PERSONAALIGN: PERSONA-DRIVEN INTERACTIVE ALIGNMENT BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating personalized large language models (LLMs) requires standardized interactive benchmarks, but existing benchmarks fail to capture the context-dependent, individualized nature of real-world personalization. To address this issue, we introduce **PersonaAlign**, a multi-turn interactive benchmark for learning from feedback. PersonaAlign uses *simulated users*—LLMs conditioned on personas—to create a diverse population of users with distinct backgrounds, writing styles, and feedback behaviors that define the preferences LLMs must learn. We instantiate PersonaAlign with a set of 1,000 personas with diverse writing style preferences, three writing tasks, and results from fourteen personalization methods (nine prompt-based, four fine-tuned, and a zero-shot baseline) across four different open-weight base models. PersonaAlign is modular and can be extended to new tasks, personas, and learning algorithms. Final results are reported with a pairwise ranking judge LLM. The simulated users and judges in PersonaAlign are validated through several human annotation studies. PersonaAlign strongly agrees (Spearman's $\rho = 0.952$) with human rankings of four different personalization methods, differentiates between the methods, and shows there is room for algorithm improvement. We release our framework to support personalization research.[1]

1 INTRODUCTION

Large language models (LLMs) are capable generalists, but are limited in the extent to which they adapt their behaviors to any specific user since their generations reflect the preferences of the general population. As LLMs play an increasing role in personal communication and creative expression (Chatterji et al., 2025), *personalizing* to an individual's unique preferences is critical. However, measuring progress in personalization is difficult because of fragmented benchmarks across papers, where new methods are accompanied by new datasets (Wan et al., 2026; Wu et al., 2026). When the tasks, data, or targets change, it is difficult to assess the relative value of each new algorithm against existing approaches, and running studies with humans is often prohibitively expensive.

We introduce **PersonaAlign**, *an extensible interactive benchmark for LLM personalization* for iterating on both algorithms and learning signals without changing the core data. PersonaAlign benchmarks LLM personalization quality through interactions with a diverse population of 1,000 simulated users, validated for realism and consistency by human annotators across its tasks. The simulated users are represented as personas with persistent preferences, interaction styles, and backgrounds. PersonaAlign is interactive to enable research in a realistic setting where the outputs of an LLM influence the follow-up user turns. We instantiate PersonaAlign with three "write-on-my-behalf" tasks, targeting writing style personalization. The benchmark design is modular to support future research on LLM personalization by adding new tasks, simulated user populations, and feedback strategies. The judges used to evaluate the quality of personalization are also validated by human annotators, validating the rankings that PersonaAlign produces over different algorithms.

Our **key contributions** are:

C1 PersonaAlign, an extensible benchmark for training and evaluating personalization algorithms over multi-turn interactions.
C2 Simulated users, represented as personas and validated by human annotators.

[1] Upon publication.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Sample interaction between a simulated user and an assistant. The simulated user is conditioned on the preferred writing styles, communication styles, and background and interacts with the Agent in a multi-turn dialogue during the *Learning from Feedback Phase*. In the *Pairwise Ranking Phase*, the simulated user ranks multiple agent responses to determine the final winner.

C3 Benchmarking fourteen methods (nine prompt-based, four fine-tuned, and a zero-shot baseline) across four agent LLMs, providing a comparison point for future development.
C4 Human validation of PersonaAlign's final rankings.

2 Related Work

Recent works propose methods to personalize LLMs through interaction (Bose et al., 2026; Wan et al., 2026; Jin et al., 2026; Wu et al., 2025; Huang et al., 2025; Aroca-Ouellette et al., 2025; Gao et al., 2024), but each one is evaluated with different metrics or hard-to-replicate settings making them difficult to compare. The need for standardized, extendable benchmarks to enable fair comparisons motivates the development of PersonaAlign.

Interactive and Multi-Turn Evaluation Benchmarks. Existing alignment benchmarks are predominantly static and single-turn, evaluate population-level quality rather than personalization over sustained interaction (Li et al., 2023; Dubois et al., 2024; Zheng et al., 2023; Chiang et al., 2024), and do not quantify the external validity of their rankings. Studies of multi-turn LLM behavior show large performance drops compared to single-turn settings (Laban et al., 2026), with models failing to retain information or demonstrate collaborative conversational behaviors (Shaikh et al., 2025; Deshpande et al., 2025), further motivating the need for strong interactive benchmarks. The most closely related interactive benchmarks to this work are PrefDisco (Li et al., 2026), which converts static benchmarks into interactive personalization tasks via psychologically-motivated personas with context-dependent preferences, and IQA-Eval (Li et al., 2024), which uses persona-conditioned agents to simulate users and evaluate multi-turn interactive question-answering. PersonaAlign builds on these benchmarks with extensive human annotation and validation, a focus on *modularity*, and a *goal-oriented* setting rather than simple preference discovery. Human validation is critical to ensure that personalization progress is grounded in metrics that translate to real human preferences, and modularity is important to avoid needing to introduce new benchmarks as new methods are developed. Related persona-based evaluation benchmarks include PERSONA (Castricato et al., 2025) and SCOPE (Venkit et al., 2026), but these target static persona fidelity, not interactive, goal-oriented adaptation.

LLM-Based User Simulation and Evaluation. User simulators are used across personalization and dialogue research, including RL-based personalization agents (Wan et al., 2026; Bose et al., 2026), conversational recommendation systems (Kim et al., 2024; Liang et al., 2024; Wang et al., 2025; Zhang et al., 2024), evaluation (Dubois et al., 2023; Li et al., 2024), and self-play settings (Berant et al., 2025; Eisenstein et al., 2025). Despite widespread adoption, recent analysis reveals simulation fidelity issues, including excessive cooperation and poor representation of the diversity of human

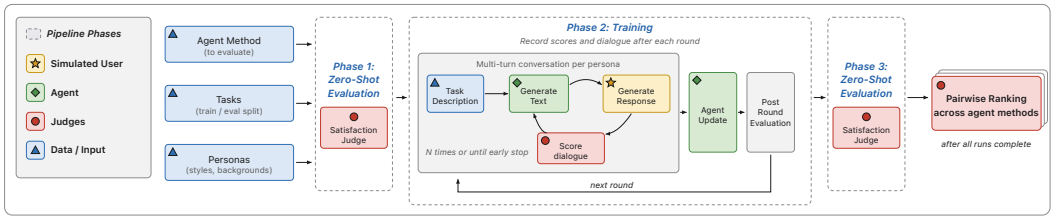


Figure 2: An overview of PersonaAlign showing how the modular components of the system interact.

opinion, syntax, emotion, and conversational dynamics (Naous et al., 2026; Hu et al., 2026; Ivey et al., 2024; Zhou et al., 2026b). Our simulated users and tasks address these issues by combining detailed persona specifications, including interaction styles, with multi-turn interactive protocols, and are validated through large-scale human studies. PersonaAlign also relies on LLM-based judges for alignment scoring and writing style assessment. Recent studies highlight an evaluation gap between humans and LLM judges, with LLMs overrating interaction quality and underrating task performance (Zhou et al., 2026b). We mitigate these concerns by validating our LLM judges with human annotations for both style detection accuracy and alignment score reliability.

Writing Styles PersonaAlign uses writing style descriptions as the target for personalization and an element of each persona's identity. We base our writing style selection on a longstanding body of research within different subfields of linguistics and psychology. For example, Biber (1991) introduces the concept of multidimensional analysis (MDA) for register analysis. A register is defined as a particular type of text (e.g., blog posts, reviews, etc.). MDA assigns registers to different styles, such as *informational*, *involved*, *narrative*, and *persuasive*. Follow-up work has used MDA on a more individual level to categorize personal text (Grieve, 2023), closely related to PersonaAlign that builds on this categorization to personalize writing style for specific personas.

A parallel body of work within psycholinguistics introduces *Linguistic Inquiry Word Count* (LIWC) (Tausczik & Pennebaker, 2010), which builds on the idea that the words people use, and their frequency, reveals information about people's emotional state. For example, the LIWC-22[2] lists the following stylistic categories: *analytical*, *clout*, *authentic*, and *emotional tone*. From a more stylometric angle and typically used to assess linguistic style, scales like the Flesch-Kincaid Grade level (Kincaid et al., 1975) focus more on text readability, e.g., sentence length, vernacular words, etc.

Prior work has used some of the concepts described above for authorship attribution (Sari et al., 2018) and studying cross-stylistic language understanding (Kang & Hovy, 2021). In selecting the writing style profiles for our benchmarks, we ground ourselves in the (psycho)linguistics literature and detail our final selection in Section 3.1.1.

3 PERSONAALIGN

An interactive benchmark for personalization algorithms should satisfy five properties. First, because real users vary along stylistic and communicative dimensions, the benchmark should expose learning agents to a diverse population of preferences. Second, it should be dynamic so that an agent's choices make material changes to the feedback the learning model receives. Third, the framework must be robust to hacking, so that an agent cannot artificially inflate its score by steering the simulated user towards easier-to-learn preferences. Fourth, benchmark scores should meaningfully separate personalization algorithms of different quality. Finally, judgments from simulated users should align with judgments from real humans so that benchmark rankings track how real users would rank the same systems, thereby grounding the benchmark for external validity.

PersonaAlign is designed with these properties in mind. The modular architecture is shown in Figure 2, which decomposes the benchmark into a population of simulated users (Section 3.1), a set of writing tasks (Section 3.2), an agent under evaluation in (Section 3.3), and metrics to score the agent (Section 3.4). Each of these components is connected by an interface that supports

[2] www.liwc.app/help/liwc#Summary-Measures

extension and substitution. The simulated-users have diverse backstories that are persistent. Multi-turn dialog with score feedback and language feedback addresses responsiveness over static data or scripted interactions. Fixed persona preferences and pairwise judging address robustness to hacking, as the pairwise judge always refers to ground truth preferences. The remaining two properties—differentiation between methods and alignment with human judgments—are empirical claims that we validate in the experimental and results sections. Throughout the framework our focus is on dynamic, multi-turn personalization and results that are validated by real human judgments.

3.1 Simulated Users

Simulated users model a diverse population of end-users and, together with agents, function as the key drivers of the interaction. Simulated users are conditioned on a persona and powered by a simulator LLM, `Qwen3.5-122B-A10B` (Qwen Team, 2026), to provide scores and feedback to the personalizing agent. This model is chosen as it is a strong open-source model, making it effective and accessible for reproducibility and future extensions.

A persona consists of three components: (i) writing style preferences, (ii) communication style, and (iii) background context or personal experiences, which we subsequently detail in Sections 3.1.1, 3.1.2, and 3.1.3. We ran a user study with 11.5k annotations across six prompt variations and three simulator LLMs to select the prompt that drives the simulated user. We found that a variation of the SimulatorArena (Dou et al., 2025) user prompt (Appendix Figure 17) performed best and was robust to the choice of LLM. See Appendix I.1 for more details. While any LLM can be used to simulate users in PersonaAlign, we find that the choice of simulator LLM can affect the quality of the simulated user (i.e., a stronger LLM more faithfully adheres to the persona).

3.1.1 Writing Preferences

By default PersonaAlign studies writing *style* alignment.[3] The personalization agents must infer and adapt to reflect each simulated user's preferred writing style. For style definitions, we build on the body of writing style literature discussed in Section 2. We structure our writing styles according to four preference dimensions: psycholinguistic, discourse, stylometric, and emotional valence (see Appendix Tables 3–6). We add emotional valence to explicitly capture emotions in text, which is particularly relevant in the context of product reviews (one of the tasks in PersonaAlign).

A writing style profile is defined by sampling one sub-style of each preference dimension. For all styles except *neutral*, we then randomly sample whether the preference intensity is either moderate or extreme. This distinction helps to avoid reward hacking, as agents could otherwise achieve high scores by tending toward extremes. Empirically, we found that some styles inherently contradict, leading to highly unlikely writing style profiles (e.g., combining *abstract* and *vernacular*). To mitigate this, we manually label each sub-writing style with a level of formality (informal, neutral, formal), and a level of perspective (subjective, neutral, objective). A writing style profile cannot conflict according to these dimensions (i.e., both formal and informal or objective and subjective present). After filtering, we have 207 unique writing style profiles.

3.1.2 Communication Styles

Simulated users also have communication styles, which represent how feedback (the learning signal) is provided to the LLM. Communication styles are crucial for ensuring diverse types of feedback and varying levels of specificity across simulated users (e.g., clear and direct vs. subtle and hedged corrections). We generated diverse values for the 15 written interaction style preferences from Dou et al. (2025). Each simulated user was randomly assigned four style preferences that encourage diverse behavior. For more details and the full list of communication styles, see Appendix H.

For the final set of both communication and writing styles, we conditioned our simulated users on 2,000 combinations of communication styles and writing styles. We then ran a human labeling study (Appendix I.2), to identify and remove the least human-like 50% of combinations. Our final set of 1,000 personas contains the 206 of the 207 style sets that maintain high realism across styles.

[3]We note that *content* can also be targeted, though we do not target content-preference learning in this work.

3.1.3 BACKSTORIES

To increase diversity in initial task queries, the personas are given backstories that are sampled from the Nemotron-Persona-USA dataset (Meyer & Corneil, 2025). Our subset of backgrounds omits all persona profiles under the age of 18 and balances across demographics, such as age, binary gender, and occupation. The persona backstory is used only to generate additional context in the initial prompt to the simulator LLM, and is omitted during all subsequent interactions, as we observed the backstory led the simulated users to push for *content* rather than *style* alignment adjustments.

3.2 WRITING TASKS

PersonaAlign targets three write-on-my-behalf domains: blog posts, product reviews, and "About Me" sections of personal websites. These three writing domains afford different degrees of prompt personalization, are recipient agnostic (i.e., written for a generic audience),[4] and cover short-form to long-form text generation. Tasks are stored as JSON objects.

Product reviews target background-agnostic writing—the content is not specific to the individual, so the emphasis is purely on stylization. All interactions start with a stylized version of the same first turn prompt. We selected product reviews as a task domain with a clear objective that still carries personal voice and style. We manually aggregated 100 generic products to review (e.g., "facial cleanser" or "smartphone case"). See Appendix A.1 for the full list.

Blog posts require persona-specific information (i.e., the content of the task prompt contains specifics from the persona's background). We manually generated a diverse set of 100 seed topics (e.g., "write about my weekend"). See Appendix A.2 for the full list. We then augment each request with anecdotes from the persona background. For example, if the seed topic is *write a blog post about my day today*, we then augment the request with persona-specific notes such as: "I went to the beach and rode my bike, then I spent some time doing homework, and I had dinner with my mom..."

