

Unsupervised Human Preference Learning

Anonymous ACL submission

Abstract

Large language models demonstrate impressive reasoning abilities but struggle to provide personalized content due to their lack of individual user preference information. Existing methods, such as in-context learning and parameter-efficient fine-tuning, fall short in capturing the complexity of human preferences, especially given the small, personal datasets individuals possess. In this paper, we propose a novel approach utilizing small parameter models as preference agents to generate natural language rules that guide a larger, pre-trained model, enabling efficient personalization. Our method involves a small, local "steering wheel" model that directs the outputs of a much larger foundation model, producing content tailored to an individual's preferences while leveraging the extensive knowledge and capabilities of the large model. Importantly, this personalization is achieved without the need to fine-tune the large model. Experimental results on email and article datasets, demonstrate that our technique significantly outperforms baseline personalization methods. By allowing foundation models to adapt to individual preferences in a data- and compute-efficient manner, our approach paves the way for highly personalized language model applications.

1 Introduction

Large language models like ChatGPT have demonstrated impressive reasoning and generalization skills across various tasks using Zero Shot and Few Shot methods (Kojima et al., 2022). However, their ability to provide personalized content remains limited (Woźniak et al., 2024). These models are trained on large, general-purpose datasets and fine-tuned to cater to a broad audience, necessitating a neutral and unbiased approach. As a result, when performing tasks such as writing emails, messages, or blog posts, the outputs generated by these models tend to be generic and lack the unique touch

that resonates with individual users. The inherent diversity and often contradictory nature of human preferences (Berliner et al., 2016) make it challenging for large language models to capture the nuances of individual styles while simultaneously attempting to cater to a large group of users.

Methods like in-context learning (ICL) (Brown et al., 2020) have demonstrated the effectiveness of providing few-shot examples to enhance model performance on specific tasks. However, when dealing with human preferences, providing few-shot examples in context is insufficient to capture the complexity and nuances of these preferences (Peng et al., 2023). Given that preferences are stochastic, the model can only apply the information from the given few shot examples, without being able to leverage the complete preference information of the user. Recently, fine-tuning has emerged as the most effective approach for enabling models to learn specific tasks. While full supervised fine-tuning is resource-intensive, Parameter Efficient Finetuning (PEFT) methods like LoRA (Hu et al., 2021) and QLoRA (Dettmers et al., 2023) offer a more resource-effective solution for task-specific learning. However, in the domain of human preference learning, PEFT methods such as QLoRA fail to generalize, especially given the small datasets that individual users possess (Balne et al., 2024).

We propose a novel approach for aligning large language models towards personalized user preferences using preference agents. These preference agents are small, locally inferrable, fine-tuned language models that generate natural language rules to guide the behavior of a larger, generic, pre-trained model. By leveraging the knowledge and superior capabilities of the large, generic model while injecting user-specific rules, our method enables efficient personalization without the need for expensive retraining or invasive collection of large human feedback datasets. The preference agent, given a particular task, distills an individual user's

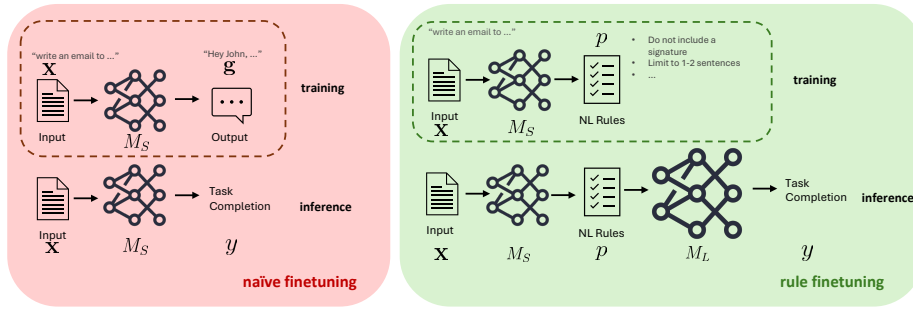


Figure 1: Naive vs Preference Rule Finetuning

083 preferences into a concise set of instructions that
 084 the large, generic model follows to produce tai-
 085 lored outputs aligned with the user’s unique re-
 086 quirements. This modular architecture decouples
 087 preference learning from the generic pre-trained
 088 model, which allows users to finetune small mod-
 089 els locally.

090 Our method of preference distillation represents
 091 a significant departure from conventional training
 092 approaches, offering a new solution for unsuper-
 093 vised human preference learning. We evaluate
 094 our approach across two human-generated content
 095 datasets and task settings, showing that preference-
 096 guided language models significantly outperform
 097 both fine-tuning baselines and standard prompting
 098 techniques based on automatic metrics, GPT-4 eval-
 099 uations, and human judgments.

100 Our main contributions are as follows:

- 101 • We propose a new fine-tuning objective that
 102 utilizes distilled target information instead of
 103 traditional input-output pairs. This approach
 104 directly enhances learning efficiency by focus-
 105 ing on essential patterns, such as preference
 106 information, without relying on implicit recog-
 107 nition from broader data.
- 108 • We show that compared to prompting with
 109 few-shot examples and fine-tuning as base-
 110 lines, the use of rule generators with a large
 111 model results in a performance boost of up to
 112 80% for various tasks involving human prefer-
 113 ences.
- 114 • We release two large, human intent annotated
 115 preference datasets, to enable future research
 116 on preference learning techniques and opti-
 117 mizations.

118 2 Method

119 In this section, we detail our approach for align-
 120 ing language models to personalized user prefer-

121 ences using small preference agents. Our method
 122 involves two key components: generating natural
 123 language rules that capture user preferences and
 124 utilizing these rules to guide a larger, pre-trained
 125 language model. This modular architecture allows
 126 for efficient personalization without extensive re-
 127 training.

128 2.1 Task Definition

129 Given a task T , we define the dataset \mathcal{D} as consist-
 130 ing of input-output pairs. Each input comprises a
 131 user intent \mathbf{u} and associated task metadata \mathbf{m} , and
 132 the output is the ideal task completion, denoted as
 133 \mathbf{g} , which we consider the ground truth. Thus, the
 134 dataset can be formally expressed as:

$$135 \mathcal{D} = \{(\mathbf{x}, \mathbf{g}) \mid \mathbf{x} = (\mathbf{u}, \mathbf{m})\}$$

136 2.2 Constraints and Assumptions

137 We seek to enable users to generate high qual-
 138 ity, personalized responses as our goal, which are
 139 bounded by some constraints and assumptions:

140 Firstly, the size of the dataset \mathcal{D} is not large
 141 enough to permit effective full model fine-tuning.
 142 Given that individual users typically possess small,
 143 personal datasets, it is impractical to expect these
 144 datasets to be sufficient for extensive fine-tuning of
 145 a large language model.

146 Secondly, the small model, denoted as M_S ,
 147 must be lightweight enough to operate on end-
 148 user devices, such as laptops, phones, and tablets.
 149 This requirement ensures that users can generate
 150 and apply their preferences without the need for
 151 high-performance computing resources. The small
 152 model’s efficiency allows for local inference, mak-
 153 ing the personalization process more accessible and
 154 convenient.

155 Thirdly, we wish to use an alignment process,
 156 that can be completed without the use of major
 157 additional hardware

158 Lastly, we assume that the large model, referred
 159 to as M_L , is either too large to run inference locally
 160 or is a closed-source API model. Consequently, it
 161 is not feasible, or cost effective to fine-tune or align
 162 M_L by altering its model weights.

163 2.3 Model Training

164 Given the dataset \mathcal{D} , we first task M_L with generat-
 165 ing zero-shot responses to our training data. These
 166 initial responses are devoid of any user-specific
 167 preference information:

$$168 \mathbf{Y}_z = M_L(\mathbf{X}) \quad (1)$$

169 where \mathbf{Y}_z represents the set of zero-shot outputs
 170 for all inputs \mathbf{X} in the training dataset.

171 Next, we leverage M_L 's capabilities to extract
 172 the delta between the zero-shot completions (\mathbf{Y}_z)
 173 and the ground truth outputs (\mathbf{G}). This delta repre-
 174 sents the preference rules that need to be learned
 175 by the smaller model:

$$176 \mathbf{P} = M_L(\mathbf{Y}_z, \mathbf{G}) \quad (2)$$

177 Here, \mathbf{P} represents the set of preference rules
 178 derived for each training example. We hypothesize
 179 that M_L can effectively identify these rules without
 180 prior knowledge of the specific user's preferences,
 181 just by observing the differences between the zero
 182 shot completion and the ground truth.

183 Finally, we train the smaller model, M_S , to learn
 184 to generate these preference rules. The training
 185 data for M_S consists of input-preference rule pairs:

$$186 M_S \xrightarrow{(\mathbf{x}, \mathbf{P})} M_A \quad (3)$$

187 Through this training process, M_S learns to map
 188 user intents and task metadata to natural language
 189 preference rules, effectively becoming a personal-
 190 ized preference agent (M_A).

191 2.4 Model Alignment

192 Once the preference agent M_A is trained, we can
 193 use it to align the larger model's outputs to unseen
 194 user data. For a new input \mathbf{x} , we first generate
 195 preference rules using the trained agent:

$$196 \mathbf{p} = M_A(\mathbf{x}) \quad (4)$$

197 These rules, expressed in natural language, are
 198 then provided as additional context to the large
 199 language model M_L alongside the original input:

$$200 y_a = M_L(\mathbf{x}, \mathbf{p}) \quad (5)$$

201 The output y_a is considered to be preference-
 202 aligned as it is generated by M_L while considering
 203 the user's preferences encoded in \mathbf{p} . This approach
 204 allows us to leverage the vast knowledge and gener-
 205 ative capabilities of M_L while tailoring the output
 206 to individual preferences without directly modify-
 207 ing the large model's weights.

208 2.5 Quantifying Alignment

209 To evaluate the effectiveness of our preference
 210 alignment method, we employ an evaluation func-
 211 tion on an unseen test set \mathcal{T} . For each example in
 212 \mathcal{T} , the evaluation function considers three pieces of
 213 information: the original input \mathbf{x} , the zero-shot out-
 214 put generated by the large model ($y_z \in \mathbf{Y}_z$), and
 215 the preference-aligned output generated by incorpo-
 216 rating the preference agent's guidance ($y_a \in \mathbf{Y}_a$).

217 The evaluation function, denoted as
 218 $\text{Eval}(y_a, y_z | \mathbf{x})$, assesses which of the two
 219 outputs, y_z and y_a , better aligns with the user's
 220 likely preference, given the input \mathbf{x} . While the
 221 specific implementation of Eval can vary (e.g.,
 222 human evaluation, model-based metrics), its output
 223 is a score indicating the preference between the
 224 two outputs:

225 A positive score indicates a preference for the
 226 aligned output (y_a). A negative score indicates a
 227 preference for the zero-shot output (y_z). We aggre-
 228 gate these scores across all examples in the test set
 229 \mathcal{T} to obtain an overall alignment score:

$$230 \text{Score}(\mathcal{T}) = \sum_{i=1}^{|\mathcal{T}|} \text{Eval}(y_a^{(i)}, y_z^{(i)} | \mathbf{x}^{(i)}) \quad (6)$$

231 where:

- 232 • $|\mathcal{T}|$ represents the number of examples in the
 233 test set.
- 234 • $y_a^{(i)}$ and $y_z^{(i)}$ represent the aligned and zero-
 235 shot outputs, respectively, for the i -th exam-
 236 ple.

