Including past attempts as negative examples and feedback in context helps language models learn personalized style

Tuning-Free Personalized Alignment via Trial-Error-Explain In-Context Learning

Hyundong Cho, Karishma Sharma, Nicolaas Jedema, Leonardo F. R. Ribeiro, Alessandro Moschitti, Ravi Krishnan, Jonathan May

Personalized Text Generation

📌 ***Task:*** *Write an email response to the following: Recently, school officials prevented a school shooting because one of the shooters posted a myspace bulletin. Do you think this was an invasion of privacy?*

✅ **Personalized (Author's text)**
... I feel as though the school officials were not invading privacy at all. The entire point of a myspace bulletin is for people - the public - to see. ... The school was doing its job ... Did the school harm the person who was going to do the shooting? No. ... since it IS available to the public- to TURN OVER ... Where should we draw the line for privacy though? ... If there are lives in danger, ... Thanks for your time.

❌ **Generic (LM + Vanilla In-Context Learning)**
... In regards to the question of whether the school officials' actions ...
I hold the firm belief that they were not.
The fundamental duty of any educational institution ...
Ultimately, ...
Hence, ...
Thank you for your inquiry.

Realistic Constraints of Personalization

Low-resource data: limited amount of per-user data
→ <10 samples per user

Private data
→ Assume access to only data from one user at a time

Fine-tuning is infeasible
→ Overwhelming overhead from per-user weights
→ Not possible for black box models, which often perform best

→ RQ: How to personalize language models for text generation given these constraints?

Naïve Solution

Vanilla In-Context Learning (ICL)

You are a stylistically consistent writer. Below are examples that exemplify your writing style.

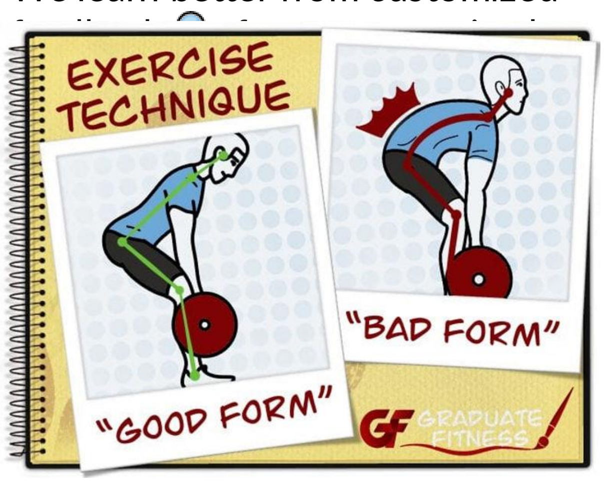
📌 Task 1 | ✅ Positive example
...
📌 Task N | ✅ Positive example

Now perform this task:

📌 Task X

✅ Positive examples are not enough for overcoming bias for generic style of formal and structural phrases!

💡 **Motivation**
We learn better from customized



Trial-Error-Explain In-Context Learning (TICL)

TICL

You are a stylistically consistent writer. Below are examples that exemplify your writing style.

📌 Task 1 | ✅ | ⛔ | 🔍 | ⛔ | 🔍
...
📌 Task N | ✅ | ⛔ | 🔍 | ⛔ | 🔍

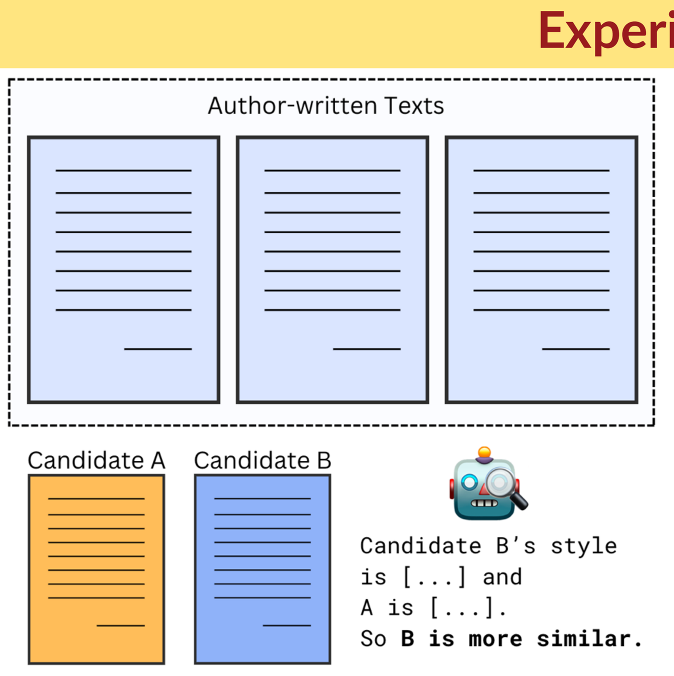
Now perform this task:

📌 Task X

High-level Algorithm for Developing TICL Prompt

1. Start with Vanilla ICL prompt that includes positive examples ✅ (but with N-1 examples)
2. Generate output for Nth example (Trial)
3. Validate whether output is similar in style to author examples, while providing an explanation 🔍
4. If not, include output as negative example ⛔ (Error) and its explanation 🔍 (Explain)
5. Repeat steps 2-4
6. Every K iterations of steps 2-5, evaluate on validation set to track best performing prompt

Experimental Setup



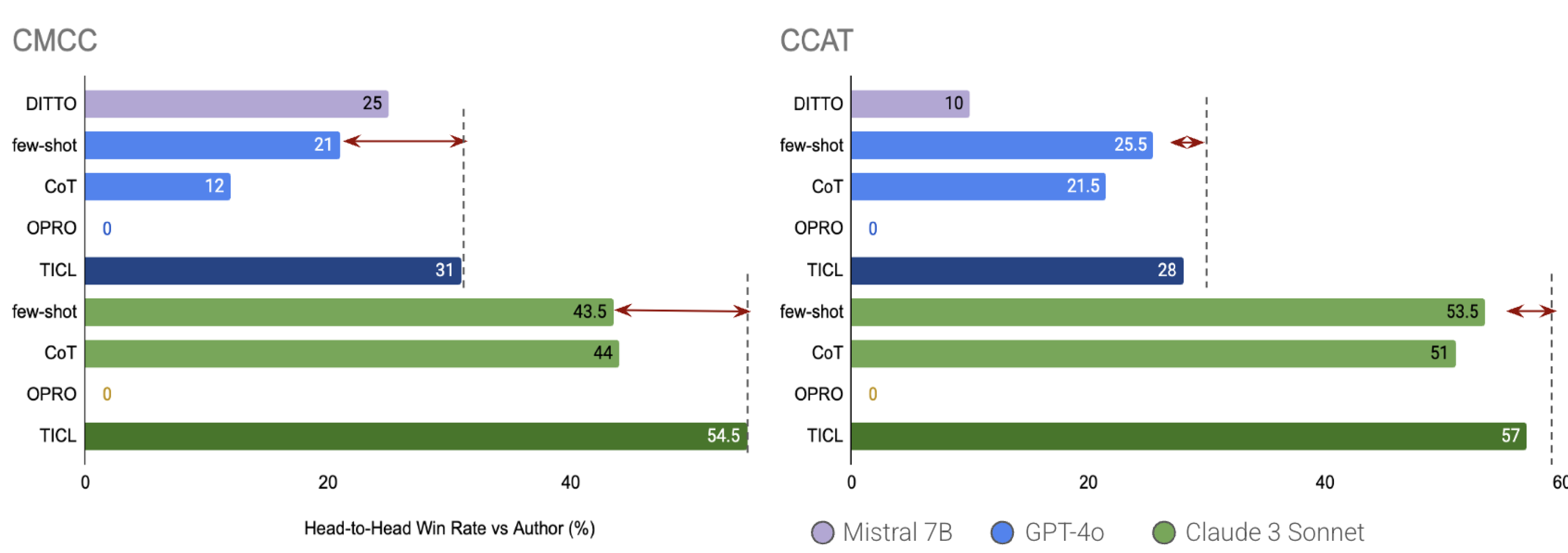
Evaluation method
- Head-head comparison on stylistic similarity against author text with LM-as-a-Judge
- >96% accuracy using human data