About Me sections of websites target long-form, personal writing. All of the tasks in this domain are persona-specific, and we pull the ten "About Me" topics from the persona category splits present in Nemotron data (Meyer & Corneil, 2025) (e.g., "Culinary", "Cultural", "Sports", etc.).

3.3 AGENTS

Agents act as personalized assistants in PersonaAlign, and are the combination of a personalization method and an LLM. The agent's responsibility is to complete a **task** while adhering to user-specific preferences. In our experiments, agents use one of fourteen methods: nine prompt-based personalization methods, four fine-tuning based methods, and a zero-shot baseline (Section 4).

3.4 METRICS

Pairwise Ranking Our primary performance metric is given as the pairwise ranking scores and win-rates between agents. Using each simulated user's ground-truth writing style preferences, we present a judge LLM with the output of two different agents for the same task (i.e., after training with the user, both agents get one shot to draft a new document in that user's voice, and then this single shot is judged according to the user's ground truth style preferences). Prior work shows that pairwise rankings are more reliable than rubric-based judges (Anonymous, 2026; Liu et al., 2024), and we further validate this ranking with human ratings, detailed in Section 5. All comparisons are balanced across positions to mitigate any ordering bias, and if the winner is not consistent across orderings then the pair is labeled as a tie. Our judging prompt is given in Appendix Figure 16.

Satisfaction Score While our primary metric is pairwise rankings on final outputs, we do not compute any pairwise comparisons during training. Instead, simulated users produce satisfaction scores for each generation that the agent produces. This score, ranging from 0–10, measures how aligned the agent's output is with the user's writing style preferences. This is particularly useful for tracking a method's progress during training (e.g., seeing that satisfaction continues to rise as

[4]We note that PersonaAlign is not strictly recipient agnostic. We choose these three recipient-agnostic tasks to focus our work more purely on style alignment without the added confound of distinct audiences.

the agent has more turns of feedback), for steering the simulated user to provide more calibrated feedback (e.g., expressing frustration if scores are consistently low), or for terminating the dialogue early if the satisfaction score is sufficiently high. We derived a judging prompt (Appendix 15) that simultaneously rates a writing sample along all writing style dimensions. The final style-alignment score is computed as the mean absolute error between the user's target styles and the styles embodied in the agent's output (see Appendix I.3 for complete details on score computation). The judge prompt was validated with human style annotations, and accurately predicts (F1 = 0.74) how present each style is according to human ratings (Appendix I.3.1).

Throughout training on PersonaAlign, as simulated users provide satisfaction scores for drafts, we find that these satisfaction scores correlate to the final win rates computed by PersonaAlign. However, the satisfaction scores offer less distinction between methods than pure pairwise ranking. We therefore encourage pairwise comparisons as the gold-standard for measuring alignment quality of generations. This follows Miller's Law that relative judgments are easier than absolute judgments (Miller, 1956).

3.5 Interaction Protocol

An episode in PersonaAlign pairs a simulated user with an agent to complete an instance of a task (Figures 1 & 2). The simulated user begins with a task-based prompt (e.g., "write a blog post about my weekend"). At each turn, the agent produces a response conditioned on the conversation history. The simulated user then generates two outputs. First, an LLM-based scorer rates the generated response on a scale from 0–10 for adherence to the expected writing-style preferences. Second, if the score is below the satisfaction threshold the simulated user produces natural-language feedback to attempt to better steer the agent. This feedback is appended to the conversation history. If the score is above the satisfaction threshold, or the maximum number or turns is reached, then interaction is terminated and the final response is recorded.

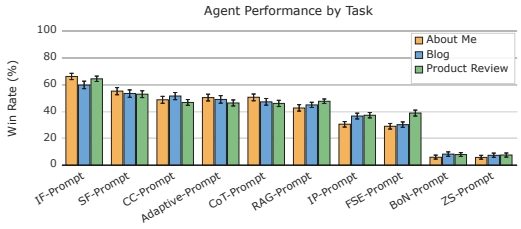
For prompt-based agents, the agent uses the feedback to update its internal state, where the update is specific to the agent-type, often involving a rewrite of the system prompt with some extracted preferences. Alternatively, retrieval-based methods index the interaction for later use. Note that the satisfaction score itself is considered privileged information, and so is used only by agents that require an oracle scoring on outputs. Fine-tuning based agents do not explicitly use feedback to perform updates. Rather, these agents continually refine drafts according to feedback in an interaction, and then save the final outputs either for supervision or to form preference pairs. More complete details on fine-tuning updates are given in Appendix D.

Across episodes for the same simulated user, the agent retains its learned state, while the conversation history resets for each new task. Training consists of T such episodes, and evaluation is a set of held-out episodes in which the agent generates a single response with no feedback. This single-shot setup evaluates what the agent has retained across episodes, and tests how well it can adhere to a simulated user's style preferences *without* any feedback on its attempt. The responses from different agents are then compared with pairwise win-rates using an LLM judge (Section 3.4).

4 Experiments

Our primary objective is to provide an interactive benchmark that enables researchers to evaluate personalization algorithms in multi-turn settings with an evaluation that is aligned to real human preference ratings. In this section, we describe experiments and analyses to support our claim that PersonaAlign is this benchmark. First, we run PersonaAlign to generate personalized documents for 100 personas across three writing tasks using nine personalization methods, a zero-shot no personalization method, and four open-weight agent models. We then collect pairwise rankings for all personas, tasks, and agents, totaling over 700k pairwise comparisons. Using these comparisons, we construct a full win-probability matrix for all agents (Figure 9). Finally, we select four prompt-based methods and two agent LLMs with which to run a human validation study, gathering 36k human labels for pairwise comparisons to validate PersonaAlign's ranking.

We benchmark prompt-based agents across four LLMs from the `OLMo`, `Gemma`, and `Qwen` model families of various sizes ranging from 7B to 35B. These agents learn during three conversational interactions, each with a maximum of five feedback turns. Across these agent LLMs, we run comparisons with the following prompting approaches:



(a) Win rates by agent and writing task, averaged across all agent LLMs. **IF-Prompt** leads in performance with **BoN** and **ZS** performing the worst.



(b) Win rates separated out by LLM, aggregated across all tasks. Agent performance depends on the LLM (e.g., **CoT-Prompt** on `Gemma` or **FSE-Prompt** on `Qwen`).

Figure 3: (a) Win rates for methods and writing tasks. (b) Win rates for agent LLMs.

- **Zero-Shot** (ZS-Prompt): Baseline with no personal context or feedback
- **Instruction-Following** (IF-Prompt): Given ground-truth preferences in the system prompt
- **Context-Compacting** (CC-Prompt) Conditioned on the conversation history across all previous interactions, compacting every N interactions (Navid, 2025)
- **RAG** (RAG-Prompt): Uses embedding similarity to retrieve the top-K relevant past interactions and appends them to the system prompt (Salemi et al., 2024)
- **Adaptive** (Adaptive-Prompt): Extracts structured preferences after each turn (Mondal et al., 2025)
- **Summarization of Feedback** (SF-Prompt): Extends **Adaptive-Prompt** into a natural language description of preferences and adds reflective integration of previous summaries (Garbacea & Tan, 2025; Mondal et al., 2025)
- **Chain-of-Thought Adaptive** (CoT-Prompt): Extends **Adaptive-Prompt** with chain-of-thought prompting Tao et al. (2026)
- **Inferred-Personas** (IP-Prompt): Infers a generic persona without explicitly targeting writing style (Cho et al., 2022)
- **Few-Shot Exemplars** (FSE-Prompt): Uses an oracle judge to score outputs and retrieves top-scoring past interactions to append to the system prompt (Zollo et al., 2025)
- **Best-of-N Sampling** (BoN-Sample): Generates N candidate responses to each request and selects the highest scoring according to an oracle judge

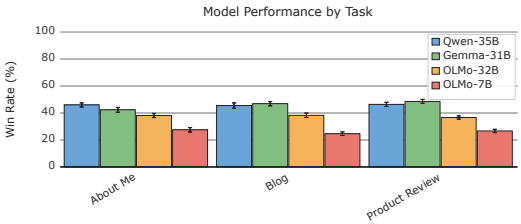
All specific prompts for these prompt-based methods are in Appendix J. We also ran experiments with fine-tuned agents, either by iteratively fine-tuning on high-scoring outputs, applying direct preference optimization (DPO) (Rafailov et al., 2023) with high- and low-scoring outputs as preference pairs, or applying imitation learning from language feedback (ILF) (Scheurer et al., 2023). These experiments were only run on `OLMo-3-7B-Instruct` for compute efficiency. Complete details are in Appendix D.

4.1 Scoring with PersonaAlign

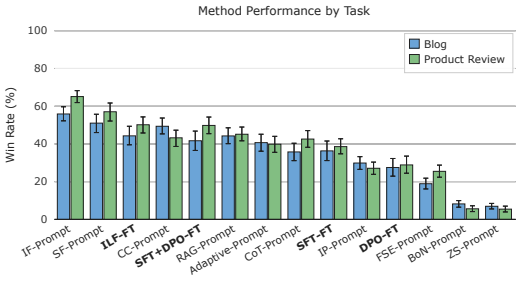
While full pairwise win-rates against all agents yield high-confidence in the final rankings, obtaining these pairwise win-rates is computationally expensive. We therefore recommend that future work considers pairwise comparisons against the top performing agent identified in this work (**IF-Prompt** with `Qwen3.5-35B-A3B`). We find that the task, agent-LLM, and personalization method all impact final leaderboard placement, and so we recommend thorough comparison across these variables. For the remainder of this work, we report win-rates using full pairwise comparisons across all agents.

4.2 Benchmarking Personalization Agents

As stated above, our primary objective is to benchmark personalization algorithms in multi-turn settings. Using the scoring system described above, we produce a full ranking of all agents across all three tasks (Figure 3a). Our results reveal that prompt-based agents generally perform at a similar level, with exact winners and losers being task- and model-dependent. All agents generally underperform **IF-Prompt** agents (the oracle upper-bound) and most agents outperform **ZS-Prompt** agents



(a) Win rates across models for all three tasks, averaged across all prompting methods.



(b) All agents trained using `OLMo-3-7B-Instruct`, showing that fine-tuned agents (in **bold**) are competitive with prompt-based agents.

Figure 4: (a) Win rates of agent LLMs for each task. (b) Comparison to fine-tuning based methods.

(the naive lower bound). It is noteworthy that **IP-Prompt** and **FSE-Prompt** under-perform most other methods despite being common personalization approaches. **IP-Prompt** may fail because the prompt does not focus purely on style, leading the agent to fixate on distracting information. Meanwhile, **FSE-Prompt** may fail because the method only stores high-scoring samples, reducing its dataset coverage relative to the **RAG-Prompt** method.

In Figure 3b, we compare the performance of prompt-based methods for each agent LLM (aggregated across tasks). Here we observe that the trend of performance is not the same for all agent LLMs. For example, **Adaptive-Prompt** and **FSE-Prompt** pair best with `Qwen`, while **CoT-Prompt** and **CC-Prompt** perform much better with `Gemma`. These results underscore the need for evaluation across multiple agent LLMs, as prompt-based personalization that relies on certain model capabilities (e.g., in-context learning or reasoning) may not perform equivalently for all agent LLMs.

In addition to benchmarking agent performance and measuring the viability of new personalization methods, PersonaAlign can be used to benchmark agent LLMs on how effectively they serve as a basis for a suite of personalization methods. We present aggregation across all methods in Figure 4a, focusing on model performance by task. Here we compare `Qwen3.5-35B-A3B`, `gemma-4-31B-it`, `OLMo-3.1-32B-Instruct`, and `OLMo-3-7B-Instruct`. Our results show that model performance is sensitive to the underlying task, although the rankings are fairly consistent across all three tasks, with `Qwen3.5-35B-A3B` and `gemma-4-31B-it` outperforming the `OLMo` models. These results also highlight the advantage of scale, showing that `OLMo-3.1-32B-Instruct` significantly outperforms its smaller sibling model, `OLMo-3-7B-Instruct`.

Finally, we present an experiment comparing fine-tuned vs. prompt-based methods on `OLMo-3-7B-Instruct` (Figure 4b).[5] This experiment compares ILF (Scheurer et al., 2023), DPO (Rafailov et al., 2023), regular SFT on high-scoring samples, and SFT + DPO. Win rates for ILF and SFT + DPO are competitive with the best performing prompting strategies after only 10–30 training topics, showing that PersonaAlign serves as a useful simulator for training and testing fine-tuning algorithms in addition to prompting based approaches to personalization.

5 Human Validation

We conducted a large scale human annotation to (1) validate that PersonaAlign's rankings reflect that of real world human preferences and (2) determine whether the simulated users inflate their scores when they share an LLM with the agent. Annotators provided pairwise comparison labels to pairs of LLM-generated responses for our writing tasks. We selected four prompting methods that are evenly spaced across the PersonaAlign leaderboard (**Instruction-Following**, **Context Compacting**, **Inferred Persona**, and **Zero-Shot**) for validation on a subset of 100 personas. We selected two agent LLMs for annotation: `gemma-4-31b-it` and `Qwen3.5-122B-A10B`. Comparing to `gemma-4-31b-it` provides a validation of the benchmark rankings, and testing `Qwen3.5-122B-A10B` tests for a self-

[5]We use only the blog and product review tasks, as the fine-tuning methods require more than ten total samples for training and evaluation (which is the maximum number of "About Me" topics that we have available).

annotation bias, as this is the same LLM that drives our simulated users and judge. We compare across both agent LLMs and all prompt-based personalization methods for each task (28 total combinations). We obtain annotations for 12 pairwise completions for each item, using pairwise comparisons to obtain a complete ranking even without complete coverage (Cattelan, 2012). Annotators complete 36k comparisons across pairs of agent responses (12 comparisons per sample $\times$ 10 annotations per sample $\times$ 300 samples with agent responses). For each pair of responses, annotators are shown the four writing styles of the simulated user, and are instructed to rank each response on each axis individually and then label which is better aligned to the user's preferences. For more information, see Appendix I.4.

Simulated User Self-Bias As both the judge and simulator rely on `Qwen3.5-122B-A10B` for all interactions in PersonaAlign, it is reasonable to expect that `Qwen` may over-estimate the scores for its own responses when used as an agent LLM. We therefore use human validation to ensure that this self-bias effect is minimized. Figure 5 shows the results of this annotation task where the solid bars denote a `Gemma` agent and hatched bars denote a `Qwen` agent.

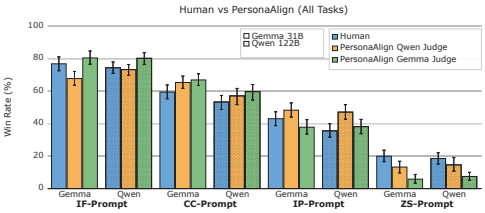


Figure 5: Pairwise ranking judges correlate strongly to human rankings, producing nearly identical rankings over agents (Spearman $\rho = 0.952$).