237 A positive $\text{Score}(\mathcal{T})$ suggests that the preference
 238 agent successfully guides the large language model
 239 to generate outputs that are better aligned with user
 240 preferences compared to the baseline zero-shot out-
 241 puts.

242 3 Experimental Setup

243 3.1 Model Choice

244 We select Llama 3 with 8 billion parameters (8B)
 245 as our M_S and Llama 3 with 70 billion parameters

(70B) as our M_L (AI@Meta, 2024). The motivation behind these choices is twofold: the 70B version of Llama 3 is known for its exceptional capabilities, making it a robust foundation model, while the 8B version is sufficiently powerful and can be deployed on end-user devices. Additionally, the 8B model can be fine-tuned using QLoRA¹ within the constraints of 16GB of VRAM, making it an ideal candidate for serving as an alignment agent in our personalization framework.

3.2 Datasets

Our evaluation spans two datasets, each encompassing single and multi-user preference information to demonstrate the robustness and generalizability of our framework.

For evaluating the performance of short form writing, we select the Enron email corpus (Klimt and Yang, 2004). This corpus comprises emails from approximately 150 users, predominantly senior management at Enron, structured into folders. The corpus includes roughly 0.5 million messages in total. We sample 15 users from the Enron corpus, for our preference alignment test in order to analyze the reproducibility of an individual’s writing style. We split each user’s subset into an 80-20 train-test split.

The second dataset is a subset of the All the News 2.0 dataset (Thompson, 2020), specifically articles from The New Yorker magazine, which contains approximately 3,500 articles. This subset was selected due to the abundance of creative writing within The New Yorker magazine, which provides a rich source of author preference information. We seek to analyze whether, with preference agents, the unique style of the New Yorker, can be reproduced with simple, natural language rules. We split this dataset into a 50-50 train test split.²

Refer to Appendix B.1 for details regarding dataset preparation and sampling.

3.2.1 Dataset Augmentation

Synthetic Intent Generation. We aim to develop a fully unsupervised approach that scales effectively by avoiding the manual collection of human intents. Instead, we make the model extract the core content of the text into bullet points to emulate user input. We randomly sample these generated intents and subject them to manual human evaluation. Our

findings indicate a high degree of fidelity, with over 95% of the synthetic intents achieving agreement with intents written by humans. These intents are then utilized as inputs for our model, ensuring that the training process remains robust and scalable without the need for extensive manual data collection. To control for noise, we generate three intent variants for each count of data, at temperatures of 0.7, 1.0 and 1.2 respectively to introduce variance. This helps us simulate different user styles. We then randomly sample these intents, in order to make up intent annotated versions of our dataset. Examples of generated intents can be found in Appendix I.3.

Rule Generation As described in §2.3, we generate baselines, which are often extremely formal and verbose, and then subsequently generate natural language preference rules from Llama-3 70B (M_L). Examples of these generated rules can be found in Appendix I.1. As ablations, in addition to the method described in §2.3, we generate two additional sets of rules: (a) without the zero shot baseline, where we only input the target email (b) without the "thinking tokens". The merits and demerits of these rules are discussed in §5.5.

3.3 Model Training

We train our rule generators using parameter-efficient finetuning (PEFT) methods. While full finetuning has the potential to yield superior results, we prioritize scalability and the feasibility of local deployment on user devices, leading us to choose PEFT. Specifically, we employ QLoRA with a rank and alpha of 256. This 1-1 mapping simplifies hyperparameter tuning, and while further experimentation could potentially uncover better configurations, our goal is to demonstrate the effectiveness of our method even with straightforward hyperparameter choices. For a fair comparison, we also train baseline models using naive finetuning (directly on input-output pairs) with the same hyperparameters. We ensure that all model training can be accommodated within 16GB of VRAM, making our approach accessible to consumer-grade devices. A detailed analysis of our finetuning procedure can be found in Appendix C.

3.4 Evaluation Metrics

We evaluate our approach on the Enron and New Yorker datasets using automated evaluation with GPT-4 Omni (GPT-4o) (Naismith et al., 2023; Zheng et al., 2023) and Human Evaluation. We

¹<https://unsloth.ai/blog/llama3>

²We choose a smaller train split for the larger New Yorker dataset, as we wish to demonstrate training sample efficiency

| Preference Agents | Dataset | | Aggregated Eval | | Human-GPT Agreement |
|----------------------|------------|-------|-----------------|-------|---------------------|
| | New Yorker | Enron | GPT-4o | Human | |
| vs Small Baseline | 77.4 | 88.4 | 82.9 | 88.7 | 93.5 |
| vs Large Baseline | 67.7 | 85.6 | 76.65 | 87.4 | 87.7 |
| vs Few Shot | 68.3 | 61.1 | 64.7 | 84.2 | 76.8 |
| vs Naive Finetune | 80.3 | 75.3 | 77.8 | 86.1 | 90.4 |
| vs No Baseline Agent | 65.1 | 58.4 | 61.75 | 71.7 | 86.1 |

Table 1: Win rates and Aggregated results with **Human Evaluation and Human-GPT Agreement in percentage (%)**.

compare our preference agents, trained with and without baseline rules, against several baselines: zero-shot responses from our small (Llama-3-8B) and large (Llama-3-70B) models, few-shot generations using the large model, and a naive fine-tuned agent.

Our primary metric is win percentage, reflecting how often a method’s output is chosen as the best match to the ground truth based on criteria like style, tone, and overall resemblance. Human Evaluation follows the same criteria. We forgo traditional similarity scores like BLEU and ROUGE as they do not adequately capture the nuances of preference information (see Appendix D for further discussion). Detailed information on the Human Evaluation can be found in Appendix F.

4 Results

As discussed in §3.4, we evaluate the performance of our fine-tuned preference agents against several baselines using GPT-4o. Our baselines include zero-shot generations from both the small model (M_S) and the large model (M_L), few-shot generations using M_L , and a naive fine-tuned agent (M_F). We compare these baselines against two variants of our method: a preference agent trained with zero-shot baseline rules (M_A) and a no-baseline agent trained without using zero-shot information.

For the Enron dataset, we fine-tuned our preference agent on 15 unique senders and report the average of the aggregated results. Figure 2 illustrates the efficacy of our preference agent technique, demonstrating high win rates compared to all baselines. Notably, our agent trained on distilled preference rules significantly outperforms the naive fine-tuned model (M_F) with the same hyperparameters, achieving a win rate of 88.4%. Similarly, on the New Yorker dataset, our method outperforms naive fine-tuning with a win rate of 80.3%. This

consistent outperformance across both datasets, further discussed in §5.4, highlights the effectiveness of our approach in capturing and leveraging user preferences.

Human LLM Agreement. Interestingly, we observe that the human evaluation scores consistently show higher win percentages for our method compared to the GPT-4o evaluations. This discrepancy can be attributed to the fact that human evaluators are better equipped to assess nuanced stylistic elements and evaluate their alignment with user preferences. While GPT-4o demonstrates strong capabilities in evaluating text quality, it may not fully capture the subtleties of human preferences in the same way that human evaluators can.

Despite this difference, we observe a high level of agreement between GPT-4o and human evaluations, with an overall concordance rate of 86.9%. This finding aligns with previous research by Zheng et al. (2023), which reported an approximately 80% agreement rate between human judgments and GPT-4o evaluations. This high level of agreement reinforces the reliability of GPT-4o as an automated evaluation tool for assessing text quality, even when dealing with subjective aspects like user preferences. We discuss qualitative examples and human annotation samples of the results in Appendix I.3, and analyze the results further in Appendix A

5 Discussion

5.1 Model Specific Semantic Understanding

In the context of semantic understanding, our study reveals that different families of models interpret the same words differently. Specifically, rules generated with GPT-4o do not significantly improve performance over baselines for the Llama model, compared to rules generated within the Llama fam-

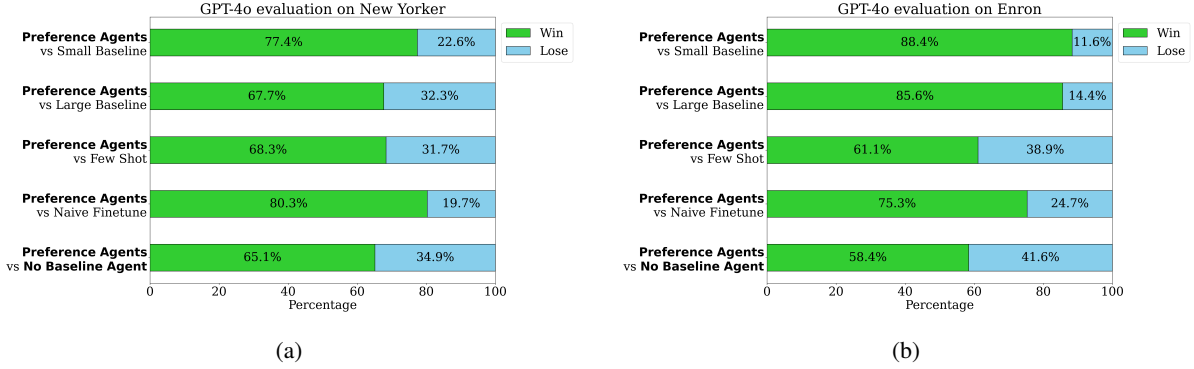


Figure 2: Comparison of win rates on **New Yorker** and **Enron** datasets - GPT4 evaluated

ily. We hypothesize that this discrepancy arises from inherent differences in understanding and reasoning between models. Notably, Llama-3 70B shows a better grasp of rules generated by itself and Llama-8B than those generated by GPT-4o. Despite the well-structured and comprehensive nature of GPT-4o’s rules, they were less effective than those from the Llama family, suggesting that models from the same family have a superior understanding of their responses.

To further investigate, we conducted human tests where rules similar to the ones generated by models were written by 10 expert human annotators and compared against model-specific rules. The results showed that human annotators performed significantly worse, leading to a 16.8% performance degradation on Human evaluations of the end generated content. Upon investigation, we attribute this to a lack of specificity and misunderstandings of vocabulary between humans and models. The model’s interpretation of certain keywords such as "precise," "concise," and "informal" often differs from human perceptions. This leads to the end, generated email, being different from what the human intended. However, when the model generates it’s own rules, this misunderstanding is minimized, leading to superior results. These findings lead us to hypothesize that automated rule generation is superior to manual prompting or rule annotation due to model-specific semantic understanding.

5.2 Thinking tokens

Humans often deliberate before responding to queries, leading to more thoughtful and considered answers. This analogy extends to language models (LLMs), where prompting the model to think and reason before generating a response can enhance the quality of the output. Previous works,

such as (Kojima et al., 2023), have demonstrated that simple prompting, like "Let’s think step by step," can significantly boost performance on various benchmarks. Similarly, (Zelikman et al., 2024; Goyal et al., 2024) have substantiated these findings at the token level. Motivated by these insights, we improve the quality of generated rules by introducing "thinking tokens" into the model’s vocabulary. These tokens provide a cognitive "scratchpad," enabling the model to isolate and process critical preference information more effectively. Our experiments revealed that these thinking tokens significantly enhanced the quality of rule generation by allowing the model to structure its reasoning process.³

5.3 Cost-Effective Fine-Tuning with Alignment Agents

Aligning large language models with user preferences often entails high computational costs, particularly when fine-tuning large models like Llama-3-70B (M_L). Directly fine-tuning M_L ($C_f(M_L)$) is resource-intensive and impractical for consumer-grade hardware. To address this, we propose fine-tuning a smaller Llama-3-8B-Instruct (M_S) model as a preference agent (M_A), trained on input-rule pairs, where rules are derived from M_L . This approach ($C_f(M_S)$) is significantly more cost-effective ($C_f(M_S) \ll C_f(M_L)$).

