Unsupervised Human Preference Learning

Sumuk Shashidhar, Abhinav Chinta, Vaibhav Sahai, Dilek Hakkani-Tür

Department of Computer Science, University of Illinois Urbana Champaign

EMNLP 2024

CONVAI

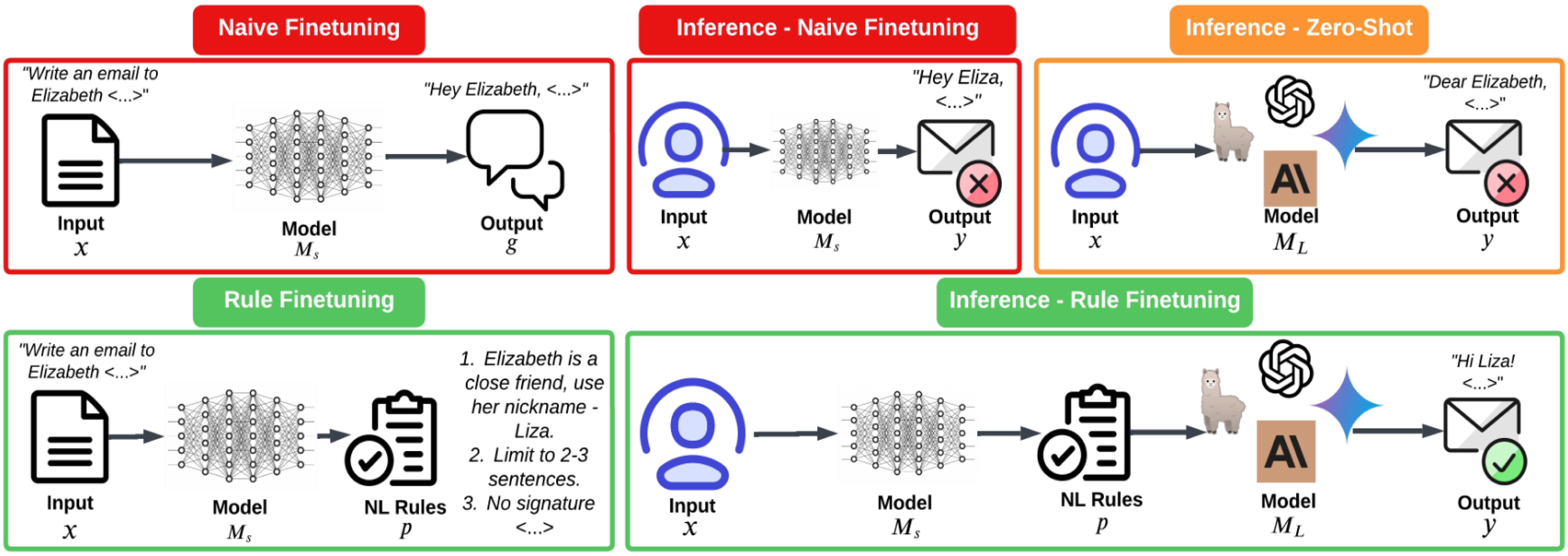
AICE Center

Amazon-Illinois Center on AI for Interactive Conversational Experiences

How Can We Efficiently Personalize Language Model Outputs?

Key Takeaways

- Preference Agents capture individual user preferences without relying on explicit human feedback or labeled data
- Small, locally trainable agents guide large LLMs, enabling cost-effective personalization on limited user data
- Outputs generated by homogenous model pairs show superior alignment compared to heterogenous model pairs



TL;DR

Problems:

- Powerful LMs produce impersonal outputs.
- Finetuning big LMs is too expensive at scale. Even LoRA!

Opportunity: People have personal preferences that can be learnt from!

Our Solution: LLM Personalization via lexicalized rules:

- Train a small, local model on personalization rules.
- Use it to guide foundational models in tasks, aligning with user preferences.