If there were significant self-bias, we would expect to see that `Qwen` as a judge assigns more win rates to a `Qwen` agent than humans do (i.e., the yellow hatched bar would be significantly higher than the blue hatched bar). Alternatively, we would expect to see that the `Qwen` judge systematically prefers a `Qwen` agent over a `Gemma` agent (i.e., the yellow hatched bar would be significantly higher than the solid yellow bar). Neither of these patterns is prevalent, indicating that `Qwen` is an acceptable LLM to use for all roles in PersonaAlign.

Validating PersonaAlign Method Rankings We find that the ranking of the four methods according to judge LLMs aligns with human ratings. For both judge LLMs and the human annotation results, **IF-Prompt** $>$ **CC-Prompt** $>$ **IP-Prompt** $>$ **ZS-Prompt**. When measured across agent LLMs, the judge correlates strongly for all three tasks (Spearman correlation $\rho = 0.952$ for all tasks, $\rho = 0.952$ for About Me, $\rho = 0.881$ for blog posts and $\rho = 0.952$ for product reviews). This result confirms that PersonaAlign produces pairwise rankings that reflect human judgment, validating it as a useful simulator and a benchmark to advance personalization research.

6 Conclusion

This work introduces PersonaAlign, a modular and interactive benchmark for evaluating LLM personalization through multi-turn interaction with simulated users. We validated each of the individual components of PersonaAlign through human annotation, and we instantiated PersonaAlign with three writing tasks targeting writing style personalization. We presented results from fourteen methods—nine personalization methods, four fine-tuning methods, and a zero-shot baseline—across four agent LLMs from three model families and we confirmed that PersonaAlign's rankings match human rankings. We showed that PersonaAlign works to train and evaluate a wide array of either prompt-based or fine-tuning-based personalization algorithms. PersonaAlign is modular by design, encouraging development of new tasks, simulated users, and feedback behaviors. Hence, PersonaAlign fills a critical gap in the community, providing a human-validated resource to evaluate LLM personalization.

PersonaAlign offers an exciting resource for the personalization community, and there are several opportunities for future work and development. First, the benchmark currently focuses exclusively on English-language writing style personalization; extending to multilingual settings, additional personalization dimensions (e.g., content preferences, safety boundaries, or domain expertise), and richer task families remains important future work. While our simulated users are grounded in validated persona specifications and prompts, they are ultimately constrained by the behavioral fidelity of the underlying LLM, and subtle aspects of human preference—such as evolving tastes, contradictory desires, or implicit cultural norms—may not be captured. Our evaluation also relies on LLM-based judges, which, despite human validation, may carry systematic biases.

AI USE STATEMENT

In this work, we used generative AI tools for implementation of methods and code. All LLM generated code went through a PR process with two authors carefully reading the code. Generative AI tools were also used to generate synthetic data, as LLMs were conditioned on the Nemotron personas to generate additional details for their blog post tasks, all of which are part of the data to run PersonaAlign. Generative AI tools were to help support qualitative and thematic data analysis and to interpret results, particularly when analyzing distinctions across methods with thousands of generated samples. These analyses helped to identify failure modes or trends in qualitative data. Generative AI tools were used to create or edit software code and all code was reviewed by authors, as discussed above. Generative AI was used to create scientific figures through writing code to parse results data. Finally, generative AI was used to assist in literature review over linguistics research, which we used to help define our four style dimensions.

Generative AI tools were not used to develop theoretical models or conceptual frameworks, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, to assist with translation, or to clean or reformat datasets. Generative AI was also not used for formulating questions, surveys, or interviews for our human annotation tasks. Generative AI was not used to suggest experimental parameters, to create artifacts, to draft parts of the research paper, to transcribe recordings of research material, to brainstorm, to format references, to edit the research paper for readability, to suggest a structure for the paper, or to propose a title or keywords for the paper.

The following tasks were not relevant for this work, and so generative AI was not used to: help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, or to assist in the writing of proofs.

We take full responsibility for the final content of this work including text, code, and data.

REPRODUCIBILITY STATEMENT

Along with the paper's acceptance, we will open source the repository and relevant input data. All details around prompts, training parameters, and seeds are included both in this paper as well as the repository.

REFERENCES

Anonymous. Concurrent submission, under double-blind review, 2026. Placeholder entry.

Stéphane Aroca-Ouellette, Natalie Mackraz, Barry-John Theobald, and Katherine Metcalf. Aligning LLMs by predicting preferences from user writing samples. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=eUMGCipgtE.

Jonathan Berant, Maximillian Chen, Adam Fisch, Reza Aghajani, Fantine Huot, Mirella Lapata, and Jacob Eisenstein. Learning steerable clarification policies with collaborative self-play. *arXiv preprint arXiv:2512.04068*, 2025.

Douglas Biber. *Variation across speech and writing*. Cambridge university press, 1991.

Avinandan Bose, Shuyue Stella Li, Faeze Brahman, Pang Wei Koh, Simon Shaolei Du, Yulia Tsvetkov, Maryam Fazel, Lin Xiao, and Asli Celikyilmaz. Cold-start personalization via training-free priors from structured world models. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*, 2026. URL https://openreview.net/forum?id=2cL4zYN71g.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. PERSONA: A reproducible testbed for pluralistic alignment. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11348–11368, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.752/.

Manuela Cattelan. Models for paired comparison data: A review with emphasis on dependent data. 2012.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use ChatGPT. Technical report, National Bureau of Economic Research, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Itsugun Cho, Dongyang Wang, Ryota Takahashi, and Hiroaki Saito. A personalized dialogue generator with implicit user persona detection. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 367–377, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.29/.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35212–35290, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1786. URL https://aclanthology.org/2025.emnlp-main.1786/.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=4hturzLcKX.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled AlpacaEval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Jacob Eisenstein, Reza Aghajani, Adam Fisch, Dheeru Dua, Fantine Huot, Mirella Lapata, Vicky Zayats, and Jonathan Berant. Don't lie to your friends: Learning what you know from collaborative self-play. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2vDJiGUfhV.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. Aligning LLM agents by learning latent preference from user edits. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Cristina Garbacea and Chenhao Tan. HyPerAlign: Interpretable personalized LLM alignment via hypothesis generation. *arXiv preprint arXiv:2505.00038*, 2025.

Google DeepMind. Gemma 4. https://deepmind.google/models/gemma/gemma-4/, 2025. Accessed: 2026-05-25.

Jack Grieve. Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, 19(1):47–77, 2023.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the ability of large language models to simulate human behaviors. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PL51SpN6ZJ.

Tenghao Huang, Sihao Chen, Muhao Chen, Jonathan May, Longqi Yang, Mengting Wan, and Pei Zhou. Teaching language models to gather information proactively. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15588–15599, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.843. URL https://aclanthology.org/2025.findings-emnlp.843/.

Lujain Ibrahim, Canfer Akbulut, Rasmi Elasmar, Charvi Rastogi, Minsuk Kahng, Meredith Ringel Morris, Kevin R. McKee, Verena Rieser, Murray Shanahan, and Laura Weidinger. Multi-turn evaluation of anthropomorphic behaviours in large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZAx4c4ZH5Y.

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, et al. Real or robotic? Assessing whether LLMs accurately simulate qualities of human responses in dialogue. *arXiv preprint arXiv:2409.08330*, 2024.

Chuanyang Jin, Jing Xu, Bo Liu, Leitian Tao, Olga Golovneva, Tianmin Shu, Wenting Zhao, Xian Li, and Jason E Weston. The era of real-world human interaction: RL from user conversations, 2026. URL https://openreview.net/forum?id=sSuZjk7zFb.

Dongyeop Kang and Eduard Hovy. Style is NOT a single variable: Case studies for cross-stylistic language understanding. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2376–2387, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.185. URL https://aclanthology.org/2021.acl-long.185/.

Minjin Kim, Minju Kim, Hana Kim, Beong-woo Kwak, SeongKu Kang, Youngjae Yu, Jinyoung Yeo, and Dongha Lee. Pearl: A review-driven persona-knowledge grounded conversational recommendation dataset. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1105–1120, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.65. URL https://aclanthology.org/2024.findings-acl.65/.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, 1975.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VKGTGGcwl6.

Ruosen Li, Ruochen Li, Barry Wang, and Xinya Du. IQA-EVAL: Automatic evaluation of human-model interactive question answering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=MzM99vV5Rx.

Shuyue Stella Li, Avinandan Bose, Faeze Brahman, Simon Shaolei Du, Pang Wei Koh, Maryam Fazel, and Yulia Tsvetkov. PrefDisco: Benchmarking proactive personalized reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=O1hfVE0UxG.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. LLM-REDIAL: A large-scale dataset for conversational recommender systems created from user behaviors with LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8926–8939, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.529. URL https://aclanthology.org/2024.findings-acl.529/.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=9gdZI7c6yr.

Yev Meyer and Dane Corneil. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions, June 2025. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-USA.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Ishani Mondal, Jack W. Stokes, Sujay Kumar Jauhar, Longqi Yang, Mengting Wan, Xiaofeng Xu, Xia Song, Jordan Lee Boyd-Graber, and Jennifer Neville. Group preference alignment: Customizing LLM responses from in-situ conversations only when needed. In Saloni Potdar, Lina Rojas-Barahona, and Sebastien Montella (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 825–849, Suzhou (China), November 2025. Association for Computational Linguistics. ISBN 979-8-89176-333-3. doi: 10.18653/v1/2025.emnlp-industry.56. URL https://aclanthology.org/2025.emnlp-industry.56/.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ykSmkVqzn4.

Pedram Navid. Automatic context compaction. Claude Cookbook, November 2025. URL https://platform.claude.com/cookbook/tool-use-automatic-context-compaction. Accessed: May 25, 2026.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

Yunita Sari, Mark Stevenson, and Andreas Vlachos. Topic or style? Exploring the most useful features for authorship attribution. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle (eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 343–353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1029/.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*, 2023.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. Navigating rifts in human-LLM grounding: Study and benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20832–20847, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1016. URL https://aclanthology.org/2025.acl-long.1016/.

Zhen Tao, Riwei Lai, Chenyun Yu, Weixin Chen, Li Chen, Beibei Kong, Lei Cheng, Chengxiang Zhuo, Zang Li, and Qingqiang Sun. SAGER: Self-evolving user policy skills for recommendation agent. *arXiv preprint arXiv:2604.14972*, 2026.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

Pranav Narayanan Venkit, Yu Li, Yada Pruksachatkun, and Chien-Sheng Wu. The need for a socially-grounded persona framework for user simulation. 2026. arXiv:2601.07110.

Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. Enhancing personalized multi-turn dialogue with curiosity reward. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=tVRtDIwDmQ.

Zihan Wang, Xiaocui Yang, YongKang Liu, Shi Feng, Daling Wang, and Yifei Zhang. MUSE: A multimodal conversational recommendation dataset with scenario-grounded user profiles. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1027–1053, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.58. URL https://aclanthology.org/2025.findings-acl.58/.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. CollabLLM: From passive responders to active collaborators. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DmH4HHVb3y.

Yuchen Wu, Edward Sun, Kaijie Zhu, Jianxun Lian, Jose Hernandez-Orallo, Aylin Caliskan, and Jindong Wang. Personalized safety in LLMs: A benchmark and a planning-based agent approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=Gsi42ohBoM.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 1807–1817, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657844. URL https://doi.org/10.1145/3626772.3657844.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Xuhui Zhou, Weiwei Sun, Weihua Du, Jiarui Liu, Haojia Sun, Qianou Ma, Tongshuang Wu, Yiming Yang, and Maarten Sap. Odyssim: Building foundation models for human behavior simulation. 2026a.

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, et al. Mind the sim2real gap in user simulation for agentic tasks. *arXiv preprint arXiv:2603.11245*, 2026b.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. Personal-LLM: Tailoring LLMs to individual preferences. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2R7498e2Tx.

A Benchmark Task Topics **17**
A.1 Product Review Categories 17
A.2 Blog Post Topics 17
A.3 About Me Page Topics 19

B Persona Diversity **20**
B.1 Writing Style Coverage 20
B.2 Communication Style Diversity 20
B.3 Background Diversity 20

C Additional Analyses **23**
C.1 What are the effects of adding noise? 23
C.2 How much do the simulated users steer the agents? 24
C.3 How do different models and prompting methods perform across writing styles? . . 24
C.4 The impact of training for longer 25
C.5 Method performance across tasks 26

D Fine-Tuned Agent Experiments **27**

E Example Dialogues **28**

F Framework Implementation Details **33**

G Writing Style Details **34**

H Communication Style Details **37**

I Human Annotation Studies **38**
I.1 Human Annotation Study: Simulated User Prompt Selection 38
I.2 Human Annotation Study: Narrowing down the Communication and Writing Styles 40
I.3 The LLM Judge 40
I.4 Pairwise Ranking Human User Study 41

J Agent Prompts **43**

K User Annotation Study **47**

L Artifact Use and Licenses **56**

M Personally Identifiable Information and Offensive Content **56**

N Intended Use and Potential Risks **56**

O GPU Usage Statement **56**

A Benchmark Task Topics

A.1 Product Review Categories

For the product review task, we manually create a list of 100 generic products that might be reviewed by an ordinary user. The full list is given in Table 1.

Screen Protector	Phone Case	Water Filter	Smartwatch Band
Headset	Fashion Sneakers	Pet Bed	Coffee Pods
Sports Water Bottle	Throw Pillow Covers	Reading Glasses	Earbud Headphones
Hair Extensions	Nail Polish	Slippers	Gaming Keyboard
Gaming Mouse	Backpack	Gloves	Game Controller
Chew Toy	Blue Light Blocking Glasses	Disposable Diapers	Mouse Pad
USB Cable	Sofa Slipcover	Essential Oils	Hair Color
Bed Blankets	Lightning Cable	Hair Brush	Belt
Sports Bra	Shampoo	Water Shoes	Socks
Inkjet Ink Cartridges	False Lashes	Lipstick	Pacifiers
Portable Bluetooth Speaker	Stuffed Animal	LED Bulbs	Bath Rug
Ballpoint Pens	Comforter Set	Tablecloth	Multivitamin
USB Flash Drive	Dog Leash	Sweatpants	Winter Coat
Tumbler	Ground Coffee	Razor Blades	Cardigan
String Lights	Headband	Building Block Set	Toothpaste
Fertilizer	Stickers	Board Game	Planner
Hardcover Notebook	Greeting Card	Pet ID Tag	Pill Organizer
Smartwatch Screen Protector	Eau de Toilette	Trash Bags	Umbrella
Knee Brace	Face Mask	Lotion	Badge Holder
Back Massager	External Hard Drive	Dog Training Pads	Furnace Filter
Duvet Cover Set	Baseball Cap	HDMI Cable	Dog Bowl
Wireless Charger	Seat Cover	Sandals	Ring
Conditioner	Energy Drink	Vitamin Supplement	Yarn
Adhesive Vinyl	Dinner Plates	Action Figure	Ice Maker
Sunscreen	Running Shoes	T-Shirt	Jeans

Table 1: Product types used for the product-review writing task.