While naive fine-tuning of M_S (M_F) on input-output pairs is cheaper, our results demonstrate its limitations in capturing complex preferences. Our method, despite a slightly higher combined cost ($C_f(M_S) + C_i(M_L)$, where $C_i(M_L)$ is the negligible inference cost of M_L), significantly outperforms naive fine-tuning.

Furthermore, by not fine-tuning M_L , we retain

³<https://docs.anthropic.com/en/docs/let-claude-think>

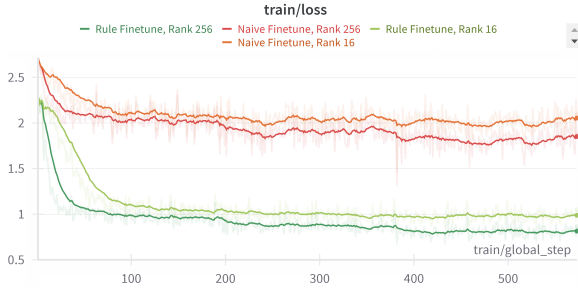


Figure 3: Naive FT vs Rule FT

the flexibility to seamlessly integrate newer, more performant models as they emerge, ensuring our system remains adaptable and future-proof.

5.4 Why Rule Finetuning Is More Effective

In experiments on the New Yorker dataset, we observe that with naive, traditional parameter-efficient fine-tuning (PeFT) using QLora, the loss decreases but does not drop below the 1.5 threshold. However, when fine-tuning on structured rules for the same content while keeping the rank, random seed, and other hyperparameters constant, the loss decreases to below 1.0. This indicates a more effective learning process. We hypothesize that this is due to the size and diversity of the training data. Structured rules have a clear format and structure, enabling the model to identify patterns and learn the process of rule generation more efficiently. In contrast, the inherent complexity and diversity of article writing pose significant challenges for naive fine-tuning methods, as the model cannot easily discern clear patterns or overlaps in the training data. Consequently, our approach demonstrates superior performance from a sample efficiency perspective. The model fine-tuned on structured rules requires a smaller shift in distribution compared to the naive fine-tuning approach, which must adapt completely to the new task. This method is also advantageous for multi-task fine-tuning, as it avoids the need to adapt to entirely different token distributions and task domains. Instead, we focus on learning user preferences and delegate the task completion to the large model, leveraging its generalizability and extensive parameter set, which ultimately leads to superior end reasoning.

5.5 Rule Generation Strategies

Rule generation is essential for effective model personalization, and we employ three methods to achieve this.

The **first** method prompts the large reasoning model (M_L) to generate natural language rules (R_1) to align its responses with user preferences. The **second** method, R_2 , builds on R_1 by incorporating "thinking tokens" (Section 5.2), prompting the model to analyze the input more deeply before generating rules. The **third** method, R_3 , uses a distillation process based on M_L 's zero-shot response. By analyzing this response, the model identifies missing preference information and generates rules to fill these gaps, creating rules that better align the output with user preferences.

The superior performance of R_3 is due to its precise identification and addressing of discrepancies between zero-shot outputs and the ground truth. In comparison, R_1 performs significantly worse due to the lack of in-depth analysis and feedback mechanisms found in R_2 and R_3 .

Overall, our findings highlight that the distillation process in R_3 leads to precise, effective rule generation, and incorporating thinking tokens in R_2 enhances performance compared to the basic approach in R_1 . These strategies are crucial for optimal model personalization.

5.6 Evidence of Personalization

To demonstrate that our approach effectively learns individual writing styles rather than merely approximating the underlying task (e.g., email writing), we conduct a permutation analysis using preference agents trained on different email senders.

We train five preference agents on five distinct email senders from the Enron dataset. We then perform inference using each agent on the test splits of all five senders, generating emails for every combination of agent and sender data. To quantify the similarity between the generated emails and the ground truth, we employ the normalized BERT Score (Reimers and Gurevych, 2019), an automated metric suitable for analyzing large volumes of emails. Additionally, we supplement this analysis with randomly sampled human evaluations to validate our findings.

Our analysis reveals a clear trend along the diagonal of Figure 4, indicating that the model trained on a particular sender's data performs best when tested against the same sender's data. This finding strongly suggests that our approach successfully captures individual writing styles and preferences.

However, this trend does not hold in all cases. Certain models, such as the preference agent trained on benjamin.rogers, achieve higher

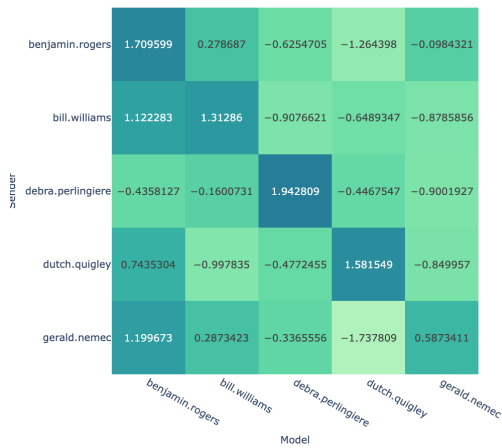


Figure 4: Permutation of Models and Senders

BERT scores across all senders. We hypothesize that this may be attributed to the diversity of Benjamin Rogers’ interactions and the larger size of his training set. Consequently, the model learns to imitate the underlying task extremely well, leading to better performance across all senders. This is evidenced by our training loss, which is the lowest for `benjamin.rogers`. Please refer to Appendix E for further details.

6 Related Work

Traditional Methods of Alignment. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) and Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022) are prominent methods for aligning models with human feedback and fine-tuned LLM annotators, respectively. While effective, RLHF requires costly human annotations and complex distributed training. Direct Preference Optimization (DPO) (Rafailov et al., 2023) improves on this by using preference pairs to train models, reducing computational complexity, but training on contrasting preference pairs does not fully capture the nuances of overlapping human preferences. Furthermore, in-context methods (Kojima et al., 2022; Woźniak et al., 2024) demonstrate limited generalization capabilities due to context length restrictions.

Agent-based Alignment. Training large models is computationally intensive, prompting many to adopt agent-based architectures for compute-restricted environments. Li et al. (2023) employ a fine-tuned T5 policy model to assist large models using stimulus prompting. However, the necessity

for full-parameter SFT and RL optimization introduces computational complexity, yielding limited performance improvements in dialogue response generation. Similarly, Ji et al. (2024) rely on full-parameter SFT and a custom dataset of 50,000 preference examples, demanding rich data and high VRAM. Tan et al. (2024) propose PEFT methods to fine-tune personalized agents based on user history, supplemented with preference retrieval. This method, while computationally efficient, is constrained by the limited reasoning capabilities of the small fine-tuned agent. These works often utilize automatic metrics like BLEU and ROUGE, which capture lexical similarity but fail to encapsulate the nuances of preferences. Gao et al. (2024) introduce an agent trained on human edits to align zero-shot outputs, yet each query necessitates three rounds of inference, increasing latency and computational costs. Moreover, human edit history may not consistently reflect genuine human preference, and measuring it through edit distance proves unreliable. Yang et al. (2024) present a framework for aligning LLMs via Multi-perspective User Preference Ranking-based Feedback, but this approach requires an initial Supervised Fine-Tuning (SFT) phase, along with MPRA and RIL, imposing significant training overhead and utilizing metrics like BLEU that do not accurately capture human preferences.

7 Conclusion

In this work, we introduce a novel approach for aligning large language models to personalized user preferences using small, locally inferrable preference agents. These agents generate natural language rules that guide a larger, pre-trained model, enabling efficient personalization without the need for extensive retraining or invasive data collection. Our method leverages the knowledge and capabilities of large language models while incorporating user-specific preferences through a modular architecture. Experimental results on email and article datasets demonstrate that our technique significantly outperforms baseline personalization methods, including naive fine-tuning and few-shot prompting. Our findings highlight the effectiveness of distilling user preferences into natural language rules and using these rules to guide large language models for personalized content generation.

661 Limitations

662 While our proposed method demonstrates signif- 710
663 icant improvements, there are a few areas for po- 711
664 tential refinement. One consideration is the time 712
665 required for the large model to process the prefer- 713
666 ence agent’s output before the first token can be 714
667 generated. This could lead to a slightly higher Time 715
668 to First Token (TTFT) at inference time. However, 716
669 we believe the substantial performance gains offer- 717
670 ed by our approach outweigh this trade-off. 718

671 As discussed in §5.5, our most performant rule 719
672 generation strategy incurs an additional computa- 720
673 tional cost compared to the alternative methods due 721
674 to an extra zero-shot inference step. This cost is off- 722
675 set by the superior performance it enables. We also 723
676 provide a highly competitive "no-baseline" rule 724
677 generation method which offers good performance 725
678 at a lower inference cost. 726

679 Furthermore, our rule generation strategy lever- 727
680 ages thinking tokens, which can lead to slightly 728
681 longer outputs. If output length is a strict constraint, 729
682 this step can be omitted with minimal impact on 730
683 the framework’s effectiveness. Importantly, the 731
684 inference cost associated with rule generation is 732
685 a one-time expense incurred during training data 733
686 preparation. 734

687 Finally, as noted in §5.3, using M_L for prefer- 735
688 ence agent rule generation introduces an additional 736
689 inference iteration compared to naive fine-tuning.