Empirical Examples

Write an email to Sylvia informing her that the NG contract is complete and ready for legal review, and that it should be available by tomorrow morning.

Dear Sylvia,
Thank you for your email regarding the status of the NG Energy contract. I'm pleased to let you know that the contract is now complete and ready for an attorney review. [...]
Best regards,
<Your Name Here>

Write an email to Sylvia informing her that the NG contract is complete and ready for legal review, and that it should be available by tomorrow morning.

Email guidelines:
- Start with talking about NG Energy Contract
- Don't use an introductory name
- Use direct tone, concise sentences

The NG Energy contract is finished. I need a lawyer to review it. It will be ready in the morning.

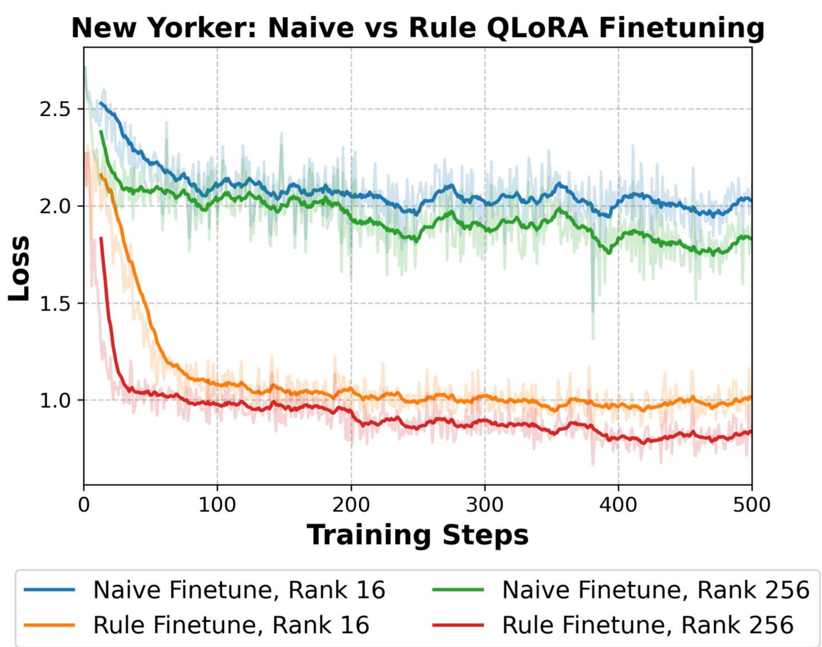
Results

Preference Agents	New Yorker			Enron			LAMP 3U			Aggregated	
$M_L \rightarrow$ vs $\downarrow$	Llama3 70B Instruct	Claude 3.5 Sonnet	Gemini 1.5 Pro	Llama3 70B Instruct	Claude 3.5 Sonnet	Gemini 1.5 Pro	Llama3 70B Instruct	Claude 3.5 Sonnet	Gemini 1.5 Pro	LLM Evaluation	Human Evaluation
Small Baseline	77.4	91.5	80.0	88.4	96.1	89.8	74.6	84.0	75.3	84.1	91.0
Large Baseline	67.7	75.2	66.9	85.6	83.7	88.2	66.5	69.5	63.8	74.1	84.5
Few Shot	68.3	62.0	66.7	61.1	68.0	57.4	58.3	57.4	59.4	62.0	73.4
Naive Finetune	80.3	82.4	81.8	75.3	87.8	81.3	85.0	92.7	89.0	83.9	92.2
No Baseline Agent	65.1	68.8	63.8	58.4	61.3	62.5	63.8	67.2	60.4	63.4	52.0