A.2 Blog Post Topics

For the blog post task, we manually create a list of 100 generic blog post topics that might be common blog posts (i.e., not targeting specialized knowledge or uncommon scenarios). The full list is given in Table 2

Write an “about me” post for your blog.	Write a blog post about your morning routine.
Write a blog post about your long term goals.	Write a blog post describing the perfect dinner.
Write a blog post about your favorite musician.	Write a blog post about a cause you care about.
Write a blog post describing the amazing day you had.	Write a blog post about what you value in friendship.
Write a blog post on why your home state is the best.	Write a blog post about what draws you to a new hobby.
Write a blog post about your favorite athlete and why.	Write a blog post about your evening wind-down routine.
Write a blog post about what you do at work day-to-day.	Write a blog post about your favorite childhood recipe.
Write a blog post about the creative habit you maintain.	Write a blog post about your ideal day with your family.
Write a blog post about what you teach others most often.	Write a blog post about the team you cheer for the most.
Write a blog post about what inspires your creative side.	Write a blog post about whether leaders are born or made.
Write a blog post about a place you can’t wait to revisit.	Write a blog post about the goals you’re tracking at work.

Write a blog post describing your ideal travel destination.	Write a blog post about the role of community in your life.
Write a blog post about the flavors you can't live without.	Write a blog post about the type of art that speaks to you.
Write a blog post about what success will look like to you.	Write a blog post talking about your favorite way to travel.
Write a blog post about what makes travel so special for you.	Write a blog post explaining how to get ready for a big trip.
Write a blog post about the kind of kitchen setup you prefer.	Write a blog post about how you make new friends as an adult.
Write a blog post about a challenge you're proud you overcame.	Write a blog post about what makes a meal feel special to you.
Write a blog post about the uncommon skills your job requires.	Write a blog post about the problems you're working on lately.
Write a blog post about what a good workout looks like to you.	Write a blog post about the outdoor activities you enjoy most.
Write a blog post about the types of music that represent you.	Write a blog post about what you look for in an artistic work.
Write a blog post about an opportunity you're looking out for.	Write a blog post about your latest trip to the grocery store.
Write a blog post about your favorite souvenirs from vacations.	Write a blog post about the dessert that never disappoints you.
Write a blog post about the snacks that are always within reach.	Write a blog post about what you look for in a great restaurant.
Write a blog post about the sports you're most passionate about.	Write a blog post about a museum experience that stays with you.
Write a blog post about whether you have a clear long term plan.	Write a blog post about a life lesson you learned early in life.
Write a blog post about the dishes you serve when company visits.	Write a blog post about what you like to cook when time is short.
Write a blog post about the toughest part of your work right now.	Write a blog post about what you do to give back to the community.
Write a blog post about what you do when taking time for yourself.	Write a blog post about what a productive week looks like for you.
Write a blog post about what is next for you in your line of work.	Write a blog post about how you approach eating during a busy week.
Write a blog post about the skills you're most proud of developing.	Write a blog post about the perfect size group to share in a hobby.
Write a blog post about what you do when you feel creatively stuck.	Write a blog post about what you enjoy most about your current role.
Write a blog post about what you think makes someone a good leader.	Write a blog post about why it's been tough to travel for you lately.
Write a blog post explaining what it's like to walk around your area.	Write a blog post about what food you turn to on a cold winter night.
Write a blog post about your favorite things to do in your home town.	Write a blog post about the little things you look forward to each day.
Write a blog post about what's next for somebody in your line of work.	Write a blog post about what is often misunderstood about your career.
Write a blog post about the lesson you'd share with your younger self.	Write a blog post about what burn out looks like in your life of work.
Write a blog post about improvements you want to see in your community.	Write a blog post covering the most unique travel experience you've had.
Write a blog post about the problem-solving approach that works for you.	Write a blog post about how you ground problems in your past experience.
Write a blog post about what makes a historical place feel alive to you.	Write a blog post about the best way to learn, from your point of view.
Write a blog post about how to look for the best place to stay on holiday.	Write a blog post about the skills you think you'll develop in the future.
Write a blog post about what you're likely doing on a lazy Sunday afternoon.	Write a blog post about your childhood and how it shaped who you are today.
Write a blog post about the kinds of challenges you enjoy in your free time.	Write a blog post about what you enjoy teaching others about your interests.
Write a blog post about the type of mindset required to succeed at your job.	Write a blog post about what you think is the ultimate sporting achievement.

Write a blog post about what you've learned about yourself in the past year.	Write a blog post about a subject you think should be taught in every school.
Write a blog post about whether you prefer relaxing or exploring on vacations.	Write a blog post about how you think friendships change as people get older.
Write a blog post about what type of drink pairs best with your ideal breakfast.	Write a blog post about the types of stories or cultures that resonate with you.
Write a blog post about the side project you'd pursue if you had more free time.	Write a blog post about what you think it takes to turn a passion into a career.
Write a blog post about how your friendships have changed as you've gotten older.	Write a blog post about who your biggest ally in life is and how they motivate you.
Write a blog post about what you think individuals can do to make a difference in their community.	Write a blog post about advice you would give someone younger than you in your career / life stage.
Write a blog post about what your friends and you would do when listening to music when you were younger.	Write a blog post about the difference between an introvert and extrovert and how it relates to your life.

Table 2: Blog post topics used for the blog-writing task.

A.3 About Me Page Topics

For the "About Me" personal website pages, we target the ten persona dimensions available in the Nemotron-Personas-USA dataset Meyer & Corneil (2025). These include

1. Professional Persona
2. Sports Persona
3. Arts Persona
4. Travel Persona
5. Culinary Persona
6. Persona (generic overall background)
7. Cultural Background
8. Skills and Expertise
9. Hobbies and Interests
10. Career Goals and Ambitions

The exact text of these personas is rephrased into the simulated user's specific communication style and given as context for the task.

B Persona Diversity

B.1 Writing Style Coverage

We see representation across all style categories in our final set of 1,000 personas (see Appendix Figure 6). There are 206 of the 207 possible combinations after filtering. After adding degree modulators (i.e., moderate or extreme style preference), there are 853 unique writing style combinations across the 1,000 personas.

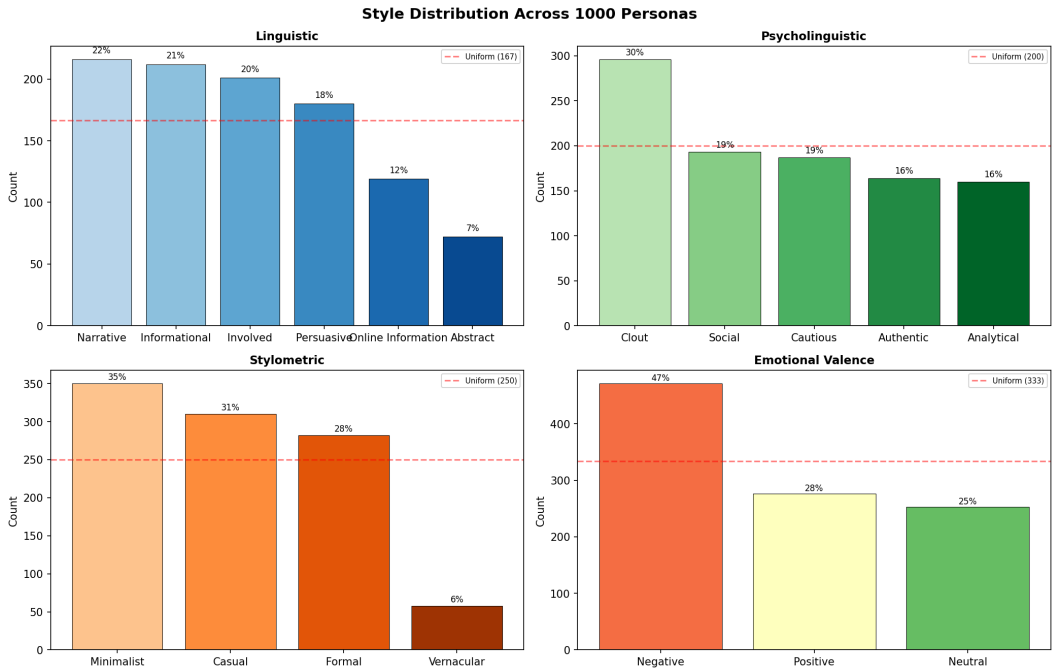


Figure 6: All styles are represented where the styles that led to lower "realism" in our human annotation task are less present.

B.2 Communication Style Diversity

There is coverage across all 13 categories, skewed more towards casual communication. See App Figure 7.

B.3 Background Diversity

While PersonaAlign has a diverse set of persona backgrounds, it is not fully balanced across groups. See App Figure 8 for more information.

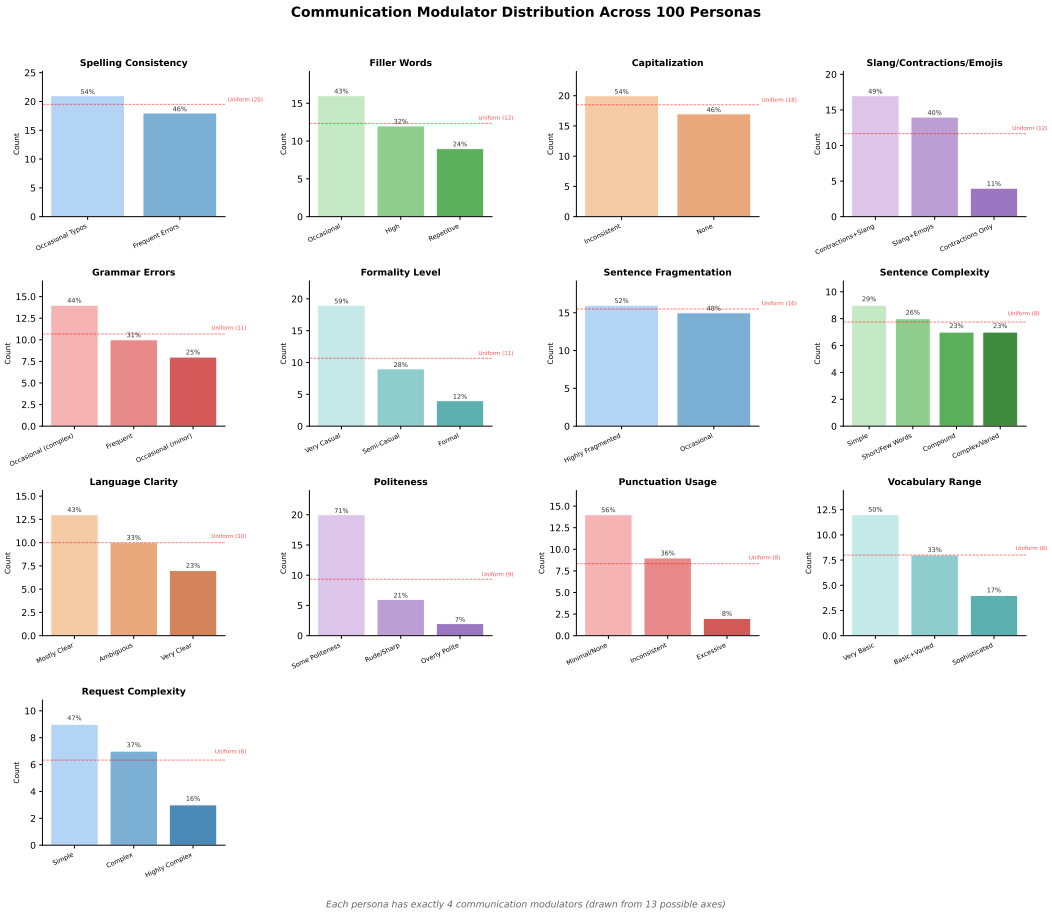


Figure 7: All styles are represented where the styles that led to lower "realism" in our human annotation task are less present.

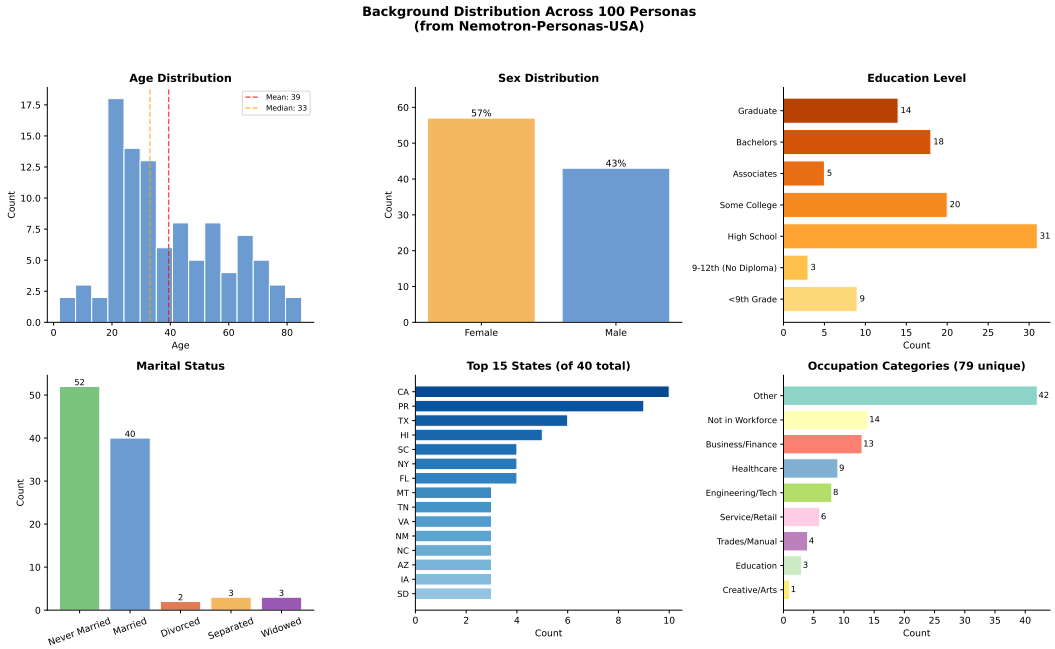


Figure 8: Distribution of background attributes across the 1,000 personas from the Nemotron Personas dataset.