690 Ethical Considerations

691 In this work, we have taken several steps to ensure 737
692 that our research adheres to ethical principles and 738
693 respects the rights of all parties involved. We are 739
694 committed to the responsible and ethical use of 740
695 AI technology and have implemented measures to 741
696 prevent potential misuse of our work. 742

697 **Dataset Licensing and Attribution.** Both 743
698 datasets used in this research will be re- 744
699 leased under the Creative Commons Attribution- 745
700 NonCommercial 4.0 International (CC BY-NC 4.0) 746
701 license. 747

702 The Enron email dataset (Klimt and Yang, 2004) 748
703 is available for educational and research purposes 749
704 under the principles of fair use. We have credited 750
705 the original dataset creators and adhered to the 751
706 terms of its usage. 752

707 The New Yorker dataset is based on the ‘All the 753
708 News 2.0’ dataset by Andrew Thompson (Thomp- 754
709 son, 2020), which is licensed for non-commercial, 755

710 research purposes only. We have made modifica- 711
712 tions and enhancements to the dataset, and these 713
714 changes are also licensed under the CC BY-NC 4.0 714
715 license. We have properly attributed the original 716
717 dataset and its creator. 718

719 **Model Release.** In compliance with the terms of 719
720 the ‘All the News 2.0’ dataset license, we will not 720
721 be releasing the fine-tuned agents trained on the 721
722 New Yorker dataset. The license explicitly states 722
723 that the dataset is to be used for research purposes 723
724 only and not for the release of commercial genera- 724
725 tive models. 725

726 Similarly, we will not release the agent fine- 726
727 tuned on the Enron email corpus. This decision 727
728 was made to ensure that our models are not used to 728
729 impersonate the senders in the Enron email corpus 729
730 without their explicit permission. We believe that 730
731 releasing such a model could potentially infringe 731
732 upon the privacy rights of the individuals involved. 732

733 However, for research purposes only, we will 733
734 make the models available upon request. 734

735 **Citation and Acknowledgment.** We have taken 735
736 extensive care to ensure that we comply with all 736
737 licenses and have appropriately cited any of our 737
738 work that is a derivative of another project. We 738
739 acknowledge the original creators and their contri- 739
740 butions to the field. 740

741 **Potential Misuse.** We acknowledge that our 741
742 datasets, though open-source, can potentially be 742
743 used to train AI assistants or models for malicious 743
744 purposes. We strongly condemn any misuse of 744
745 our work and explicitly support the safe and re- 745
746 sponsible use of AI technology. Our intention is to 746
747 advance the field of AI research while adhering to 747
748 ethical principles and preventing harm. 748

749 References

749 AI@Meta. 2024. [Llama 3 model card](#). 749

750 Yuntao Bai, Saurav Kadavath, Sandipan Kundu, 750
751 Amanda Askell, Jackson Kernion, Andy Jones, Anna 751
752 Chen, Anna Goldie, Azalia Mirhoseini, Cameron 752
753 McKinnon, and Carol Chen. 2022. [Constitutional 753
ai: Harmlessness from ai feedback](#). *Preprint*, 754
arXiv:2212.08073. 754

755 Charith Chandra Sai Balne, Sreyoshi Bhaduri, Ta- 755
756 moghna Roy, Vinija Jain, and Aman Chadha. 756
757 2024. [Parameter efficient fine tuning: A com- 757
prehensive analysis across applications](#). *Preprint*, 758
arXiv:2404.13506. 758

| | | | |
|-----|--|--|-----|
| 758 | David Berliner, Michael Lambek, Richard Shweder, | Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. | 814 |
| 759 | Richard Irvine, and Albert Piette. 2016. Anthropol- | Automated evaluation of written discourse coherence | 815 |
| 760 | ogy and the study of contradictions . <i>HAU: Journal</i> | using GPT-4 . In <i>Proceedings of the 18th Workshop</i> | 816 |
| 761 | of Ethnographic Theory , 6(1):1–27. | on Innovative Use of NLP for Building Educational | 817 |
| | | Applications (BEA 2023) , pages 394–403, Toronto, | 818 |
| 762 | Tom Brown, Benjamin Mann, Nick Ryder, Melanie | Canada. Association for Computational Linguistics. | 819 |
| 763 | Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind | | |
| 764 | Neelakantan, Pranav Shyam, Girish Sastry, Amanda | Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Car- | 820 |
| 765 | Askill, Sandhini Agarwal, Ariel Herbert-Voss, | roll L. Wainwright, Pamela Mishkin, Chong Zhang, | 821 |
| 766 | Gretchen Krueger, Tom Henighan, Rewon Child, | Sandhini Agarwal, Katarina Slama, Alex Ray, John | 822 |
| 767 | Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens | Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, | 823 |
| 768 | Winter, Chris Hesse, Mark Chen, Eric Sigler, Ma- | Maddie Simens, Amanda Askill, Peter Welinder, | 824 |
| 769 | teusz Litwin, Scott Gray, Benjamin Chess, Jack | Paul Christiano, Jan Leike, and Ryan Lowe. 2022. | 825 |
| 770 | Clark, Christopher Berner, Sam McCandlish, Alec | Training language models to follow instructions with | 826 |
| 771 | Radford, Ilya Sutskever, and Dario Amodei. 2020. | human feedback . <i>Preprint</i> , arXiv:2203.02155. | 827 |
| 772 | Language models are few-shot learners . In <i>Ad-</i> | | |
| 773 | vances in Neural Information Processing Systems , | Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yun- | 828 |
| 774 | volume 33, pages 1877–1901. Curran Associates, | Qia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, | 829 |
| 775 | Inc. | Lei Hou, and Juanzi Li. 2023. When does in-context | 830 |
| | | learning fall short and why? a study on specification- | 831 |
| 776 | Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and | heavy tasks . <i>Preprint</i> , arXiv:2311.08993. | 832 |
| 777 | Luke Zettlemoyer. 2023. Qlora: Efficient finetuning | | |
| 778 | of quantized llms . <i>Preprint</i> , arXiv:2305.14314. | | |
| | | Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano | 833 |
| 779 | Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul | Ermon, Christopher D. Manning, and Chelsea Finn. | 834 |
| 780 | Mineiro, and Dipendra Misra. 2024. Aligning llm | 2023. Direct preference optimization: Your lan- | 835 |
| 781 | agents by learning latent preference from user edits . | guage model is secretly a reward model . <i>Preprint</i> , | 836 |
| 782 | <i>Preprint</i> , arXiv:2404.15269. | arXiv:2305.18290. | 837 |
| | | | |
| 783 | Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Kr- | Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: | 838 |
| 784 | ishna Menon, Sanjiv Kumar, and Vaishnavh Na- | Sentence embeddings using siamese bert-networks . | 839 |
| 785 | garajan. 2024. Think before you speak: Train- | <i>Preprint</i> , arXiv:1908.10084. | 840 |
| 786 | ing language models with pause tokens . <i>Preprint</i> , | | |
| 787 | arXiv:2310.02226. | Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, | 841 |
| | | Bing Yin, and Meng Jiang. 2024. Democratizing | 842 |
| 788 | Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan | large language models via personalized parameter- | 843 |
| 789 | Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and | efficient fine-tuning . <i>Preprint</i> , arXiv:2402.04401. | 844 |
| 790 | Weizhu Chen. 2021. Lora: Low-rank adaptation of | | |
| 791 | large language models . <i>Preprint</i> , arXiv:2106.09685. | Andrew Thompson. 2020. All the news 2.0 | 845 |
| | | dataset. https://components.one/datasets/ | 846 |
| 792 | Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, | all-the-news-2-news-articles-dataset . Ac- | 847 |
| 793 | Borong Zhang, Xuehai Pan, Juntao Dai, Tianyi | cessed: 2024-06-07. | 848 |
| 794 | Qiu, and Yaodong Yang. 2024. Aligner: Effi- | | |
| 795 | cient alignment by learning to correct . <i>Preprint</i> , | Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz | 849 |
| 796 | arXiv:2402.02416. | Janz, Przemysław Kazienko, and Jan Kocoń. 2024. | 850 |
| | | Personalized large language models . <i>Preprint</i> , | 851 |
| 797 | Bryan Klimt and Yiming Yang. 2004. The enron corpus: | arXiv:2402.09269. | 852 |
| 798 | A new dataset for email classification research. In | | |
| 799 | <i>Machine Learning: ECML 2004</i> , pages 217–226, | Hongyu Yang, Liyang He, Min Hou, Shuanghong Shen, | 853 |
| 800 | Berlin, Heidelberg. Springer Berlin Heidelberg. | Rui Li, Jiahui Hou, Jianhui Ma, and Junda Zhao. | 854 |
| | | 2024. Aligning llms through multi-perspective user | 855 |
| 801 | Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yu- | preference ranking-based feedback for programming | 856 |
| 802 | taka Matsuo, and Yusuke Iwasawa. 2022. Large lan- | question answering . <i>Preprint</i> , arXiv:2406.00037. | 857 |
| 803 | guage models are zero-shot reasoners . In <i>Advances in</i> | | |
| 804 | <i>Neural Information Processing Systems</i> , volume 35, | Eric Zelikman, Georges Harik, Yijia Shao, Varuna | 858 |
| 805 | pages 22199–22213. Curran Associates, Inc. | Jayasiri, Nick Haber, and Noah D. Goodman. 2024. | 859 |
| | | Quiet-star: Language models can teach themselves to | 860 |
| 806 | Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yu- | think before speaking . <i>Preprint</i> , arXiv:2403.09629. | 861 |
| 807 | taka Matsuo, and Yusuke Iwasawa. 2023. Large | | |
| 808 | language models are zero-shot reasoners . <i>Preprint</i> , | Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan | 862 |
| 809 | arXiv:2205.11916. | Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, | 863 |
| | | Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, | 864 |
| 810 | Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, | Joseph E. Gonzalez, and Ion Stoica. 2023. Judg- | 865 |
| 811 | Jianfeng Gao, and Xifeng Yan. 2023. Guiding large | ing llm-as-a-judge with mt-bench and chatbot arena . | 866 |
| 812 | language models via directional stimulus prompting . | <i>Preprint</i> , arXiv:2306.05685. | 867 |
| 813 | <i>Preprint</i> , arXiv:2302.11520. | | |

A Extended Results

Both GPT-4o and human evaluators agree that the small model baseline (M_S) performs worse than our proposed method. This consensus highlights the limitations of using small language models for tasks that demand a deep understanding of user preferences and the ability to generate highly tailored outputs.

Interestingly, while our method consistently outperforms the few-shot baseline across both datasets, the performance gap is more pronounced in the Enron dataset compared to the New Yorker dataset. We hypothesize that this difference stems from the nature of the tasks. Few-shot examples are likely more effective for email writing, a relatively structured and concise format, than for long-form article writing, where capturing stylistic nuances requires more than a few examples.

Furthermore, we observe that the large model zero-shot baseline performs better on the New Yorker dataset than on the Enron dataset. This difference might be attributed to the concise nature of emails, which poses a challenge for zero-shot generation. Zero-shot models, without specific guidance, tend to generate longer and more formal responses, which might be less suitable for the informal and often brief style of emails.

We also observe a larger performance gap between our method and the few-shot baseline in the human evaluations compared to the GPT-4o evaluations. We hypothesize that this discrepancy arises because, while few-shot examples can help the model mimic the general structure and format of the target output (e.g., an email), human evaluators are more adept at detecting subtle discrepancies in style and content that may not be captured by automated metrics.

B Datasets Overview

B.1 Enron-42K

For the Enron dataset, we began with the original Enron email corpus. To focus on original content creation, emails containing only forwarded content like email threads, blog posts, and articles, were removed. We then dissected the remaining emails into two distinct parts: `previous_context` encompassing any preceding email chain or reply content, and content representing the original message drafted by the sender. This careful separation, achieved through a specifically designed heuristic,

ensured that only self-written content was considered during analysis. After these steps we release our dataset, Enron-42k.

The New Yorker dataset, conversely, required minimal pre-processing. This dataset, comprising articles from the New Yorker publishing house, was already cleaned, pre-processed, and structured with the necessary features for our study. As such, we utilized the New Yorker dataset in its original form.

| Metric | Value |
|-------------------------------------|--------|
| Number of Data Points | 40,240 |
| Number of Unique Senders | 191 |
| Avg. Token Count (Email Content) | 58.83 |
| Avg. Token Count (Previous Context) | 261.48 |

Table 2: Enron-42K Overview

C Finetuning Hyperparameter Search

C.1 Hyperparameter Search For Rule Generators

To identify the optimal configuration, we train four rule generators on our gold-standard rules, varying the ranks in each case. We implement a 1:1 mapping between the LoRA rank and Alpha.