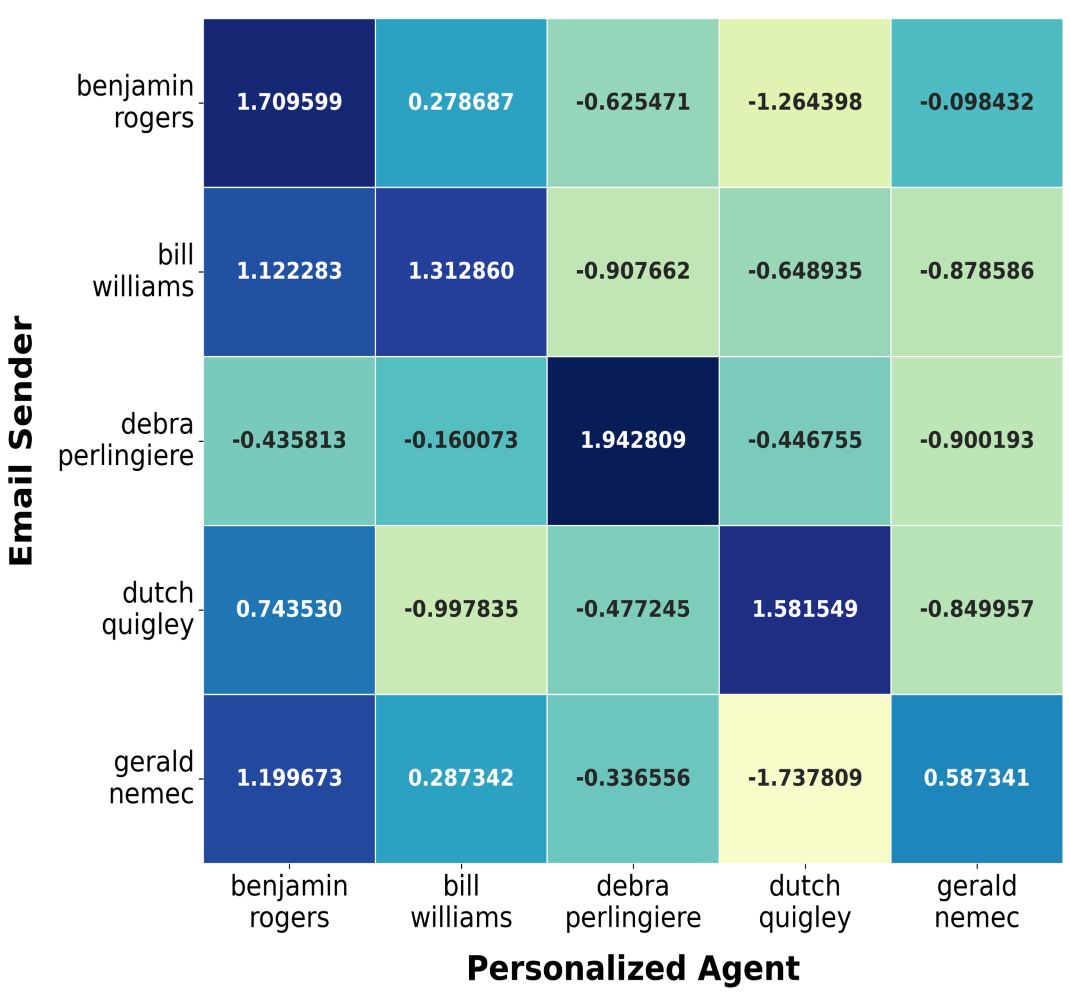
Table 2: **Win Rates** of Llama3 8B M_s combined with various M_L, evaluated by GPT4o and human evaluation.

PeFT on User Data vs Rules

- Ground truth content is diverse and varies significantly based on task context.
  - It also varies significantly across different tasks
- Preference rules have repeating patterns and structures making it easier for instruction finetuned models to learn them from limited demonstrations



Ablation Study



Personalization Test:

- Top 5 senders from Enron corpus
- Distilled and trained preference agents for each sender
- Tested each agent on all senders in a permutation analysis
- Evaluated using BertScore

Results:

- Strong diagonal trend
- Agents perform best on the sender they were trained on