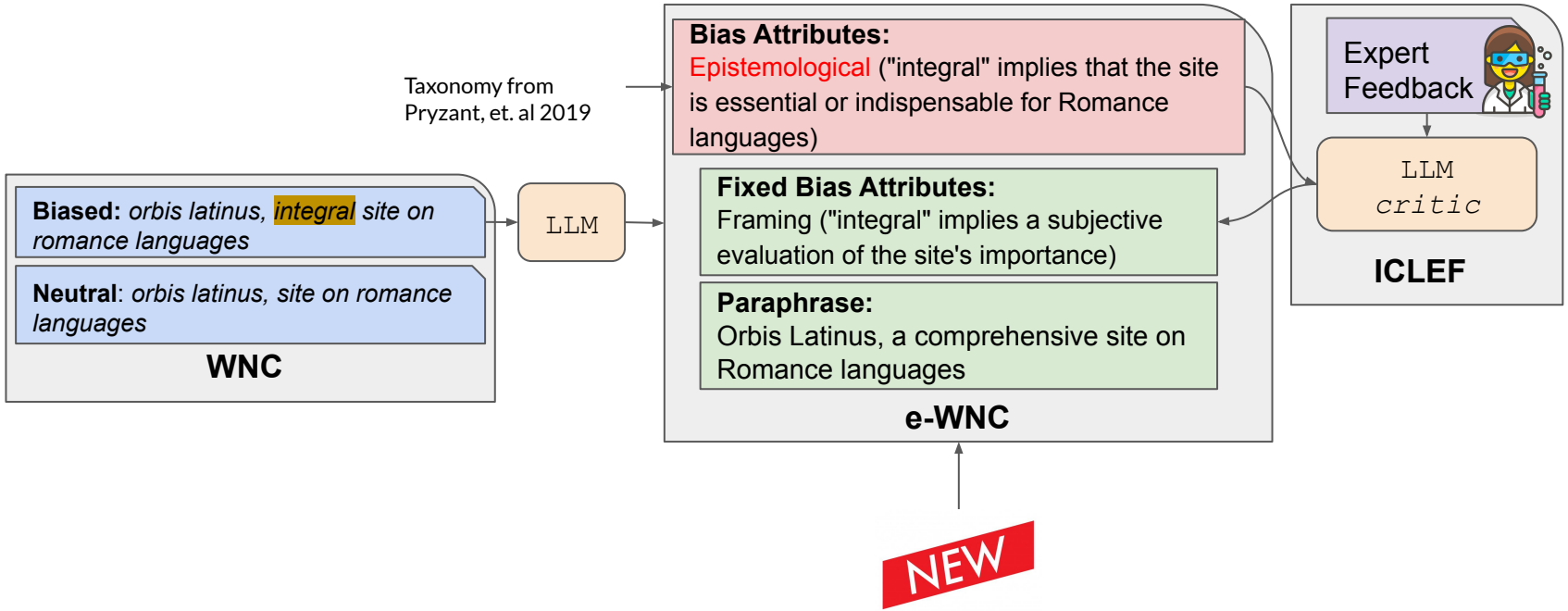
Few-shot prompt with demonstrations of expert annotator corrections:

e-GYAFC:

…

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)
…

Similarly for *biased -> unbiased* style transfer



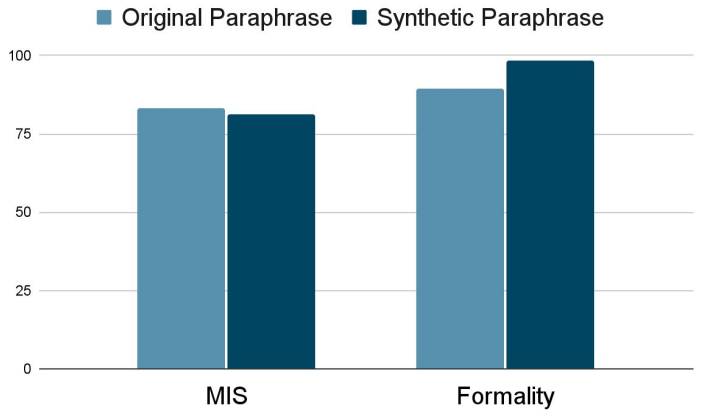
Examples of ICLEF editing:

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.

Dataset Quality

- Automatic Evaluation
  - **MIS:** semantic accuracy of the paraphrase
  - **Formality**: style accuracy



- Human Evaluation
  - **Acceptability:** does the explanation list correct attributes and does the paraphrase have the same meaning?
  - **Preference:** is the iclef-explanation preferred to the original one? Is the synthetic paraphrase preferred to the original one?