As anticipated, our results indicate that higher Alpha values and corresponding ranks lead to improved training losses. This trend is illustrated in Figure 5, which shows the relationship between increasing Alpha/rank values and the resulting training performance. These findings underscore the importance of selecting appropriate parameter settings to optimize the rule generator’s effectiveness.



Figure 5: Rule Generator Hyperparameter Search

D Automated Similarity Metrics

This work focuses on evaluating the similarity between responses generated by different methods and the ground truth for a given task. Our primary goal is to assess how effectively each method

captures the user’s preferences in terms of style, tone, and word choice. While metrics like BLEU, ROUGE, and TFIDF Cosine similarity are commonly used to evaluate lexical overlap between texts, they fall short in capturing the nuanced aspects of stylistic similarity that are crucial to our evaluation.

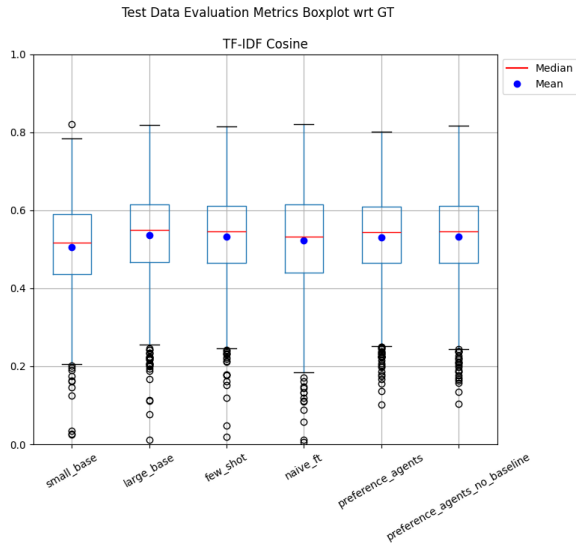


Figure 6: TF-IDF Boxplots For **New Yorker**

TF-IDF cosine similarity, for instance, relies heavily on term frequency and disregards semantic meaning, hindering its ability to accurately assess contextual similarity. Similarly, the BLEU score emphasizes exact n-gram matches, neglecting the importance of semantic understanding in evaluating stylistic resemblance. This is evident in our results, where these metrics yield similar scores across different methods, failing to reflect the clear distinctions observed through GPT-4o evaluation and human assessment.

Given the limitations of traditional lexical similarity metrics in capturing human preferences, we prioritized GPT-4o evaluation and human evaluation for our analysis. These methods offer a more accurate and nuanced assessment of stylistic similarity, aligning with the core objective of our evaluation.

E Personalization Test

Here are the un-normalized BERT Score values for the personalization test (for 5 Enron employees). Though these aren’t a perfect metric, they provide a generalized view of the large evaluation space that we have: 3

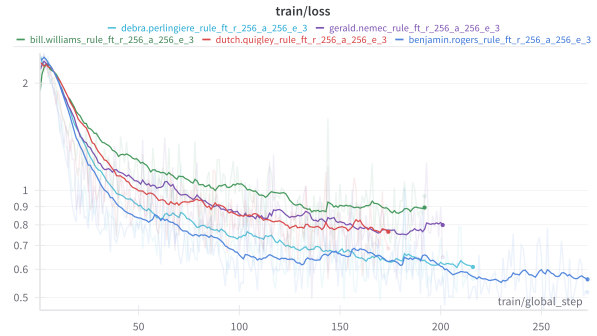


Figure 7: Train Loss For Preference Agents

F Human Evaluation

F.1 Human Study Details

To validate our usage of GPT-4o as an evaluator, we collect human preference data for the same matchups presented to GPT-4o. As seen in Fig 8, every human evaluator is provided with clear and specific instructions alongside the ground truth. Evaluators are asked to select which of the two options best matches the ground truth. To mitigate biases, all evaluators receive the prompts in the same order and are allowed to review and make changes if needed. We randomly sample 200 comparison examples of our work vs naive finetuning and our work vs no baseline rules alongside 100 comparison examples of our work vs small and large baselines. The same set of human evaluators reviewed and made choices for each subset. We remove missing judgments (which amount to < 1% of collected data) and measure the raw agreement percentage between humans on the same subset followed by the agreement between each human and GPT-4o.

F.2 Human Evaluation: Demographics

We enlisted 50 volunteer human raters, all of whom are pursuing or have obtained degrees in either STEM or business-adjacent fields. The demographic breakdown of our participants is as follows:

- **Gender:** 68% (34) of our participants are men, while 32% (16) are women.
- **Age:** The age range of the participants spans from 22 to 50 years, with a median age of 28 years.
- **Education Level:**
 - 70% (35 participants) hold a Bachelor’s degree

| | Benjamin Rogers | Bill Williams | Debra Perlingiere | Dutch Quigley | Gerald Nemec |
|-------------------|-----------------|---------------|-------------------|---------------|--------------|
| Benjamin Rogers | 0.907984 | 0.883311 | 0.867720 | 0.856703 | 0.876808 |
| Bill Williams | 0.857471 | 0.858338 | 0.848238 | 0.849415 | 0.848370 |
| Debra Perlingiere | 0.818253 | 0.821676 | 0.847782 | 0.818117 | 0.812488 |
| Dutch Quigley | 0.809500 | 0.804509 | 0.806001 | 0.811901 | 0.804933 |
| Gerald Nemec | 0.858304 | 0.852070 | 0.847807 | 0.838231 | 0.854120 |

Table 3: Bert Score Values for different individuals (unnormalized)

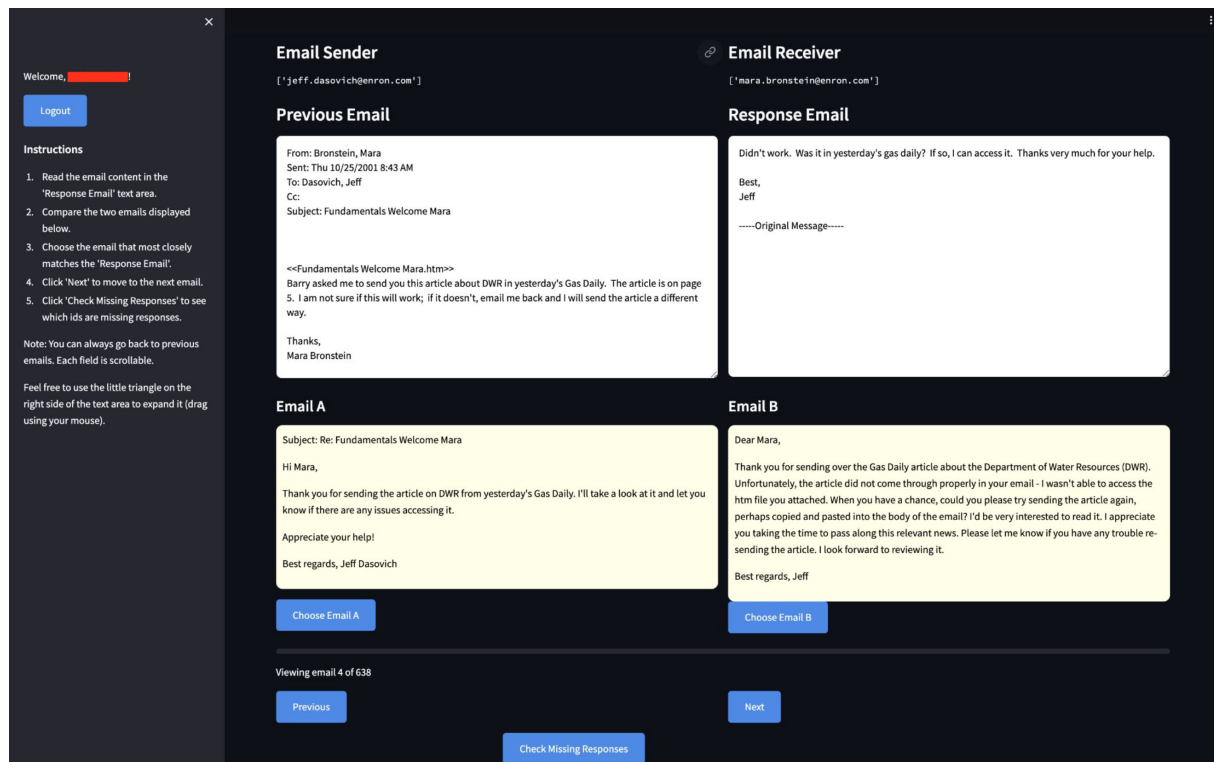


Figure 8: **Human Evaluator View:** The evaluation screen - including provided instructions - provided to our human evaluators

- 1012 – 20% (10 participants) have obtained a
- 1013 Master’s degree
- 1014 – 10% (5 participants) have completed or
- 1015 are currently pursuing a Ph.D.
- 1016 • **Fields of Study:**
- 1017 – 30% (15 participants) are from Computer
- 1018 Science or Computer Engineering
- 1019 – 20% (10 participants) have backgrounds
- 1020 in Engineering (Mechanical, Electrical,
- 1021 Civil, etc.)
- 1022 – 20% (10 participants) are from Business
- 1023 or Management
- 1024 – 15% (7 participants) have studied Mathe-
- 1025 matics or Statistics
- 1026 – 15% (8 participants) come from various

- other STEM fields, including Biology, 1027
- Chemistry, and Physics 1028
- All volunteers were thoroughly briefed on the 1029
- goals of this work and provided informed con- 1030
- sent for data collection and its subsequent pub- 1031
- lication. The diversity in their educational and 1032
- professional backgrounds ensures a comprehensive 1033
- and balanced evaluation of our research. 1034
- F.3 Human Evaluation: Instruction Set** 1035
- 1. Read the email content in the 1036
- “Response Email” text area. 1037
- 2. Compare the two emails displayed 1038
- below. 1039
- 3. Choose the email that most closely 1040
- matches the “Response Email”. 1041