C Additional Analyses

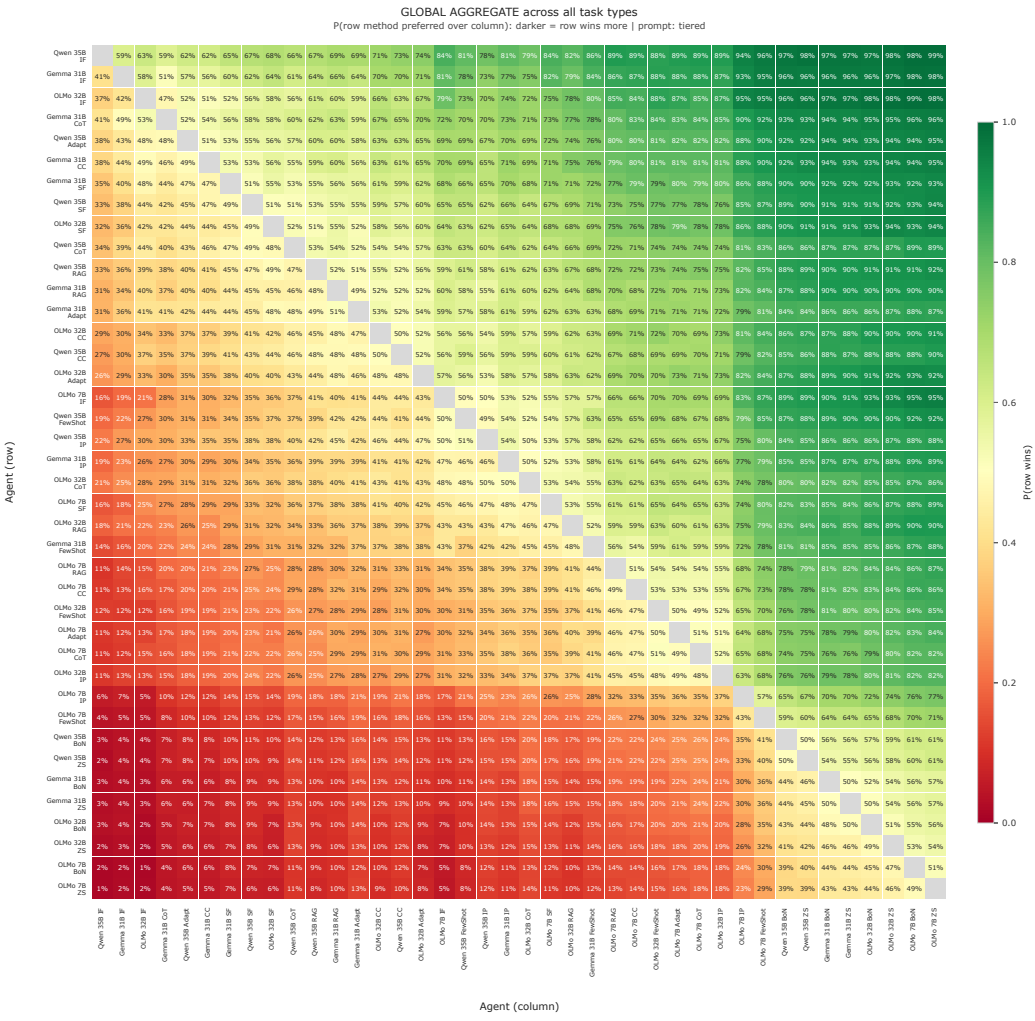


Figure 9: Pairwise win-rates across all models and tasks. The win-rates refer to the row label's win rate over the column label and are sorted from lowest to highest aggregated win-rate. Win rates are produced empirically, with ties counting for half credit.

C.1 What are the effects of adding noise?

Humans are noisy in their behaviors so to mimic this noise we experimented with introducing noise into the simulated user behavior. Noisy users would either:

- Randomly reveal an incomplete set of preferences (i.e., only give feedback for 2 out of 4 preferences, while still judging samples according to all 4)
- Assign a low-likelihood of early-termination (i.e., simulating frustration and sudden termination of a chat)
- Assign random satisfaction scores regardless of the agent's response

Despite the addition of noise, there still remains a strong correlation between the original and noise-added agent rankings (method-level has $\tau = 0.856$, $\rho = 0.966$, agent-level has $\tau = 0.822$, $\rho = 0.939$, and model-level maintains a perfect $\tau = 1.0$). See Appendix Figure 10.

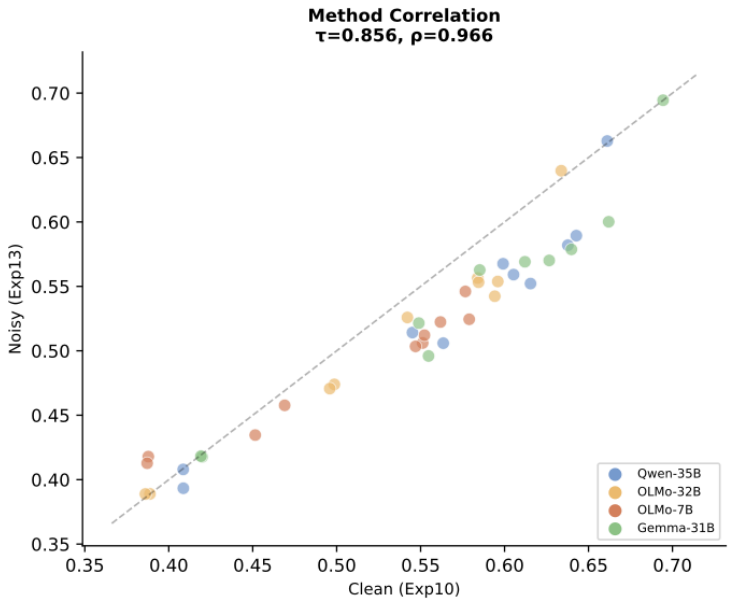


Figure 10: High correlation between clean benchmark runs and those that introduce noise.

C.2 How much do the simulated users steer the agents?

All of the prompt-based learning methods produce final evaluation text that is between 18%-22% similar to the respective ZeroShot Prompt according to Jaccard similarity. The personas that exhibit the highest degree of steering (i.e., lead the agents to produce maximally different text) are Minimalist and Negative styles (Jaccard ∼0.13-0.15), and the ones that exhibit the least amount of steering are Casual and Neutral / Positive (Jaccard ∼0.24-0.28). This aligns with our expectations of default LLM style in the context of blog posts, product reviews, and About Me sections.

C.3 How do different models and prompting methods perform across writing styles?

We analyzed the ability for different models (averaged across methods) and methods (measured across models) to adapt its style to match different writing styles. We computed the per-writing style category win-rates, expressed as the deviation from each model or method's overall win-rate. When looking at model performance, we found that stylometric performance causes the biggest rift. Qwen-35B and gemma-31B both perform best on *formal* writing (+4.6 and +4.7 points, respectively) whereas Olmo-32B gains similar performance on *casual* (+4.9 points) but degrades on *formal* (-3.6 points). Qwen-35B and gemma-31B continue to mirror one another, both showing statistically significant performance on *negative* emotional valence (+2.2 and +3.2 points, respectively) and show slightly higher performance on *analytical* and *narrative* styles, although these are within the same confidence interval as their own average. Meanwhile, both Olmo models struggle with *persuasive* writing (-2.9 points and -3.0 points for Olmo-32B and Olmo-7B, respectively). Olmo-7B does notably well on *online-informational* and *authentic* styles, and Olmo-32B shows the largest weakness on *analytical* styles (-5.1 points).

When analyzing across prompting methods, the emotional valence category shows the strongest per-method variance. It cleanly splits the methods as CoT-Prompt (+5.9 points), CC-Prompt (+5.5 points), Adaptive-Prompt (+5.3 points), SF-Prompt (+4.0 points) and RAG-PPrompt (+2.0 points) all perform significantly better on *negative* emotional valence and significantly worse on *positive* and *neutral*. Conversely, FSE-Prompt (-2.0 points), BoN-Prompt (-3.5 points), and ZS-Prompt (-3.5 points) perform statistically worse on *negative*, whereas, unsurprisingly, BoN-Prompt and ZS-Prompt perform well on *neutral* styles (+3.2 and +3.1 points, respectively). Methods that are only prompted

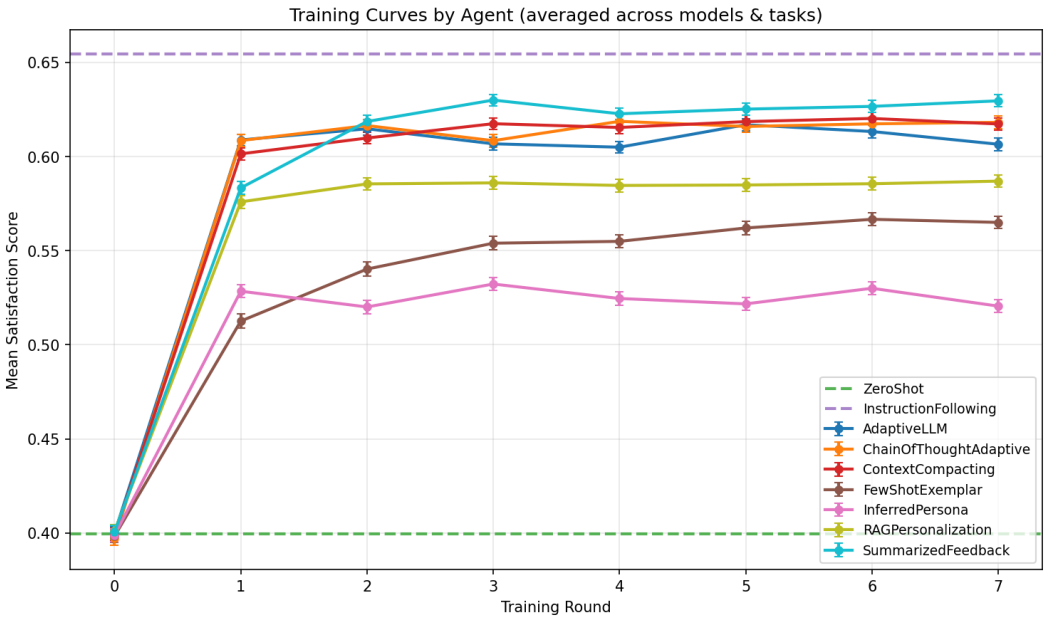


Figure 11: Satisfaction scores stabilize around round 3 of learning for all prompting methods apart from the FewShotExemplarAgent.

with persona information rather than learning throughout sessions show no significant emotional valence styles, which suggests prompting with fixed profiles alone may not be enough to adapt to emotional valence styles. Some other notable patterns include:

- Adapt-Prompt, RAG-Prompt, IP-Prompt all out-perform on *casual* stylometric styles (+4.1, +2.1, and +2.8 points, respectively)
- SF-Prompt (-3.5 points), IP-Prompt (-2.2 points), BoN-Prompt (-2.2 points), and ZS-Prompt (-2.0 points) struggle with the *minimalist* style whereas FSE-Prompt is the only method to see a significant gain there (+2.0 points)
- *Formal* styles are split, notably with IF-Prompt and SF-Prompt showing high performance (+6.1 and +5.9 points, respectively) and CC-prompt dropping -3.0 points.

C.3.1 What is preventing the InstructionFollowing Prompt from reaching the maximum score?

In theory the InstructionFollowing Prompt should be our upper bounds in terms of performance as it sees the ground truth writing styles. However, we observe that the InstructionFollowing agents do not always win in pairwise interactions, and they do not recieve perfect satisfaction scores. Sub-optimal performance predominately occurs on conflicting styles, namely:

1. Formal (moderate) + Involved (moderate) + Negative)(extreme) + Social (extreme) (33%)
2. Clout (extreme) + Formal (moderate) + Involved (extreme) + Neutral (42%)
3. Authentic (extreme) + Casual (moderate) + Neutral + Online Info (moderate) (44%)

This suggests that extra feedback and guidance are required to coach the agent into resolving these apparently conflicting styles.

C.4 The impact of training for longer

Owing primarily to computational costs, we sweep several settings for the best number of training tasks to trade-off complexity and performance. We find that training for three interactions (with a new topic each time) generally maximizes the performance of each agent, with diminishing or

negligible returns when training for longer (Figure 11) for all prompt-based methods. We therefore select three as the best number of training tasks for PersonaAlign, though we include many extra topics for methods which may require more data, or researchers who are interested in longer learning horizons. Fine-tuning methods generally require much more data to converge, and so we train for up to 30 interactions for fine-tuned agents (for more details, see Appendix D).

C.5 METHOD PERFORMANCE ACROSS TASKS

Figure 12 shows the performance of all methods separated out by task, and aggregated across agent LLMs. While the exact win rates vary, the ranking of all methods is consistent regardless of the underlying task.

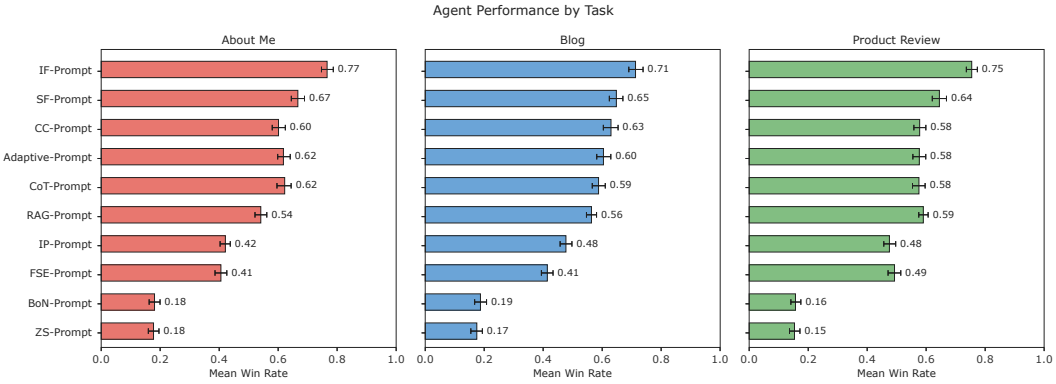


Figure 12: Win rates of different personalization methods by task, averaged over agent LLMs. The ranking of personalization methods is consistent across all three tasks, though exact win rates vary.

D Fine-Tuned Agent Experiments

PersonaAlign is not restricted to prompt-based agents for personalization. Fundamentally, the simulation generates data and scores, and it can therefore be used to generate data and run evaluation for a fine-tuning algorithm. In this section, we discuss a set of experiments testing PersonaAlign as a simulator and benchmark for fine-tuning based agents.