	Acceptability	Preference
Informal explanation	87	90
Formal paraphrase	77	77
Formal explanation	98	N/A

Evaluation on the Explainable Style Transfer Task

Teacher models:

- GPT models

Student models:

- LLaMA-7B
- Alpaca-7B

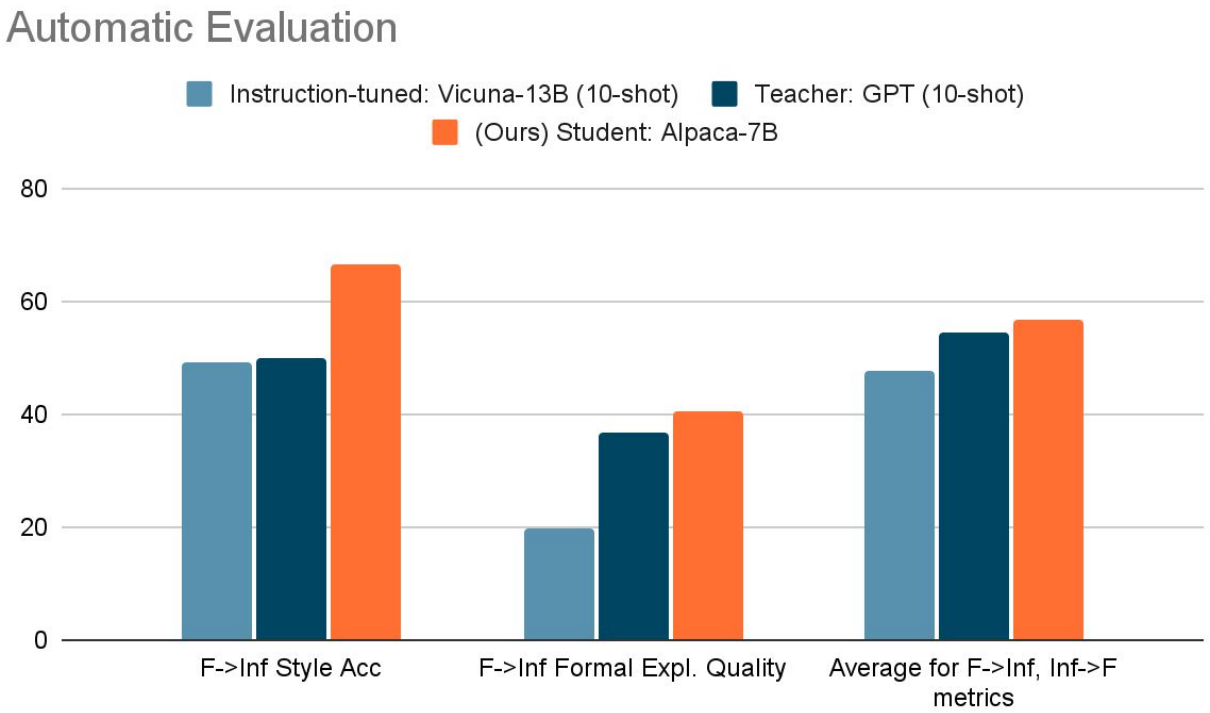
Other instruction-tuned models

- Vicuna-13B
- More baselines in the paper

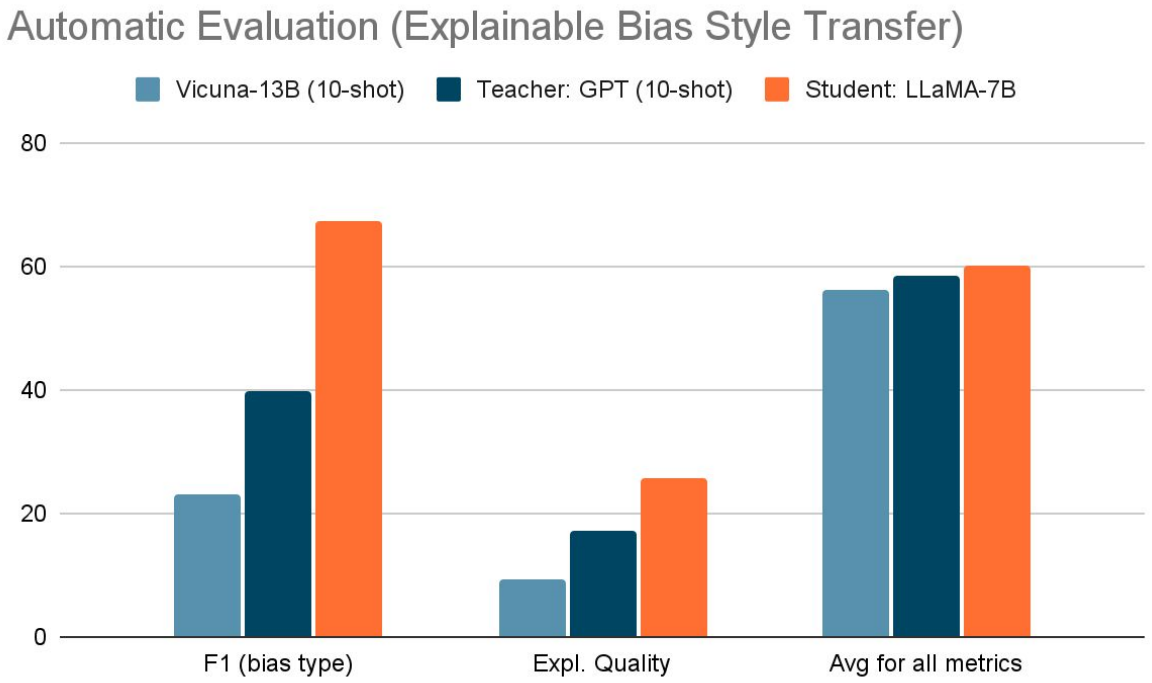
Settings:

- One-shot
- Few-shot with ICLEF-improved data
  - 5-shot
  - 10-shot

Results: Formality Explainable Style Transfer



Results: Bias Style Transfer

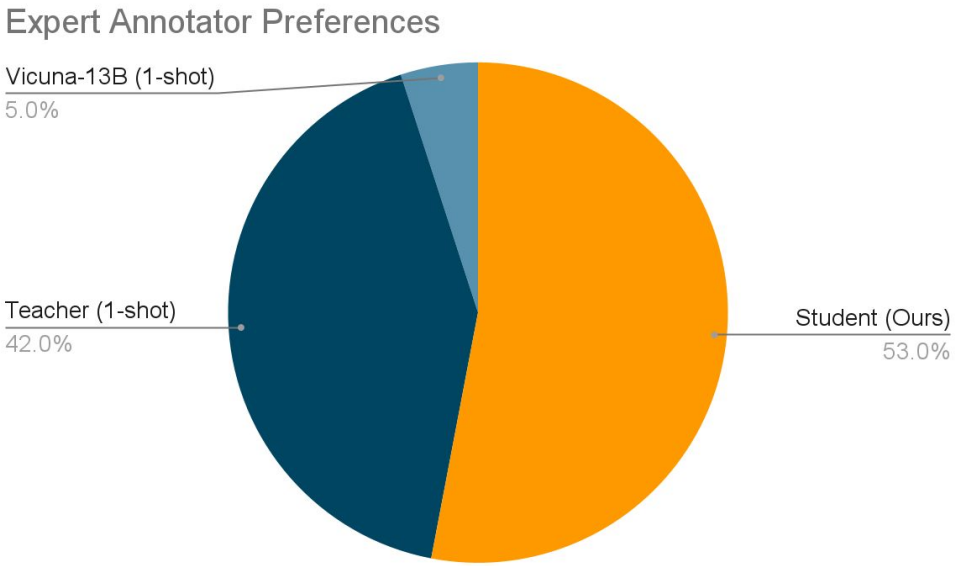


Human Evaluation

Formal (Input)
If it is feasible, allow love to prevail.

Teacher (1-shot)
`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")

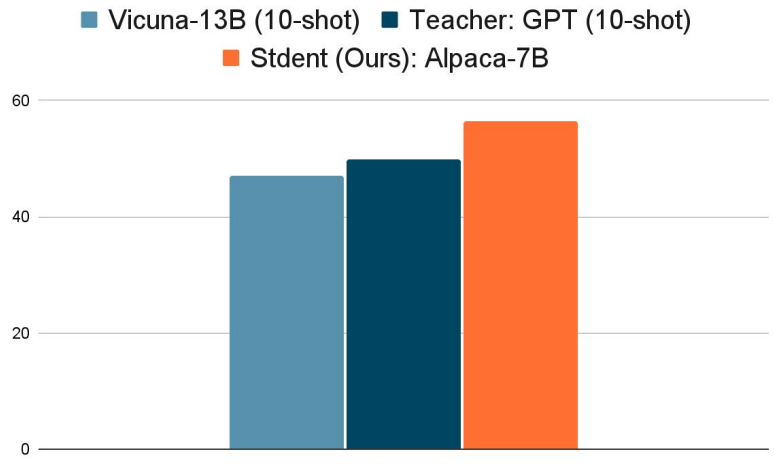
Student (Ours):
`Formal Attributes:`
precise language ("feasible"), use of conditional clause ("if it is feasible"), lexical sophistication ("prevail")
`Informal Paraphrase:`
let love win if it's possible.
`Informal Attributes:`
contraction ("it's"), casual tone ("let love win")



Extrinsic Evaluation: Authorship Verification

- **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”*



Bonus: Interpretable Adversarial Attacks on AI-generated text detection

- How well can informal paraphrase be detected as AI-generated?
  - GPT-F: Formal (AI-generated)
  - GPT-Inf: ChatGPT informal paraphrase
  - Our model