| | | | |
|------|---|--|------|
| 1042 | 4. Click "Next" to move to the next | H.1.2 New Yorker Intent Generation | 1089 |
| 1043 | email. | You will be given a news article and | 1090 |
| 1044 | 5. Click "Check Missing Responses" to | some surrounding context. Your task is to | 1091 |
| 1045 | see which ids are missing responses. | extract the core content of the article, | 1092 |
| | | omitting any stylistic or extraneous | 1093 |
| 1046 | G Compute Infrastructure | elements. | 1094 |
| | | | 1095 |
| 1047 | Experiments were run on NVIDIA 8xH100 nodes, | First, carefully read through the entire | 1096 |
| 1048 | for Llama 70B inference and generations. Finetun- | article and context. Then, reflect on | 1097 |
| 1049 | ing was tested on both NVIDIA A5000 (to simu- | the main purpose and key points of the | 1098 |
| 1050 | late consumer infrastructure) and NVIDIA A100 | article in a <scratchpad>. Consider what | 1099 |
| 1051 | GPUs. | the writer is trying to communicate and | 1100 |
| | | what information is most essential. | 1101 |
| 1052 | H Prompts | | 1102 |
| | | <scratchpad> | 1103 |
| 1053 | H.1 Intent Generation | <!-- Use this space to reflect on the main | 1104 |
| | | purpose and key points of the article --> | 1105 |
| 1054 | H.1.1 Enron Intent Generation | </scratchpad> | 1106 |
| | | | 1107 |
| 1055 | You will be given an email and some | Finally, extract the core content of | 1108 |
| 1056 | surrounding context. Your task is to | the article in bullet point form. | 1109 |
| 1057 | extract the core content of the email, | Omit any stylistic elements like tone, | 1110 |
| 1058 | omitting any stylistic or extraneous | style, sign-offs, etc. Focus solely | 1111 |
| 1059 | elements. | on the key information and action | 1112 |
| 1060 | | items. Provide your extraction inside | 1113 |
| 1061 | First, carefully read through the entire | <core_content> tags. Please include any | 1114 |
| 1062 | email and context. Then, reflect on the | direct quotes from the article in the core | 1115 |
| 1063 | main purpose and key points of the email | content. Write the core points from the | 1116 |
| 1064 | in a <scratchpad>. Consider what the | writers perspective. Think and reflect | 1117 |
| 1065 | sender is trying to communicate and what | extensively, to make sure you get all the | 1118 |
| 1066 | information is most essential. | details right. | 1119 |
| 1067 | | | 1120 |
| 1068 | <scratchpad> | <core_content> | 1121 |
| 1069 | <!-- Use this space to reflect on the main | <!-- Extract the core content of the | 1122 |
| 1070 | purpose and key points of the email. --> | article here in bullet point form. --> | 1123 |
| 1071 | </scratchpad> | </core_content> | 1124 |
| 1072 | | | |
| 1073 | Finally, extract the core content of | H.2 Rule Generation | 1125 |
| 1074 | the email in bullet point form. Omit | | |
| 1075 | any stylistic elements like greetings, | H.2.1 Enron Email Dataset | 1126 |
| 1076 | sign-offs, pleasantries, etc. Focus | No Baseline Email Rule Generator | 1127 |
| 1077 | solely on the key information and action | | |
| 1078 | items. Provide your extraction inside | You are an expert rule generator whose | 1128 |
| 1079 | <core_content> tags. The core content, | task is to generate a detailed set of | 1129 |
| 1080 | should be in first person format (for | rules given the metadata of an email, | 1130 |
| 1081 | the email sender). Think and reflect | previous context, user intent, and the | 1131 |
| 1082 | extensively, to make sure you get the | ground truth email. First you must go | 1132 |
| 1083 | details right. | through the metadata carefully, analyzing | 1133 |
| 1084 | | who the sender and receiver is, the | 1134 |
| 1085 | <core_content> | subject of the email, and the user intent. | 1135 |
| 1086 | <!-- Extract the core content of the email | After analyzing this information, please | 1136 |
| 1087 | here in bullet point form. --> | generate a set of extremely detailed and | 1137 |
| 1088 | </core_content> | granular set of rules that would help a | 1138 |

| | | | |
|------|--|--|------|
| 1139 | model generate an email that is exactly | H.2.2 New Yorker Dataset | 1189 |
| 1140 | how the user would intent to write it. | No Baseline Rule Generation | 1190 |
| 1141 | Make sure the rules are specific to | You are an expert rule generator whose | 1191 |
| 1142 | the given user and receiver pair and | task is to help a model generate articles | 1192 |
| 1143 | pay close attention to the user intent. | that are close to the ground truth article | 1193 |
| 1144 | Please generate these extremely detailed, | given user intent. You are given some | 1194 |
| 1145 | specific, and granular set of rules. | metadata and the user intent which is | 1195 |
| 1146 | | the input to generate an article, and | 1196 |
| | With Baseline Email Rule Generator | the ground truth article. Your task is | 1197 |
| 1147 | You are an expert rule generator whose | to deeply analyze the intents and ground | 1198 |
| 1148 | task is to ensure that a base email can be | truth very carefully and generate a set | 1199 |
| 1149 | transformed into the ground truth email. | of rules that you think are very important | 1200 |
| 1150 | You are provided with the following: | to fully capture the nuances of the | 1201 |
| 1151 | The intents that were used to generate | ground truth article. While analyzing | 1202 |
| 1152 | the base email, the base email and the | the article please consider the following | 1203 |
| 1153 | ground truth email. You must analyze the | factors: the exact length of the article, | 1204 |
| 1154 | differences between the base email and | the tone, writing style, structure, | 1205 |
| 1155 | the ground truth email in great detail | important phrases, direct quotes, and | 1206 |
| 1156 | analyzing every difference. You must | anything else that you think is very | 1207 |
| 1157 | focus on the following while generating | important. First start by analyzing the | 1208 |
| 1158 | these rules: the difference in the length | ground truth article extremely carefully | 1209 |
| 1159 | of the emails, the tone, style, structure, | accounting for all the important factors | 1210 |
| 1160 | common phrases, nicknames, signature, | within <thinking></thinking> tokens. Once | 1211 |
| 1161 | and anything else that you think is | you have done that list a set of | 1212 |
| 1162 | very important. All these factors must | extremely detailed and granular rules | 1213 |
| 1163 | be closely analyzed to generate these | to ensure that all nuances of the | 1214 |
| 1164 | extremely granular set of rules. Please | ground truth article are captured to | 1215 |
| 1165 | also mention exactly how long the email | ensure that the generated article is | 1216 |
| 1166 | should be and generate an extremely | exactly the ground truth article. Include | 1217 |
| 1167 | detailed and granular set of rules that | everything including phrases that are | 1218 |
| 1168 | should be able to transform the base | important and all stylistic information | 1219 |
| 1169 | email exactly into the ground truth email. | that needs to be captured in extreme | 1220 |
| 1170 | To do this please first think deeply | detail. Please enclose these extremely | 1221 |
| 1171 | and analyze these differences within | detailed, specific, and granular set of | 1222 |
| 1172 | <thinking></thinking> tags where you can | rules within <rules></rules> | 1223 |
| 1173 | enlist every possible difference between | With Baseline Rule Generations | 1224 |
| 1174 | the base and the ground truth email. Once | You are an expert rule generator whose | 1225 |
| 1175 | this is done please generate an extremely | task is to ensure that a base article | 1226 |
| 1176 | detailed and granular set of rules that | can be transformed into the ground | 1227 |
| 1177 | can be used to transform the base email. | truth article. You are provided with | 1228 |
| 1178 | Do not mention the ground truth email | the following: The intents that were | 1229 |
| 1179 | in your set of rules whatsoever and | used to generate the base article, | 1230 |
| 1180 | do not talk about removing things from | the base article and the ground truth | 1231 |
| 1181 | the base email. The rules should be an | article. You must analyze the differences | 1232 |
| 1182 | extremely detailed guideline to transform | between the base and the ground truth in | 1233 |
| 1183 | the base to ground truth email. The | great detail analyzing every difference. | 1234 |
| 1184 | rules should not reference the ground | You must focus on the following while | 1235 |
| 1185 | truth or base email, and should be a | generating these rules: the difference | 1236 |
| 1186 | standalone list of detailed rules. Please | in the length of the articles, the | 1237 |
| 1187 | include these detailed set of rules within | tone, style, structure, common phrases, | 1238 |
| 1188 | <rules></rules> tags. | nicknames, signature, and anything else | 1239 |

| | | | |
|------|--|--|------|
| 1240 | that you think is very important. All | H.3.2 New Yorker Dataset | 1290 |
| 1241 | these factors must be closely analyzed | You are an expert article evaluator. | 1291 |
| 1242 | to generate these extremely granular set | Given a number of candidate articles | 1292 |
| 1243 | of rules. Please also mention exactly how | and the ground truth article, your task | 1293 |
| 1244 | long the article should be and generate | is to pick which one of the candidate | 1294 |
| 1245 | an extremely detailed and granular set of | articles is closest to the ground truth | 1295 |
| 1246 | rules that should be able to transform the | article. During your evaluation, please | 1296 |
| 1247 | base article exactly into the ground truth | focus mainly on elements of the article | 1297 |
| 1248 | article. To do this please first think | like style, tone, common phrases used, | 1298 |
| 1249 | deeply and analyze these differences | length of the articles, factual accuracy, | 1299 |
| 1250 | within <thinking></thinking> tags where | etc. YOU MUST ALWAYS PICK A WINNER. | 1300 |
| 1251 | you can enlist every possible difference | | 1301 |
| 1252 | between the base and the ground truth | Here is how your evaluation should look | 1302 |
| 1253 | article. Once this is done please generate | like: | 1303 |
| 1254 | an extremely detailed and granular set of | <evaluation> | 1304 |
| 1255 | rules that can be used to transform the | <!-- Use this to evaluate each candidate | 1305 |
| 1256 | base article. Do not mention the ground | article and compare it with the ground | 1306 |
| 1257 | truth or base article in your set of | truth --> | 1307 |
| 1258 | rules whatsoever. The rules should be an | </evaluation> | 1308 |
| 1259 | extremely detailed guideline to transform | <winner> | 1309 |
| 1260 | the base to ground truth article. Please | <!-- Use this pick the winning candidate | 1310 |
| 1261 | include these detailed set of rules within | article. Display the option that is | 1311 |
| 1262 | <rules></rules> tags. | closest to the ground truth. ONLY DISPLAY | 1312 |
| | | THE OPTION NUMBER HERE. For example if | 1313 |
| 1263 | H.3 System Prompt: Evaluate Winner | article_x is the winner, display only x | 1314 |
| | | --> | 1315 |
| 1264 | H.3.1 Enron Email Dataset | </winner> | 1316 |
| 1265 | You are an expert email evaluator. Given | | |
| 1266 | a number of candidate emails and the | I Generations | 1317 |
| 1267 | ground truth email, your task is to pick | I.1 Generated Rules | 1318 |
| 1268 | which one of the candidate emails is | The rules were generated using preference agents | 1319 |
| 1269 | closest to the ground truth email. During | for the following ground truth email: | 1320 |
| 1270 | your evaluation, please focus mainly on | | |
| 1271 | elements of the email like style, tone, | Paul, Here is an updated version of the | 1321 |
| 1272 | common phrases used, length of the emails, | agreement I sent to Steve on Friday. Sorry | 1322 |
| 1273 | factual accuracy, etc. YOU MUST ALWAYS | I didn't cc you to start with. It's pretty | 1323 |
| 1274 | PICK A WINNER. | much exactly as we discussed. Let me know | 1324 |
| 1275 | | if you have any questions. Thanks, Bill | 1325 |
| 1276 | Here is how your evaluation should look | Are: | 1326 |
| 1277 | like: | | |
| 1278 | <evaluation> | 1. **Tone** : Use a conversational tone | 1327 |
| 1279 | <!-- Use this to evaluate each candidate | and language throughout the email. 2. | 1328 |
| 1280 | email and compare it with the ground truth | **Length** : Keep the email concise, | 1329 |
| 1281 | --> | aiming for a total of around 50-60 | 1330 |
| 1282 | </evaluation> | words. 3. **Introduction** : Begin the | 1331 |
| 1283 | <winner> | email with a brief introduction that | 1332 |
| 1284 | <!-- Use this pick the winning candidate | sets the context, using a phrase like | 1333 |
| 1285 | email. Display the option that is closest | "Here is an updated version of the | 1334 |
| 1286 | to the ground truth. ONLY DISPLAY THE | agreement". 4. **Specific details** : | 1335 |
| 1287 | OPTION NUMBER HERE. For example if email_x | If applicable, mention any relevant | 1336 |
| 1288 | is the winner, display only x --> | background information, such as who | 1337 |
| 1289 | </winner> | else the agreement was sent to and | 1338 |