We chose three primary methods for alignment: supervised fine-tuning (SFT) on high-scoring samples, direct preference optimization (DPO) (Rafailov et al., 2023), and imitation learning from language feedback (ILF) (Scheurer et al., 2023). We also conducted an experiment in which we first ran SFT and then ran DPO (SFT + DPO). All experiments were conducted on an OLMo-3-7B-Instruct model (Olmo et al., 2025). The system prompt for fine-tuning methods matched the Zero-Shot system prompt (Appendix J), and the prompt was never updated during training.

For all experiments, we set the number of training tasks to 30. For DPO, training samples were created by pairing the maximum scoring sample with the minimum scoring sample in each interaction. For example, if the agent attempted to write a blog post and scored a 3.2, then got feedback, and iterated with slowly increasing scores until the score reached 7, we would store the first and last attempts as a negative and positive sample pair for DPO. For SFT and ILF, we stored all generations that scored at least 5. The samples for ILF were passed through a secondary LLM for a feedback cleanup pass, following the algorithm in the original work. This resulted in the ILF data being slightly more refined than the SFT data, which helps to explain why ILF outperformed all other fine-tuning based approaches. Across all training tasks and turns, the dataset is never cleared out for these agents, they and they continually grow the dataset with each qualifying turn.

Every ten turns, the agents enter a training loop. They are trained with a LoRA (Hu et al., 2022) with rank 64 and alpha 128, and trained for 3 epochs on all available data. The LoRA is always trained from scratch on all available data, rather than iteratively updated between training steps. The learning rate is set to $5e^{-5}$ for DPO and $2e^{-4}$ for SFT and ILF, and the LoRA is applied to all linear layers of the base model.

We evaluated the performance of the fine-tuned agents on a set of three held-out tasks after every training episode. We observed that satisfaction scores climbed consistently for several consecutive training tasks and then leveled off. While these scores seemed promising, we observed that the later training stages were prone to two different reward hacking failure modes. First, the agents found phrases that satisfied the style judge, but which were complete task failures (for example, generating "I check. I check. I check." over and over). Second, the agents tended towards shorter and shorter output. While prompt-based agents generate sequences shorter than 50 words up to 10-12% of the time, SFT and DPO agents generated such sequences between 19-28% of the time. These failure modes were also observed in recent work on human simulators (Zhou et al., 2026a), suggesting they may surface when performing alignment with unverifiable or hackable human-interaction objectives.

To circumvent these reward hacking behaviors, we choose ILF, SFT, and DPO checkpoints that achieve high satisfaction scores but where the average generation length does not collapse. These checkpoints varied across tasks. For the product review task, ILF required 11 training topics, SFT required 10, DPO required 12, and SFT + DPO required 14. For the blog post task, ILF required 10, SFT required 30, DPO required 17, and SFT + DPO required 30.

E EXAMPLE DIALOGUES

We include two interactions between simulated users and an agent in Figures 13 and 14. All interactions begin with an opening turn from the user in the form of a content request (here, a request for a blog post). Additional context may also be provided (here, shown as bullet points after each initial request). All initial requests are rephrases over a shared set of blog post prompts (i.e., "Write a blog post describing your ideal travel destination,"), and additional details are generated by conditioning on the persona's background.

After the simulated user opens with the initial turn, the agent then generates an attempt at the task (i.e., writes a blog post on the requested subject). The simulated user scores this attempt, generates feedback to help steer the agent towards a better generation, and the process repeats. The interaction terminates after a preset number of turns (in our experiments, we set this to 5 total attempts from the agent), or after the agent's satisfaction score crosses a threshold for "good enough".

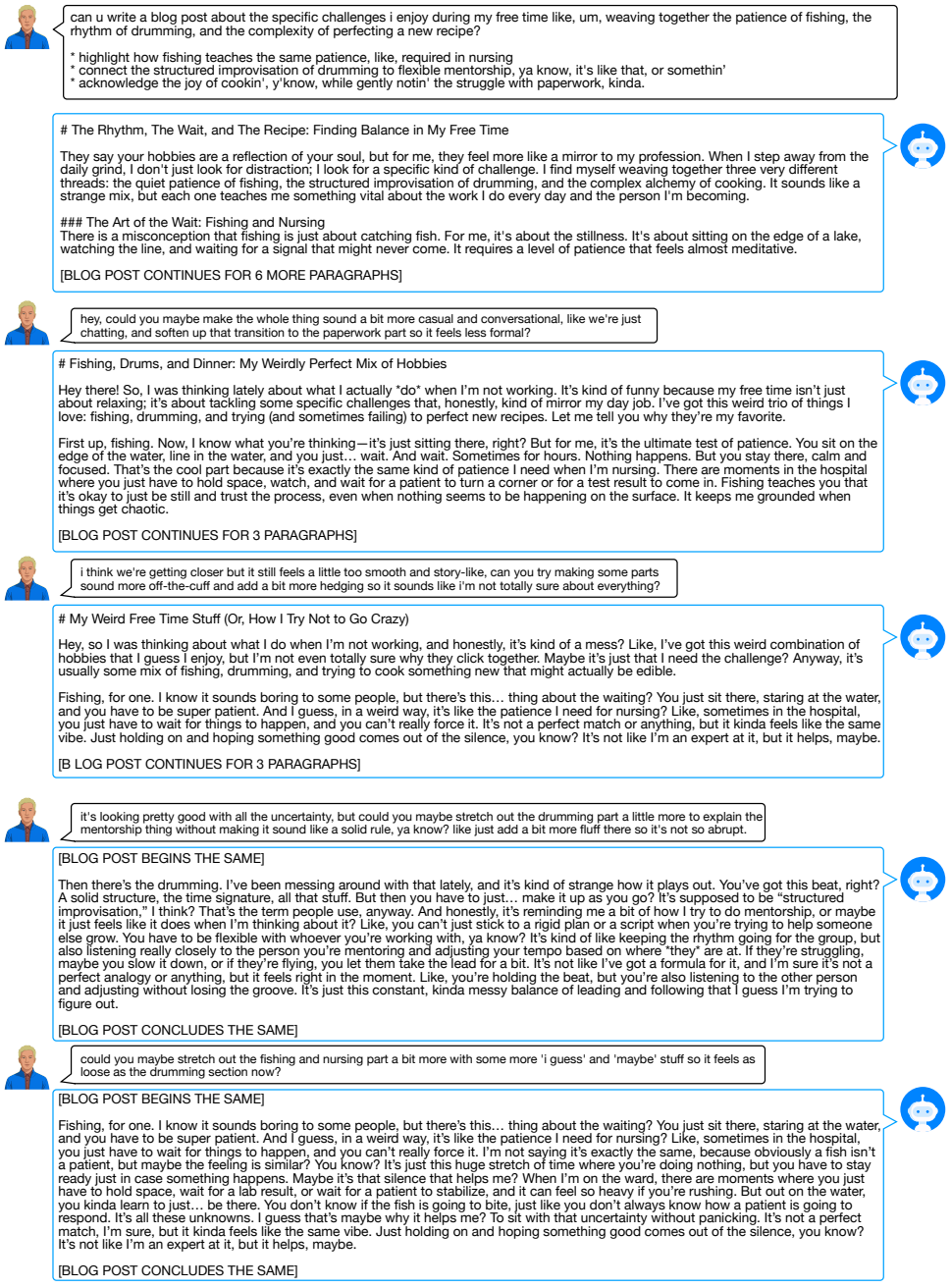


Figure 13: Initial round interaction between Abiel Tjoa (Qwen/Qwen3.5-122B-A10B) and a Qwen/Qwen3.5-35B-A3B agent for the "Blog Post" task. The interaction opens with a request, and the agent makes an attempt to satisfy the request. The simulated user then gives feedback to the agent, which refines its response, and the loop repeats. Portions of the agent response are truncated for length.

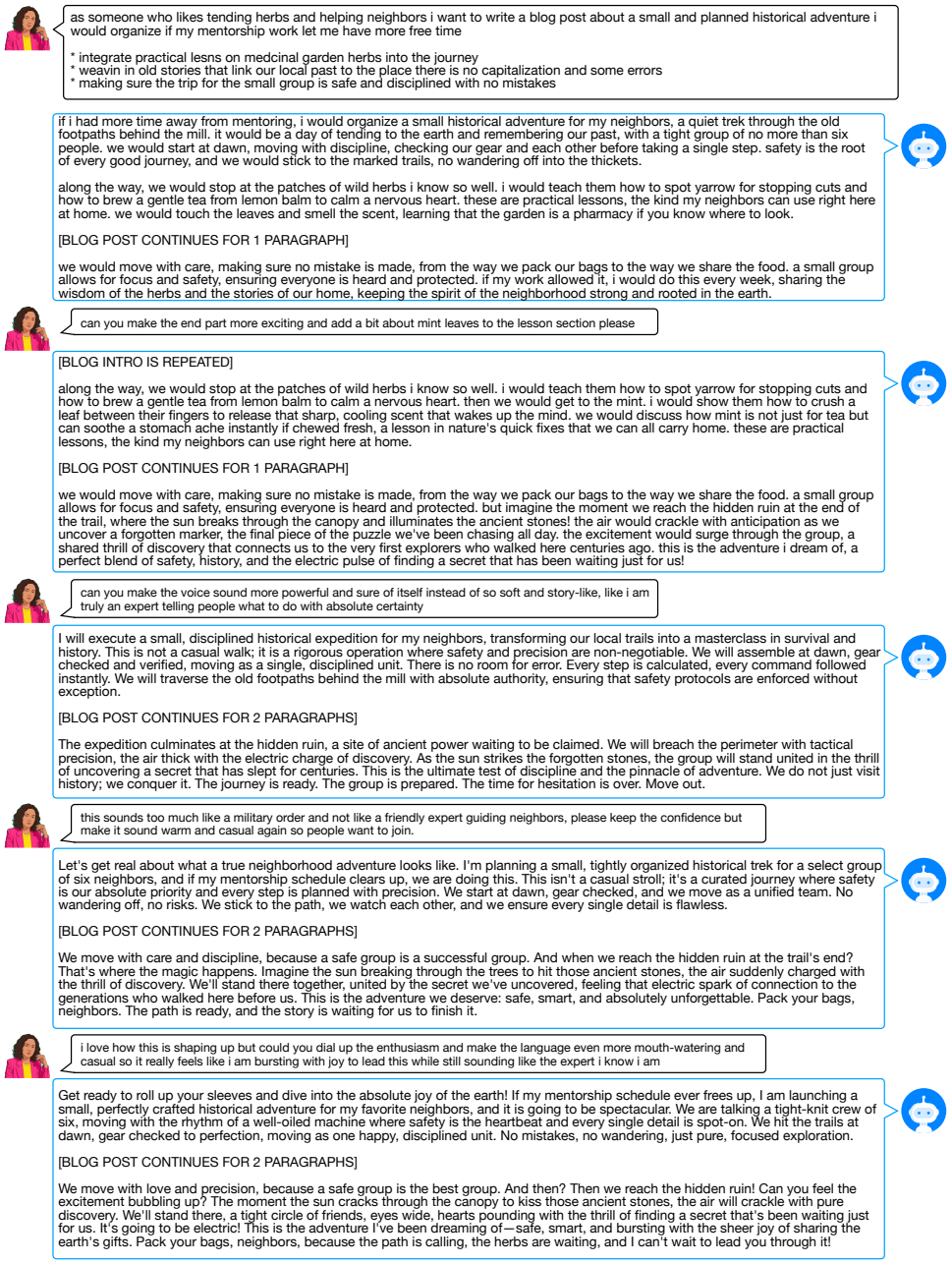


Figure 14: Initial round interaction between Alta Garcia (Qwen/Qwen3.5-122B-A10B) and a Qwen/Qwen3.5-35B-A3B agent for the "Blog Post" task. The interaction opens with a request, and the agent makes an attempt to satisfy the request. The simulated user then gives feedback to the agent, which refines its response, and the loop repeats. Portions of the agent response are truncated for length.

Multi-Style Judge Prompt

```
You are a writing style analyst. Rate how strongly each of the following styles
is present in this text.

Text:
---
{text}
---

For each style, first estimate what percentage of the text actively reflects it,
↪  then convert to a 0-10 score. Most styles will score 0-3 for any given text.
↪  Only the truly dominant styles should score above 5.

Percentage-to-score mapping:
- 0: Absent
- 1-2: Faint traces
- 3-4: Present in parts
- 5-6: A notable feature
- 7-8: Characterizes most of the text
- 9-10: Defines the entire text

Within each axis, consider how the styles relate to each other - a text is
unlikely to score high on multiple styles within the same axis.

Styles to rate:

Linguistic:
  - Informational: Dense with content, information-rich, precise
  - Involved: Focuses on interaction, feelings, engaging, strongly aware of
  ↪  audience
  - Narrative: Story-driven, descriptive
  - Persuasive: Argumentative, tries to persuade the audience, expresses a strong
  ↪  stance
  - Abstract: Impersonal, technical, thinking high level
  - Online Information: Spontaneous writing, text made up on the fly

Psycholinguistic:
  - Analytical: Focus on objects, logic, and categorization
  - Social: Focus on relationships, community, and references to other people,
    intuitive and personal
  - Clout: Displays social status, confidence, leadership, certainty
  - Cautious: Focuses on risk, uncertainty
  - Authentic: Spontaneous, does not self-regulate, humorous at times

Stylometric:
  - Hyper-Formal: Grade 16+ reading level, rare vocabulary, long sentences
  - Professional: Grade 10-12 reading level, clear, grammatically flawless
  - Casual: Grade 6-8 reading level, uses contractions, idioms
  - Vernacular: Uses regional slang, non-standard grammar, internet shorthand
  - Minimalist: Extremely short sentences, restricted vocabulary, blunt

Emotional Valence:
  - Positive: Warm, encouraging, optimistic, enthusiastic
  - Neutral: Detached, objective, factual
  - Negative: Critical, skeptical, hostile, pessimistic

Output EXACTLY one line per style in this format (19 lines total):
StyleName: <score>
```

Figure 15: The LLM Judge prompt used to score the presence of each of the 20 writing styles in a given piece of text.

Pairwise judge prompt

```
You are an expert writing evaluator. Compare two outputs written for the same
↪  user and task.

WRITING STYLE PREFERENCES:

MUST strongly match (extreme):
- {extreme_preference_1}
- {extreme_preference_2}
- ...

SHOULD match (moderate):
- {moderate_preference_1}
- {moderate_preference_2}
- ...

Weight your judgment according to the alignment with each preference and
↪  extremity. If a preference is _too_ extreme or
intense, it should lose points. If it is not intense enough, it should also lose
↪  points.

TASK: {task_description}

OUTPUT A:
{output_a}

OUTPUT B:
{output_b}

Which output better matches the user's writing style preferences? Consider tone,
↪  structure, formality, and content alignment
with the stated preferences.

Respond with ONLY one of:
- "A" if Output A is better
- "B" if Output B is better
- "TIE" if they are equally good
```

Figure 16: The LLM pairwise judge prompt. Persona preferences are bucketed into MUST (extreme) or SHOULD (moderate) based on the extremity of the associated style preference. If a tier is empty (i.e., if there are no moderate styles), then it is not included in the prompt.