Dataset
- CMCC: Student-written essays/emails
- CCAT: News articles
- Sample size: 7/2/3 for train/validation/test
- 10 authors from each dataset

Models
- GPT-4o (`gpt-4o-2024-0806`)
- Claude 3 Sonnet

Baselines
- DITTO: Previous fine-tuned SOTA (Shaikh et al., 2024)
- Few-shot: Vanilla ICL
- CoT: Chain-of-Thought (Wei et al., 2022)
- OPRO: Prompt optimization (Yang et al., 2024)

Main Results



TICL outperforms all baselines, including test-time compute scaling methods (CoT, OPRO) and previous SOTA (DITTO)

TICL Ablation

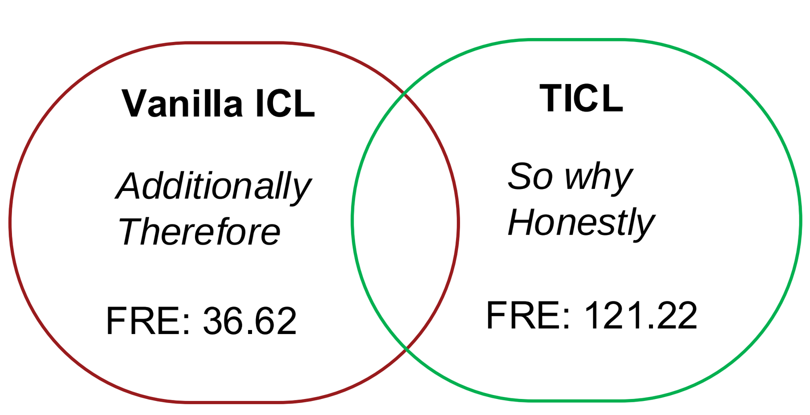
Model	Ablation	Win rate
GPT-4o	TICL	$\mathbf{31.00}_{1.18}$
	– Initial ICL examples	$28.50_{1.50}$
	– Explanations	$23.50_{1.24}$
	– Checkpointing	$22.50_{0.90}$
	→ with Claude 3 S expl.	$32.50_{1.13}$
	→ with Claude 3 S TICL	$18.30_{0.25}$
Claude 3 S	TICL	$\mathbf{54.50}_{1.67}$
	– Initial ICL examples	$52.00_{1.22}$
	– Explanations	$46.00_{1.41}$
	– Checkpointing	$54.00_{1.67}$
	→ with GPT-4o expl.	$55.00_{1.28}$
	→ with GPT-4o TICL	$42.50_{0.92}$

- Explanations are important for TICL performance
- But they don't have to be from the same model
- The negative examples, however, need to be from the same model! i.e., models need to see their own failures.

Qualitative/Lexical Analysis of TICL > ICL

❌ **LM + Vanilla ICL → Generic**
... In regards to the question of whether the school officials' actions ... I hold the firm belief that they were not. The fundamental duty of any educational institution ... Ultimately, ... Hence, ... Thank you for your inquiry.

✅ **LM + TICL → Personalized**
... I genuinely believe that ... It's simple - ... If someone posts about harmful intentions, expecting privacy is a bit ironic, don't you think? The school officials did their job, ... Privacy is essential, yes, ... If a message was genuinely private, ... So, let's focus ...
Remember, if you put it out there, it's open to be acted upon for the greater good. Thanks for hearing me out on this matter!



Fightin' Words model (Monroe et al., 2008)
- Surfaces significant frequency differences of n-grams between two distributions
- N-grams more frequent to TICL: casual and opinionated phrases, higher FRE
- N-grams more frequent to ICL: formal and structural phrases, lower FRE
- FRE: Flesch Readability Ease, ↑ = easier to read.

Summary

- Self-generated negative examples and their corresponding explanations are effective even in low-resource settings for personalized text generation!
- TICL is effective in helping models overcome bias for formal and structural language and instead adopt the opinionated and casual phrases from user examples
- LMs need to be shown their own negative outputs in their context for TICL to be effective!

USC INFORMATION SCIENCES INSTITUTE

CUTE LAB NAME

amazon

Contact: hd.justincho@gmail.com
Project page: justin-cho.com/ticl