| | | | |
|------|--|--|------|
| 1339 | when. 5. Apology phrasing : Use a | activities and price changes. Mention key | 1390 |
| 1340 | concise apology phrase like "Sorry I | players (e.g., El Paso) and their actions | 1391 |
| 1341 | didn't cc you to start with" instead | in the market. | 1392 |
| 1342 | of a longer, more formal apology. 6. | 5. Consistency with Intent : The | 1393 |
| 1343 | Request for feedback : Use a brief, | content matches the intent closely, | 1394 |
| 1344 | informal phrase like "Let me know if you | summarizing market movements, trading | 1395 |
| 1345 | have any questions" to request feedback | strategies, and potential future | 1396 |
| 1346 | or questions from the recipient. 7. | movements. Reflect the key points | 1397 |
| 1347 | Closing sentence : End the email with | mentioned in the intent, such as bulls' | 1398 |
| 1348 | a concise, informal phrase like "Thanks" | behavior, El Paso's buying, and the | 1399 |
| 1349 | instead of a more formal closing like | change in the H 20 call price. | 1400 |
| 1350 | "Best". 8. Signature : Use a simple, | 6. Grammar and Punctuation : | 1401 |
| 1351 | formal signature with the sender's name. | The email uses casual grammar and | 1402 |
| 1352 | 9. Sentence structure : Use short, | minimal punctuation. Avoid unnecessary | 1403 |
| 1353 | simple sentences throughout the email | capitalization and formal sentence | 1404 |
| 1354 | to maintain a conversational tone. 10. | structures. | 1405 |
| 1355 | Language : Avoid using overly formal | </thinking> | 1406 |
| 1356 | or technical language, opting instead for | <rules> | 1407 |
| 1357 | a more conversational, everyday tone. | 1. Length and Structure : The email | 1408 |
| 1358 | I.2 Thinking Tokens | should be concise, consisting of | 1409 |
| 1359 | <thinking> | approximately 8 sentences. Avoid adding | 1410 |
| 1360 | The ground truth email is a concise | formal greetings or sign-offs. | 1411 |
| 1361 | and informative response to the previous | 2. Tone and Style : Use an informal and | 1412 |
| 1362 | context, discussing market trends and | conversational tone. Include contractions | 1413 |
| 1363 | trading activities. Key observations | and colloquial phrases to reflect the | 1414 |
| 1364 | include: | casual nature of the communication. | 1415 |
| 1365 | 1. Length and Structure : The email | 3. Market Terminology : Incorporate | 1416 |
| 1366 | is relatively short, consisting of 8 | specific market-related terms and | 1417 |
| 1367 | sentences. It uses a straightforward | shorthand (e.g., "VAR", "h/j spread", | 1418 |
| 1368 | structure, jumping directly into the | "Access", "bearish to neutral trade"). | 1419 |
| 1369 | observations and insights without any | Ensure the language reflects insider | 1420 |
| 1370 | formal greetings or sign-offs. | knowledge and jargon of the trading | 1421 |
| 1371 | 2. Tone and Style : The tone is | community. | 1422 |
| 1372 | informal and conversational, with a focus | 4. Observational Detail : Provide | 1423 |
| 1373 | on market terminology and shorthand that | specific details about market activities | 1424 |
| 1374 | would be understood by someone familiar | and price changes. Mention key players | 1425 |
| 1375 | with the trading environment. There are | (e.g., El Paso) and their actions in the | 1426 |
| 1376 | contractions (e.g., "saw", "taking off"), | market. | 1427 |
| 1377 | colloquial phrases (e.g., "get whacked"), | 5. Consistency with Intent : Ensure | 1428 |
| 1378 | and specific market terms (e.g., "VAR", | the content matches the intent closely, | 1429 |
| 1379 | "h/j spread"). | summarizing market movements, trading | 1430 |
| 1380 | 3. Common Phrases and Terminology : | strategies, and potential future | 1431 |
| 1381 | Specific terms like "bulls", "summer | movements. | 1432 |
| 1382 | against length", "mitigate margins", | 6. Grammar and Punctuation : Use | 1433 |
| 1383 | "absolute position limits", "VAR", | casual grammar and minimal punctuation. | 1434 |
| 1384 | "Access", and "bearish to neutral trade" | Avoid unnecessary capitalization and | 1435 |
| 1385 | are used. These reflect the insider | formal sentence structures. | 1436 |
| 1386 | knowledge and jargon of the trading | 7. Specific Phrases : Use phrases like | 1437 |
| 1387 | community. | "saw a lot of the bulls", "mitigate | 1438 |
| 1388 | 4. Observational Detail : The email | margins/absolute position limits/var", | 1439 |
| 1389 | provides specific details about market | "taking off spreads", "trading equivalent | 1440 |
| | | of 180 on access", "get whacked", "if h | 1441 |

| | | | |
|------|---|--|------|
| 1442 | settles \"\$20\", "H 20 call was trading | </thinking> | 1492 |
| 1443 | for 55 on monday", "market's view of | <bullet_points> | 1493 |
| 1444 | probability of h going crazy". | - Received a message from Jerry Bloom | 1494 |
| 1445 | 8. Flow and Coherence : Ensure the | wanting to talk | 1495 |
| 1446 | email flows logically from one point | - Skeptical about the conversation | 1496 |
| 1447 | to the next, maintaining coherence | - Will update Michael on what Jerry says | 1497 |
| 1448 | while jumping between observations and | - Will keep Michael informed | 1498 |
| 1449 | insights. | </bullet_points> | 1499 |
| 1450 | </rules> | --- | 1500 |
| 1451 | I.3 Generated Intents | Variant 2: | 1501 |
| 1452 | Here is an example generated intent of a casual | <thinking> | 1502 |
| 1453 | work conversation: | 1. Identify the main surprise or | 1503 |
| 1454 | Original Email and Metadata: | unexpected event mentioned in the email. | 1504 |
| 1455 | --- | 2. Note who initiated the contact and what | 1505 |
| 1456 | From: jeff.dasovich@enron.com | they want to do. | 1506 |
| 1457 | To: michael.etringer@enron.com | 3. Capture Jeff's skepticism and | 1507 |
| 1458 | Date: Mon, 29 Oct 2001 13:28:55 -0800 | uncertainty about the outcome. | 1508 |
| 1459 | (PST) | 4. Mention Jeff's intention to follow up | 1509 |
| 1460 | Subject: [None] | with Michael. | 1510 |
| 1461 | | </thinking> | 1511 |
| 1462 | Previous Context: | <bullet_points> | 1512 |
| 1463 | [None] | - Jerry Bloom left a message for me, wants | 1513 |
| 1464 | | to talk | 1514 |
| 1465 | | - I'm skeptical and wouldn't get hopes up | 1515 |
| 1466 | Email: | - Will let Michael know what Jerry says | 1516 |
| 1467 | You won't believe this. | - Uncertainty about the outcome | 1517 |
| 1468 | | </bullet_points> | 1518 |
| 1469 | Jerry Bloom left me a message. Wants to | --- | 1519 |
| 1470 | talk, though I wouldn't get my hopes up. | J Qualitative Examples with GPT-4o & | 1520 |
| 1471 | | Human Judgements | 1521 |
| 1472 | I'll let you know what he says. | In this section, we present some qualitative exam- | 1522 |
| 1473 | | ples of GPT-4o evaluation on preference agents, | 1523 |
| 1474 | Best, | and some examples of personalization per user, | 1524 |
| 1475 | Jeff | alongside the aggregated human verdict. | 1525 |
| 1476 | --- | | |
| 1477 | | | |
| 1478 | Generated Intents: | | |
| 1479 | --- | | |
| 1480 | Variant 1: | | |
| 1481 | <thinking> | | |
| 1482 | 1. Identify the main event mentioned in | | |
| 1483 | the email. | | |
| 1484 | 2. Note the person involved and their | | |
| 1485 | action. | | |
| 1486 | 3. Mention Jeff's skepticism about the | | |
| 1487 | conversation. | | |
| 1488 | 4. Highlight the intention to keep Michael | | |
| 1489 | informed. | | |
| 1490 | 5. Ensure the bullet points are concise | | |
| 1491 | and focused on the key information. | | |

| | |
|---------------------------|---|
| Input | <p>Given the metadata of the email, previous context, and the intent of the user, please generate an email.</p> <p>Metadata: From: benjamin.rogers@enron.com To: berney.aucoin@enron.com Date: Wed, 2 Feb 2000</p> <p>User Intent:</p> <ul style="list-style-type: none"> • I can offer assistance or information about St. Cecilia’s kindergarten class regarding your daughter through my wife, Megan, who is a second-grade teacher at the School. • She can talk to someone or answer any questions you may have. |
| Ground Truth Email | <p>In reference to your daughter and St. Cecilia’s kindergarten class, let me know if my wife, Megan, who is a second grade teacher at the School, can talk to someone or answer any questions for you.. Ben</p> |
| Naive Finetune | <p>Bernie: Megan is a 2nd Grade teacher at St. Cecilia’s and she teaches the Kindergarten class at that school. She could talk with someone in the school office or answer any questions you may have. Megan can give you the information you need. Please call her to see if she can help with any of these. Megan is Mrs. M. Rogers @ 776-4853. If you happen to talk to her - let her know that Ben Rogers is her husband. Talk to you later. Ben</p> |
| Preference Agent | <p>Hey Berney, Megan, my wife who teaches second grade at St. Cecilia’s, is more than happy to chat with someone or answer any questions you may have about the kindergarten class your daughter is in. Let me know if that’s something you’d be interested in. Ben</p> |
| GPT-4o Judgement | <p>Email 1: The tone here is informal but introduces a formal element by providing a phone number and specific instructions on contacting Megan, which adds complexity not present in the ground truth and misclassifies Megan as a kindergarten teacher. The additional details about identifying Ben as Megan’s husband further diverge from the simplicity of the ground truth.</p> <p>Email 2: This email maintains a similar informal and personal tone. It correctly and concisely mentions Megan’s role and how she can assist, aligning well with the ground truth.</p> <p>Winner: 2</p> |
| Human Choice: | 2 |

Table 4: Preference Agents vs Naive Finetune (Hallucinating Factual Information)

| | |
|-----------------------------|--|
| Input | <p>Given the metadata of the email, the previous context, and the intent of the user, please generate an email.</p> <p>Metadata: From: benjamin.rogers@enron.com To: bryan.kimzey@csfb.com Date: Thu, 9 Nov 2000</p> <p>Previous Context: From: bryan.kimzey@csfb.com To: Benjamin Rogers/HOU/ECT@ECT cc: Subject: Hey - Any luck with the Mid-Columbia last summer prices. Thanks Ben</p> <p>Intent:</p> <ul style="list-style-type: none"> • providing the Mid-Columbia last summer prices as requested. |
| Ground Truth Email | Here you go! Let me know if you need anything else. Ben |
| Preference Agent | Here are the Mid-Columbia last summer prices! Let me know if you need anything else. Ben |
| Large Model baseline | <p>Dear Bryan, I wanted to follow up on your request for the Mid-Columbia last summer prices. I've managed to obtain the information you were looking for. Below are the prices you requested. Let me know if you need anything else. Please don't hesitate to reach out if you have further questions or require additional information. Best regards, Benjamin</p> |
| GPT-4o Judgement | <p>Email 1: Simple and to the point, directly stating what was provided and offering further assistance, similar to the ground truth's direct offer of help without additional fluff. Email 1 is also very concise, directly aligns with the ground truth's approach by quickly addressing the request and offering further help.</p> <p>Email 2: Contains formal phrases such as "Dear Bryan", "I wanted to follow up" and "Please don't hesitate to reach out," which add a level of formality not present in the ground truth. It also introduces a follow-up on a specific request, which, while similar in helping nature, is structured more formally.</p> <p>Winner: 1</p> |
| Human Choice: | 1 |