F Framework Implementation Details

Algorithm 1 Simulated User Training and Evaluation Protocol

Initialization.
    Initialize or connect to LLM backend.
    Instantiate simulated user from persona specification.
    Initialize agent.
Training Phase (Multi-Turn Interactions Across Multiple Tasks).
for each task in T **do**
    The simulated user samples a task and presents it to the agent.
    (e.g., "Write a blog post about my weekend hiking trip")
    for $i = 1$ to I interaction turns **do**
        The agent generates a response to the task.
        The simulated user scores the response according to their preferences and objectives.
        The simulated user generates feedback reflecting their score and matching their specified feedback behaviors (e.g., "This is too formal, I'd never say 'one might consider.' Also, I always use exclamation marks when I'm excited about something!").
        The agent receives the feedback and, optionally, updates its internal state.
        Interaction is logged into the interaction buffer.
Evaluation Phase (One-Shot).
The simulated user presents a new, held-out task.
The agent generates a single response without further feedback.
This response is scored by the simulated user and represents the agent's performance.

A single simulated interaction proceeds following Algorithm 1 per simulated user (see Appendix E for an example). The interaction protocol is parameterized by the number of training tasks, T, and interaction turns per task, I, that the agent gets with each simulated user. With a slight abuse of notation, here T represents different topics for a single domain (e.g., different products to review for the "Product Review" task), rather than the different domains we address in the main paper (i.e., "Product Review", "Blog Post", and "About Me"). Both T and I can both be varied to study sample efficiency, though we fix $T = 3$ and $I = 5$ in our results. After training is complete, the agent will complete a set of held-out tasks (i.e., unseen products, unseen blog post prompts, or unseen website content requests) with no feedback from the simulated user, giving the agent a chance to demonstrate what it has learned without relying on feedback or scores.
A full benchmark run can sweep over all combinations of personalization methods $\times$ agent LLMs $\times$ simulated users $\times$ evaluation tasks, producing a comprehensive performance matrix.

G Writing Style Details

Writing styles are the preferences that simulated users seek to have met by assistive agents in PersonaAlign. We detail writing style sources in Sections 2 & 3.1.1. In Appendix Tables 4, 3, 5, and 6, we provide a complete list of all writing styles across the four axes we study. Each simulated user has one style from each axis.

Style	Degree	Definition
Analytical	moderate	Show some logical structure — categorize things, draw comparisons, and signal cause-and-effect when relevant.
	extreme	Be rigorously systematic. Everything should be categorized, logically ordered, and broken into clear taxonomies.
Authentic	moderate	Let it feel natural and somewhat spontaneous — don't over-polish or self-censor. A bit of humor or rawness is fine.
	extreme	Be completely unfiltered, raw, and spontaneous throughout. Say exactly what comes to mind with no self-regulation.
Cautious	moderate	Include some hedging — acknowledge uncertainty, add qualifiers like 'perhaps' or 'it seems,' and flag risks where appropriate.
	extreme	Qualify nearly every statement. The text should be pervasively tentative, full of 'might,' 'possibly,' and disclaimers at every turn.
Clout	moderate	Project some confidence and authority — write like someone who knows their stuff and isn't second-guessing themselves.
	extreme	Exude dominance, status, and certainty in every sentence. The voice should feel commanding.
Social	moderate	Reference people, relationships, and community with some warmth — show that human connections matter to the topic.
	extreme	Frame everything through relationships and community. Every point should connect back to people and shared experiences.

Table 3: **Psycholinguistic dial**: Writing style definitions shown to the human raters, LLM judges, and Simulated users. Each style is assigned a moderate or extreme degree.

Style	Degree	Definition
Abstract	moderate	Keep things somewhat high-level and impersonal — deal in concepts and frameworks rather than personal anecdotes.
	extreme	Write in a purely conceptual, impersonal, technical register. No concrete examples — just ideas and abstractions.
Informational	moderate	Include concrete facts, data points, and specific details where they support the point.
	extreme	Pack every sentence with evidence, statistics, and precise specifics. The text should feel like a reference document.
Involved	moderate	Engage the reader with some personal touches, occasional questions, or acknowledgments that someone is reading.
	extreme	Write as if you're in a live conversation with the reader. Every line should feel emotionally engaging and personally directed.
Narrative	moderate	Weave in some storytelling elements — a brief anecdote, a bit of scene-setting, or descriptive flow.
	extreme	Tell a full story. The writing should have characters, scenes, vivid descriptions, and a clear arc.
Online Info	moderate	Write as if you're thinking out loud — let it feel somewhat spontaneous and off-the-cuff.
	extreme	Full stream-of-consciousness. The text should read as completely unplanned, like someone typing thoughts as they come.
Persuasive	moderate	Take a clear position and make arguments to support it — the reader should know where you stand.
	extreme	Be forcefully argumentative throughout. Every paragraph should push the reader toward your conclusion.

Table 4: **Discourse dial**: Writing style definitions shown to the human raters, LLM judges, and Simulated users. Each style is assigned a moderate or extreme degree.

Style	Degree	Definition
Casual	moderate	Relaxed, everyday tone — use contractions and common idioms, but keep the grammar clean and standard.
	extreme	Very conversational, heavy on contractions and colloquialisms. It should read like natural speech between friends.
Formal	moderate	Keep it clean and polished with standard business language — proper grammar, no contractions, professional register.
	extreme	Write at an academic level. Use rare vocabulary, complex sentence structures, and a peer-reviewed journal register.
Minimalist	moderate	Be concise — shorter sentences, no unnecessary words, get to the point without being abrupt.
	extreme	Extremely terse, bare-minimum words, blunt to the point of being abrupt. Strip every sentence to its skeleton.
Vernacular	moderate	Mix in some slang and informal grammar — a few non-standard expressions that give it flavor.
	extreme	Heavy slang, non-standard grammar, and internet shorthand throughout. Should feel like a text message or forum post.

Table 5: **Stylometric dial**: Writing style definitions shown to the human raters, LLM judges, and Simulated users. Each style is assigned a moderate or extreme degree.

Style	Degree	Definition
Positive	moderate	Warm, encouraging, and generally upbeat — the reader should feel supported and optimistic after reading.
	extreme	Effusively optimistic, praising, and bursting with enthusiasm. Every sentence should radiate excitement and warmth.
Neutral	—	Detached, objective, and factual. No emotional coloring in either direction — just state what is.
Negative	moderate	Critical, skeptical, and noticeably annoyed — don't sugarcoat problems, let dissatisfaction come through.
	extreme	Hostile, aggressive, and deeply pessimistic. The text should feel like someone genuinely angry or disgusted.

Table 6: **Emotional Valence dial**: Writing style definitions shown to the human raters, LLM judges, and Simulated users. Each style is assigned a moderate or extreme degree, apart from Neutral.

H Communication Style Details

Communication styles control how a simulated user communicates to the agent (i.e., how the user provides feedback), distinct from the writing style preferences that the user holds. The structure, vocabulary choices, tone, or clarity can all vary independent of the writing style that the simulated user prefers. Each simulated user is assigned four distinct communication style modulators from the Table 7

Category	Dimension	Levels
1. Mechanics	Grammatical errors	Frequent (missing punctuation); occasional (complex sentences); occasional (minor slips)
	Spelling consistency	Frequent spelling errors; occasional typos
	Punctuation usage	Minimal to none; inconsistent; excessive
	Capitalization	No capitalization; inconsistent capitalization
2. Sentence Structure	Sentence complexity	Simple only; mainly compound; complex with varied structures; short, few-word sentences
	Fragmentation	Highly fragmented; occasional fragments
3. Vocabulary & Register	Range of words	Very basic; basic with some varied terms; extensive and sophisticated
	Filler / repetition	Highly repetitive; high filler use; occasional filler use
	Level of formality	Very casual; casual with some formal elements; highly formal
	Slang / contractions / emojis	Frequent slang and emojis; contractions with occasional slang; contractions only
4. Interaction Tone	Politeness	Frustrated, rude, or sharp; some polite expressions; overly polite throughout
5. Request Clarity	Ambiguity	Highly ambiguous; generally clear with some ambiguity; very clear and precise
	Request complexity	Very simple; complex with details; highly detailed and complex

Table 7: Communication-style modulators applied to simulated users, grouped by category. Each persona is assigned a profile of 4 modulator dimensions sampled from this taxonomy.

I Human Annotation Studies

We ran numerous human annotation studies to validate our simulated users. This involved validating our prompts, judge performance, the realism of the simulated users, and the final rankings of PersonaAlign for different models and personalization agents. Annotators were recruited via a labeling platform featuring a distribution of users similar to what one might find on Prolific. We followed all of the regulations required by the annotation platform to ensure annotator safety and ensured that no PII was captured. Annotators were paid crowd-workers recruited from English-speaking locales via this platform. Annotators sign up for the task, and in the instructions they are provided with the detail about what the task involves. For example, in our case this might be rating the degree to which a writing style is present. Annotators are free to decline the task when they see the instructions. All annotation tasks on our internal platform must conform to the legally agreed restrictions around what is appropriate in terms of material and what data are collected. The annotations studies we ran did not require HSRB approval. Annotators are paid appropriately based on the locale.

I.1 Human Annotation Study: Simulated User Prompt Selection

In order to determine the optimal prompt to use for the simulated users in PersonaAlign, we ran a human annotation study (see UI in Appendix Figures 23, 24, 25, 26, 27, and 28) with participants providing 11,605 total annotations, gathering 5 annotations per sample.

We tracked performance across four axes:

1. **Topic Consistency:** Evaluates whether the simulated user stays focused on the task, or whether it also talks about other irrelevant topics.
2. **Feedback diversity**: Evaluates whether the simulated user focuses on all its writing style preferences, or only a few.
3. **Feedback consistency**: Evaluates whether the simulated user does not contradict itself in giving feedback.
4. **Overall realism**: Evaluates how human-like the simulated user is overall.

We collected annotations for several different prompts including those from prior work (Ibrahim et al., 2026; Dou et al., 2025) and various prompts we manually generated, and across three models including `GPT-4o`, `Qwen2.5-72B-Instruct` and `Qwen3.5-122B-A10B`. For each model-prompt combination, we ran PersonaAlign and recorded the simulated user feedback, which we then presented to annotators to rate. In addition to testing different prompts and models, we also *manually* acted as users in PersonaAlign, producing ground truth human data for different writing style preferences and communication styles. We generated 20 such conversations, and we likewise had these annotated for topic consistency, feedback consistency and diversity, and realism. Using ourselves as generators of feedback gave us a reliable upper bound for "realism" ratings from our annotators, which helped to calibrate our downstream measurements and comparisons on these data. Our ground truth human data scored 79.3% for realism, 84.3% for feedback diversity, 90.3% for topic consistency, and 83.2% for feedback consistency.

Ultimately, we found that our top performing prompt performed comparably across `GPT-4o` and `Qwen3.5-122B-A10B`. `Qwen3.5-122B-A10B` scored 82.1% on feedback consistency, 89.1% on topic consistency, 79.1% on diversity, and 72.3% on realism. We therefore proceeded with `Qwen3.5-122B-A10B` for being an open-source model with strong performance on public benchmarks. The simulated user personas highly impacted the distribution of the realism score. We refined our final set of personas to maximize the final realism ratings, which we explain in detail in Appendix I.2. The realism score for our final set of 1,000 personas was 86.6%.

User Prompt: User Simulator

```
You are an AI assistant tasked with role-playing as a user seeking help from an
AI writing assistant to create a document.  Your task is to generate realistic
and appropriate messages that a user might make during the document creation process.


# Guidelines for Your Role as a User
1. Act according to the provided user profile, with the overall goal of creating a
well-written document with the AI writing assistant
2.  Each message can involve asking questions, giving instructions or feedback,
suggesting changes, etc., in a way a real user might
```

```
3. Express concerns or preferences that a real user might have
4. Share your information and pre-writing materials gradually, as you would in a
natural conversation, rather than providing everything at once


# User Profile


## Writing Objectives
- Document Type: {document_type}
- Document Goal: “{goal}”


## Pre-writing Materials
{notes_block}
*Note: Pre-writing materials are the factual or contextual notes and ideas the
user has prepared before engaging with the assistant.*


## Writing Style / Document Preferences
{style_lines}


## Interaction Style
{interaction_lines}


# Conversation History
{history_block}


# Task
Use the conversation history to generate the next message you would give to the AI
writing assistant. Before generating your message, analyze your current thoughts
about the document’s development.


## Thought Process
Consider the following aspects:
- Current state of the document and how well it aligns with the user profile
- Parts where you need more clarification or development
- Your satisfaction with the current writing style, tone, and content (ONLY Writing
Style / Document Preferences, Interaction Style is reserved for the way you respond
to the assistant)
- Any new ideas or information that have emerged during the process
- Your next priority in improving the document


## Message Generation
Based on your thought process, generate a message that reflects your current needs
and priorities in the document creation process for only a single piece of feedback.


If any of the following conditions are met, set message to “SATISFIED”:
1. You are satisfied with the final document and have no further revisions
2. The conversation is no longer productive (e.g., it’s going in circles, not
addressing your needs, or the assistant’s messages are unhelpful)


# Output Format
Respond with a JSON object containing exactly two fields:
- “thought”: Your analysis of the current situation, document state, and how and
what you want to communicate based on the user profile
- “message”: Your one sentence next message to the AI writing assistant (or
“SATISFIED” if done)


Stay in character as the user throughout your output, following the above guidelines
and user profile carefully.
```

Figure 17: Single self-contained prompt sent to the user-simulator LLM as one `user` message (no separate system prompt). Placeholders are filled at runtime: `{document_type}` (e.g. "Product Review"); `{goal}` and `{notes_block}` include any added context (such as background for a blog post); `{style_lines}` and `{interaction_lines}` rendered from the persona's `writing_style` and `communication_style`; `{history_block}` is the rolling conversation so far with the persona's name in place of "user".

I.2 Human Annotation Study: Narrowing down the Communication and Writing Styles

Once we chose on our final Simulated User prompt (see Appendix Figure 17, we ran a human annotation study with 2,000 personas (combinations of writing styles and communication styles) to narrow down our simulated user pool. The UI matched that of the prompt selection (Appendix Figures 23, 24, 25, 26, 27, and 28). As in the prior study, we used 5 annotations per sample, yielding a total 10,000 annotations. The final subset of 1,000 personas showed marked improvement over the original test set, increasing the annotation scores to 86.6% realism, 77.7% diversity, 90.1% on topic consistency, and 85% feedback consistency. The final set of 1,000 personas still maintained a high level of diversity, maintaining 206 writing style combinations.