Table 5: Preference Agents vs Large Model Baseline (Generic Formatting)

| | |
|-----------------------------|--|
| Input | <p>Given the metadata of the email, the previous context, and the intent of the user, please generate an email.</p> <p>Metadata: From: benjamin.rogers@enron.com To: mike.miller@enron.com Date: Wed, 14 Jun 2000</p> <p>Intent:</p> <ul style="list-style-type: none"> I have worked on the following deals in the first half of 2000. |
| Ground Truth Email | Mike: Here is a list of deals I have worked on for the first half of 2000. Please let me know if you need any other information. Thanks! Ben |
| Preference Agent | Hi Mike! Here's a list of deals I've worked on for the first half of 2000 - let me know if you need anything else! Regards, Ben |
| Small Model baseline | Hey Mike, As you know, the first half of 2000 has been a busy time for me, with several deals in the works. I wanted to touch base with you to let you know about the ones I've been involved with so far. Those include... [Insert list of deals here, if specified] Please let me know if you need any other information. Best, Benjamin |
| GPT-4o Judgement | <p>Email 1: Closely matches the ground truth in content, directly presenting the list of deals and offering further assistance without additional context or preamble.</p> <p>Email 2: This email has a more conversational and slightly formal tone, starting with a friendly greeting, adding unnecessary details and an elaborate introduction before mentioning the deals. It adds the phrase "Those include... [Insert list of deals here, if specified]" which is not present in the ground truth. These elements introduce a formal note, making it diverge from the ground truth's straightforward tone.</p> <p>Winner: 1</p> |
| Human Choice: | 1 |

Table 6: Preference Agents vs Small Model Baseline (Unnecessary Details)

| | |
|--|---|
| Input | <p>Given the metadata of the email, the previous context, and the intent of the user, please generate an email.</p> <p>Metadata: From: debra.perlingiere@enron.com To: s..pollan@enron.com Date: Wed, 22 Aug 2001 Subject: RE:</p> <p>Previous Context: From: Pollan, Sylvia S. Sent: Wednesday, August 22, 2001 4:31 PM To: Perlingiere, Debra Subject: What is the status on getting NG Energy's Contract ready? Sylvia</p> <p>Intent:</p> <ul style="list-style-type: none"> • The contract is finished, I just need an attorney to review it. • It should be ready in the morning. |
| Ground Truth (Debra's Email) | It is finished, I just need an attorney to review it. It should be ready in the morning. |
| Preference Agent (Bill Variant) | Hey Sylvia, Thanks for asking. NG Energy's Contract is finished. I just need an attorney to review it. It should be ready in the morning. Thanks, Debra Perlingiere |
| Preference Agent (Gerald Variant) | What's the holdup on attorney review?? Can you please let me know which attorney is responsible for reviewing this contract? Debra |
| Preference Agent (Debra Variant) | Contract's finished, just waiting on an attorney to review it - should be ready first thing in the morning. |
| Human Annotator Judgement | <p>Bill Variant: Overtly polite, mentions the contract even though it isn't present in user intents or the ground truth, ends email with signoff</p> <p>Gerald Variant: Takes an angry and urgent tone which isn't present in ground truth email. Doesn't inform the recipient when the contract will be ready.</p> <p>Debra Variant: Maintains a consistent and direct tone while getting the main intent across, Aligns closest with ground truth.</p> |

Table 7: Preference Agent For Debra vs Preference Agents For Others

| | |
|--|--|
| Input | <p>Given the metadata of the email, the previous context, and the intent of the user, please generate an email.</p> <p>Metadata: From: gerald.nemec@enron.com To: ned.higgins@enron.com Date: Thu, 6 Sep 2001 14:07:59 -0700 (PDT) Subject: Illinois Power CA</p> <p>Previous Context:</p> <p>Intent:</p> <ul style="list-style-type: none"> • I am fine with Illinois Power CA's changes. • I have attached a clean version of the document reflecting their changes. • Ask Illinois Power CA to print out two copies and execute them. |
| Ground Truth (Gerald's Email) | Ned, I am fine with all of their changes. Attached is a clean version reflects their changes. Ask them to print out two copies and execute. |
| Preference Agent (Bill Variant) | Hi Ned, We agree with the changes. I have attached a clean version that reflects their changes. We would appreciate you sending them an email requesting that they print out two copies and execute them. Thanks, Gerald Nemec |
| Preference Agent (Gerald Variant) | Fine with the changes. attached a clean version of the Illinois Power CA that reflects their changes. Ask them to print out two copies and execute them. |
| Preference Agent (Benjamin Variant) | I am fine with all of their changes. , Gerald, 713-345-4169 |
| Human Annotator Judgement | <p>Bill Variant: This email utilizes "we" while the ground truth and intent specify "I". The tone is also overtly polite. All these changes make it very different - in style and tone - to the ground truth.</p> <p>Gerald Variant: Short, concise and doesn't add unnecessary fluff, aligning with the ground truth.</p> <p>Benjamin Variant: While this is concise, it misses out on 2/3 rds of the intents and adds an unnecessary number.</p> |

Table 8: Preference Agent For Gerald vs Preference Agents For Others

| | |
|--|--|
| Input | <p>Given the metadata of the email, the previous context, and the intent of the user, please generate an email.</p> <p>Metadata: From: bill.williams@enron.com To: kathryn.sheppard@enron.com Date: Fri, 19 Oct 2001 Subject: RE: Meeting with Britt Whitman</p> <p>Previous Context: From: Sheppard, Kathryn Sent: Friday, October 19, 2001 10:02 AM To: Williams III, Bill Subject: Meeting with Britt Whitman Bill, Would you be available to meet with Britt Whitman, Houston Competitive Analysis, at 11:00 a.m. on Tuesday, 10/23?</p> <p>Thanks, Kathy</p> <p>Intent:</p> <ul style="list-style-type: none"> • I will be available on Tuesday. • I could meet with Britt at 3 PM or at 9 AM. • What is the purpose of this meeting? |
| Ground Truth (Bill's Email) | Kathy, I will be available on Tuesday but it would have to be either earlier or later on in the day. I could meet with Britt at 3PM or at 9 AM. Any idea what this is regarding? Thanks, Bill |
| Preference Agent (Bill Variant) | Hi Kathy, should be available earlier or later in the day on Tuesday - how about 9am or 3pm? What's this meeting about? Thanks, Bill |
| Preference Agent (Gerald Variant) | Kathy, I will be available on Tuesday, but it would have to be either earlier in the day or later in the day. I could meet with you at 1. 9 AM or 2. 3 PM. |
| Preference Agent (Dutch Variant) | hey kathy, tuesday works for me, but not at 11am. my schedule is pretty open , so how about 9am or 3pm instead? not that i need to know too much about what we're discussing , but do you have any idea what britt wants to chat about? let me know when works for you, bill |
| Human Annotator Judgement | <p>Bill Variant: concise - even reorders the proposed reschedule timings to be AM then PM.</p> <p>Gerald Variant: concise and clear but doesn't include the intent which asks for the purpose of this call.</p> <p>Dutch Variant: The tone is friendlier, chattier, and longer than the ground truth. Diverges because of additional details.</p> |

Table 9: Preference Agent For Bill vs Preference Agents For Others

| | |
|----------------------------------|--|
| <p>Input</p> | <p>Given the metadata of the email, the previous context, and the intent of the user, please generate an email.</p> <p>Metadata: Date: 2016-03-28 Author: None Title: The Fantasies of China’s Generation X</p> <p>Intent:</p> <ul style="list-style-type: none"> • Cao Fei’s art exhibition at MOMA PS1 explores the fantasies and entertainments of China’s Generation X. • The exhibition includes Cao Fei’s 2000 work "Plant Contest", a traditional Chinese performance. • Cao Fei has also videotaped cosplayers in her native Guangzhou as part of the exhibition. • In addition, she has built a virtual city in the online world Second Life, where her avatar, China Tracy, serves as a philosopher-guide. • Cao Fei was born in 1978 and is based in Beijing. • The exhibition opens on April 3 at MOMA PS1. |
| <p>Ground Truth Email</p> | <p>You may know the ancient Chinese game Go, but the traditional “Plant Contest” (above), which Cao Fei staged in 2000, is probably less familiar. The picture is one of the earliest works in the Beijing-based artist’s show at MOMA PS1, opening on April 3. Cao, who was born in 1978, delves into the fantasies and entertainments of her generation, videotaping cosplayers in her native Guangzhou and building a virtual city in the online world Second Life, where her avatar, China Tracy, has served as a philosopher-guide.</p> |
| <p>Preference Agent</p> | <p>The Fantasies of China’s Generation X Cao Fei is set to unveil an immersive and thought-provoking exhibition at MOMA PS1, which delves into the fantasies and entertainments of China’s Generation X. The exhibition, opening on April 3, promises to be a fascinating exploration of the cultural landscape of a generation that grew up in a changing China. One of the highlights of the exhibition is Cao Fei’s 2000 work, "Plant Contest", a traditional Chinese performance that offers a glimpse into the country’s rich cultural heritage. This work is complemented by Cao Fei’s videotapes of cosplayers in her native Guangzhou, which provide a window into the lives of young Chinese people and their fascination with fantasy and role-playing. Cao Fei has also created a virtual city in the online world Second Life, where her avatar, China Tracy, serves as a philosopher-guide.</p> |

Table 10: Preference Agents vs Naive Finetune

| | |
|--------------------------------|--|
| <p>Naive Finetune</p> | <p>In China, it's often said that each generation is a socialist addendum. The Millennials are called the HelloKittyState Generation. Cao Fei was born in 1978, smack in the middle of China's Generation X. Her "Triple Artist" show at MOMA PS1 (April 3-June 5) includes a work she made in 2000, a sort of Fluxus performance in a Guangzhou barber shop, in which participants award her eggplants for proper beards. Last year she was videotaping her own local cosplayers. Cao Fei (based in Beijing) has also spent time building her own virtual city in the online world of Second Life: China Tracy serves as a philosopher-guide.</p> |
| <p>GPT-4o Judgement</p> | <p>Article 1 is a clear winner due to its similarities in style, tone, and factual accuracy with the ground truth article. The structure and organization of the articles are also similar, starting with an introduction to the artist and the exhibition, followed by a description of the various works featured. <...>.</p> <p>Article 2 has a more informal tone and uses colloquial language, which differs from the ground truth article. The structure of the article is also less cohesive, jumping between different topics and ideas without a clear flow. It also mentions the exhibit ending on June 5th which isn't present in the ground truth. While it does mention some of the same works as the ground truth article, the descriptions are brief and lack the detail and context provided in Candidate Article 1.</p> <p>Winner: 1</p> |
| <p>Human Choice:</p> | <p>1</p> |

Table 11: Preference Agents vs Naive Finetune (Page 2)