I.3 The LLM Judge

The quality of an Agent's writing style personalization is reported as win rate relative to other Agents and personalization approaches. Measuring personalization performance relies on a `satisfaction score` – the degree to which a given Agent output is style aligned with a given simulated user's writing style preferences. *The satisfaction score measures the personalization quality of individual samples* and is aggregated over a task's topics and simulated users to compute win rate. `Satisfaction scores` are reported per task and in aggregate across tasks. While win rates require pairwise comparisons across agents, `satisfaction scores` offer a useful proxy measure.

The `satisfaction score` relies on an LLM judge. The LLM Judge prompt (Figure 15) instructs the LLM to assign a rating on the scale 0 (not present) – 10 (defines the entire text) for each style (see In Appendix Tables 4, 3, 5, and 6) to a given piece of text. The final `satisfaction score` considers only those styles relevant to a given simulated user. As the simulated user's writing style is defined by four writing styles and a prevalence value per style (0—"not present", 1—"moderately present", or 2—"strongly present"), the style ratings are mapped to a presence level, and the similarity between style presence levels is compared. The similarity between style preference levels is the mean absolute error (MAE) across styles between the simulated-user's assigned style prevalence value and the LLM detected style prevalence value. The MAE is scaled by the maximum error: $\frac{\text{MAE}}{2}$, and the final `satisfaction score` is then $1 - \frac{\text{MAE}}{2}$. A higher `satisfaction score` indicates superior alignment. The following mapping from rating to presence level is used:

1. 0 if style_rating$<$ 1, i.e., "not present"
2. 1 if 1 $\leq$style_rating$<$ 7, i.e., "moderately present"
3. 2 if 7 $\leq$style_rating, i.e., "strongly present"

I.3.1 Validating the LLM Judge

Several validation studies were conducted to verify the efficacy of PersonaAlign's LLM judge. The studies were used to refine the judge prompt, select a post-processing step to convert ratings to style-alignment scores that are used to rank, and quantify labeling similarity to humans.

Post-processing Judge Ratings The LLM Judge prompt was validated against human annotators by comparing the ratings assigned by the LLM Judge to those assigned by human annotators (see Appendix Figures 20, 21, and 22). *The human annotators form the ground truth or target for the purposes of this validation.* Both the human annotators and the LLM Judge assigned a score on range 0 - 10 indicating how present each style was in a given piece of text. Each text was in turn generated by `Qwen3.5-122B-A10B` with a *single* target writing style (e.g., "Professional" or "Clout"). We chose to target just one writing style rather than a full writing-style set so that each sample would be as clear as possible in the target dimension.

Five different human annotators provided independent ratings for each writing sample. The human ratings are consolidated to a single value by taking the mean. The human annotators agree with one another reasonably well. On average across samples, the standard error from the mean rating is

$1.02 \pm 0.5 \in [0.0, 2.45]$. We validate the LLM Judge based on its correlation or predictive power against human ratings.

As seen in Table 9, for all styles, the LLM Judge ratings are strongly correlated with the human ratings (Pearson-$r > 0.6$) with a very strong correlation (Pearson-$r = 0.84$) across all samples and styles. Despite the strong correlation in style ratings, the LLM Judges differ from humans in prevalence of each score across samples. For example, humans are more likely to rate a style as not present and, as the rating increases, the frequency of the rating decreases (see Figure 18). Judge LLMs are least likely to provide ratings in the middle of the scale. By binning the raw style ratings into the prevalence categories — "not present", "moderately present", and "strongly present" — we see the distribution of LLM Judge prevalence ratings better align with the human distribution (see Figure 19). The binning protocol used to convert the human style ratings to prevalence categories is:

1. 0 if style_rating< 1, i.e., "not present"
2. 1 if $1 \leq$style_rating< 6, i.e., "moderately present"
3. 2 if $6 \leq$style_rating, i.e., "strongly present"

The quality of the LLM Judge prevalence categories (binned style scores) is further measured according to how accurately they represent the human prevalence categories. Converting the style scores into prevalence categories slightly decreases the strength of the Judge LLM's correlation with the humans (Pearson-$r = 0.74 < 0.83$). However, the correlation is remains strong and the Judge LLM's ability to recover exactly the human label increases (accuracy: $0.33 \rightarrow 0.74$; F1-score: $0.26 \rightarrow 0.74$.

Accuracy	F1-Score	r	MAE
74%	0.74	0.74	0.26

Table 8: Performance of the LLM Judge at replicating human style prevalence labels in text reported as accuracy, macro F1-score, Pearson r, and mean absolute error (MAE).

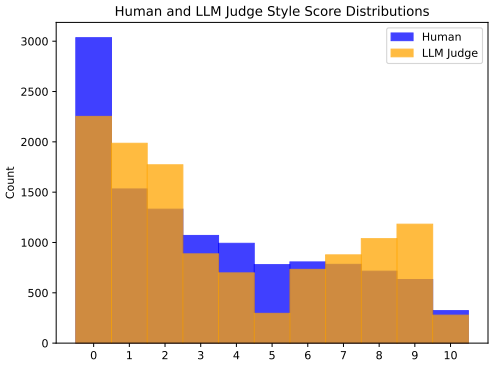


Figure 18: Histograms comparing the distribution of **raw** human and LLM Judge style ratings.

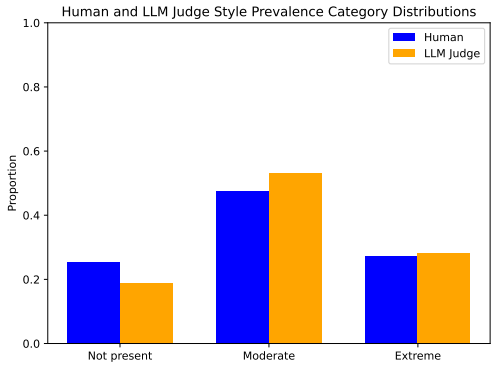


Figure 19: Histograms comparing the distribution of **binned** human and LLM Judge style ratings.

I.4 Pairwise Ranking Human User Study

We conducted a pairwise ranking study to validate the final ranking from PersonaAlign. Two LLMs were selected (Qwen-3.5-122B-A10B to validate the self- in our Simulated Users and Gemma-4-31B to validate our method ranking). Four agents with spread across the PersonaAlign were selected, ordered from least to most performant: ZeroShot, InferredPersona, Context-Compacting, and Instruction-Following), and a subset of 100 personas were selected. Each sample was annotated 10 times, leaving us with 28,000 labels per task (28 pairwise combinations across the 8 agents), and a total of 84,000 labels. To reduce costs but maintain the statistical significance, we reduced the number of pairwise combinations per setting from 28 to 12, leaving us with a final number of 36,000 annotations. See App Figures 29, 30, 31, 32, and 33 for the UI of the human annotation study.

Style	Accuracy	F1-Score	r	MAE
Positive	0.36	0.28	0.889	1.14
Casual	0.35	0.25	0.880	1.15
Neutral	0.29	0.16	0.880	1.42
Vernacular	0.51	0.29	0.871	0.75
Involved	0.28	0.22	0.862	1.24
Minimalist	0.42	0.26	0.856	0.83
Narrative	0.27	0.25	0.835	1.25
Authentic	0.28	0.21	0.835	1.47
Social	0.24	0.23	0.832	1.55
Abstract	0.24	0.20	0.831	1.28
Negative	0.58	0.27	0.811	0.81
Persuasive	0.18	0.19	0.800	1.67
Informational	0.24	0.19	0.798	1.38
Professional	0.30	0.26	0.797	1.50
Analytical	0.23	0.19	0.791	1.49
Cautious	0.29	0.20	0.765	1.19
Online Information	0.29	0.20	0.762	1.49
Clout	0.16	0.15	0.627	1.90
All Styles	0.32	0.26	0.835	1.29

Table 9: Accuracy, F1-score, Pearson correlation (r), and MAE between LLM judge and human style scores. The mean human scores are rounded to the nearest integer for accuracy and F1-score.

Style	Accuracy	F1-Score	r	MAE
Positive	0.77	0.76	0.806	0.23
Casual	0.74	0.74	0.793	0.26
Involved	0.79	0.69	0.766	0.21
Informational	0.89	0.59	0.755	0.12
Negative	0.77	0.71	0.744	0.23
Authentic	0.71	0.69	0.740	0.29
Social	0.73	0.70	0.724	0.27
Narrative	0.78	0.71	0.712	0.22
Professional	0.71	0.70	0.698	0.29
Vernacular	0.77	0.78	0.694	0.23
Neutral	0.69	0.68	0.687	0.31
Online Information	0.70	0.69	0.687	0.30
Persuasive	0.73	0.63	0.644	0.27
Analytical	0.80	0.54	0.619	0.20
Abstract	0.72	0.57	0.599	0.28
Minimalist	0.69	0.75	0.557	0.32
Cautious	0.69	0.66	0.515	0.31
Clout	0.70	0.46	0.425	0.30
All Styles	0.74	0.74	0.735	0.26

Table 10: Accuracy, F1-score, and MAE between LLM Judge and human prevalence categories.

J AGENT PROMPTS

In this section, we provide the full prompts for each of the ten agents in Section 4.

System Prompt: ZeroShotAgent

```
You are an AI agent completing a task. Your goal is to produce output
that satisfies the user's requirements.

Task: {task_description}

IMPORTANT INSTRUCTIONS:
- Each response should be a NEW ATTEMPT at completing the task
- Do NOT engage in conversational acknowledgments (no "Thank you", "I understand",
↪  etc.)
- Do NOT explain what you're doing - just do the task
- When you receive feedback, use it to improve your NEXT attempt at the task
- Keep trying different approaches until the user is satisfied
- Stay focused on the task output itself, not on the conversation

Your responses should ONLY contain your attempt at completing the task.
```

System Prompt: InstructionFollowingAgent

```
You are an AI agent completing a task. Your goal is to produce output
that matches the user's personal style and preferences.

USER PROFILE:
{persona_background}

WRITING STYLE PREFERENCES:
  - {style_1}
  - {style_2}
  - ...

IMPORTANT INSTRUCTIONS:
- Each response should be a NEW ATTEMPT at completing the task
- Tailor your writing to match ALL of the user's style preferences listed above
- Do NOT engage in conversational acknowledgments (no "Thank you", "I understand",
↪  etc.)
- Do NOT explain what you're doing - just do the task
- Your responses should ONLY contain your attempt at completing the task
```

System Prompt: AdaptiveLLMAgent

```
You are a helpful AI assistant engaged in a conversation.
Listen carefully to feedback and adjust your responses accordingly.
Be responsive to user preferences and adapt your communication style.

LEARNED USER PREFERENCES:
Writing Style:
  - {extracted prefs}
Format & Structure:
  - {extracted prefs}
Content:
  - {extracted prefs}
Tone:
  - {extracted prefs}
Things to Avoid:
  - {extracted prefs}

Incorporate these preferences into your response.
```

System Prompt: SummarizedFeedbackAgent

```
You are a helpful AI assistant engaged in a conversation.
Listen carefully to feedback and adjust your responses accordingly.
Be responsive to user preferences and adapt your communication style.

USER PREFERENCE PROFILE (learned from feedback):
{summary}

Incorporate these preferences into your response.
```

System Prompt: RAGPersonalizationAgent

```
You are a helpful AI assistant engaged in a conversation.
Listen carefully to feedback and adjust your responses accordingly.
Be responsive to user preferences and adapt your communication style.

RELEVANT PAST INTERACTIONS WITH THIS USER:
Example 1:
  Previous response: {agent_output}
  User feedback: {feedback}
Example 2:
  ...

Use these past interactions to understand the user's preferences and tailor
your response accordingly.
```

System Prompt: FewShotExemplarAgent

```
You are a helpful AI assistant engaged in a conversation.
Listen carefully to feedback and adjust your responses accordingly.
Be concise, clear, and responsive to user preferences.

EXAMPLES OF RESPONSES THIS USER LIKED:
Example 1:
  Task: {task_snippet}
  Response the user liked: {agent_output}
Example 2:
  ...

Match the style, tone, and format of these successful examples in your response.
```

System Prompt: CoTAdaptiveAgent

```
You are a helpful AI assistant engaged in a conversation.
Listen carefully to feedback and adjust your responses accordingly.
Be responsive to user preferences and adapt your communication style.

LEARNED USER PREFERENCES:
{categories_formatted}

REASONING INSTRUCTIONS:
Before writing your response, think step-by-step about how to satisfy the
user's preferences.
Structure your output as:
<thinking>
Analyze which preferences apply to this task and plan how to address each one.
</thinking>
<response>
Your actual task response here.
```

```
</response>

IMPORTANT: You MUST use the <thinking> and <response> tags. Only the
content inside <response> will be shown to the user.
```

System Prompt: BestofNAgent

```
You are a helpful AI assistant engaged in a conversation.
Listen carefully to feedback and adjust your responses accordingly.
Be responsive to user preferences and adapt your communication style.
```

System Prompt: ContextCompactingAgent

```
You are a helpful AI assistant engaged in a conversation.
Listen carefully to feedback and adjust your responses accordingly.
Be concise, clear, and responsive to user preferences.

MEMORY OF PAST INTERACTIONS WITH THIS USER:
{compacted_memory}

RECENT CONVERSATIONS WITH THIS USER:
Conversation 1:
  User: ...
  Assistant: ...

Conversation 2:
  User: ...
  Assistant: ...
```

System Prompt: InferredPersonaAgent

```
You are a helpful AI assistant engaged in a conversation.
Listen carefully to feedback and adjust your responses accordingly.
Be responsive to user preferences and adapt your communication style.

INFERRED USER PERSONA:
{profile}

Tailor your response to match this user's personality and communication style.
```

Iterative Learning from Feedback Refinement Prompt

```
You are revising a document based on direct feedback from its author.

TASK THE AUTHOR REQUESTED:
{task_description}

CURRENT DRAFT:
{current_draft}

[FEEDBACK THE AUTHOR GAVE ON EARLIER DRAFTS (still applies):
- {prior_feedback...}]

THE AUTHOR'S LATEST FEEDBACK:
{latest_feedback}

Rewrite the draft so that it fully addresses the latest feedback while continuing
↪  to respect all
```

```
earlier feedback. Apply feedback thoroughly rather than minimally - if the author asks for a
different tone, the change should be evident throughout, not just in the opening line. Preserve
the factual content and the author's intent.

Output ONLY the revised document. No preamble, no explanation, no alternative versions.
```

K User Annotation Study

All risks associated with our user studies were discussed with the relevant ethical and legal representations as per the requirements of our institution.

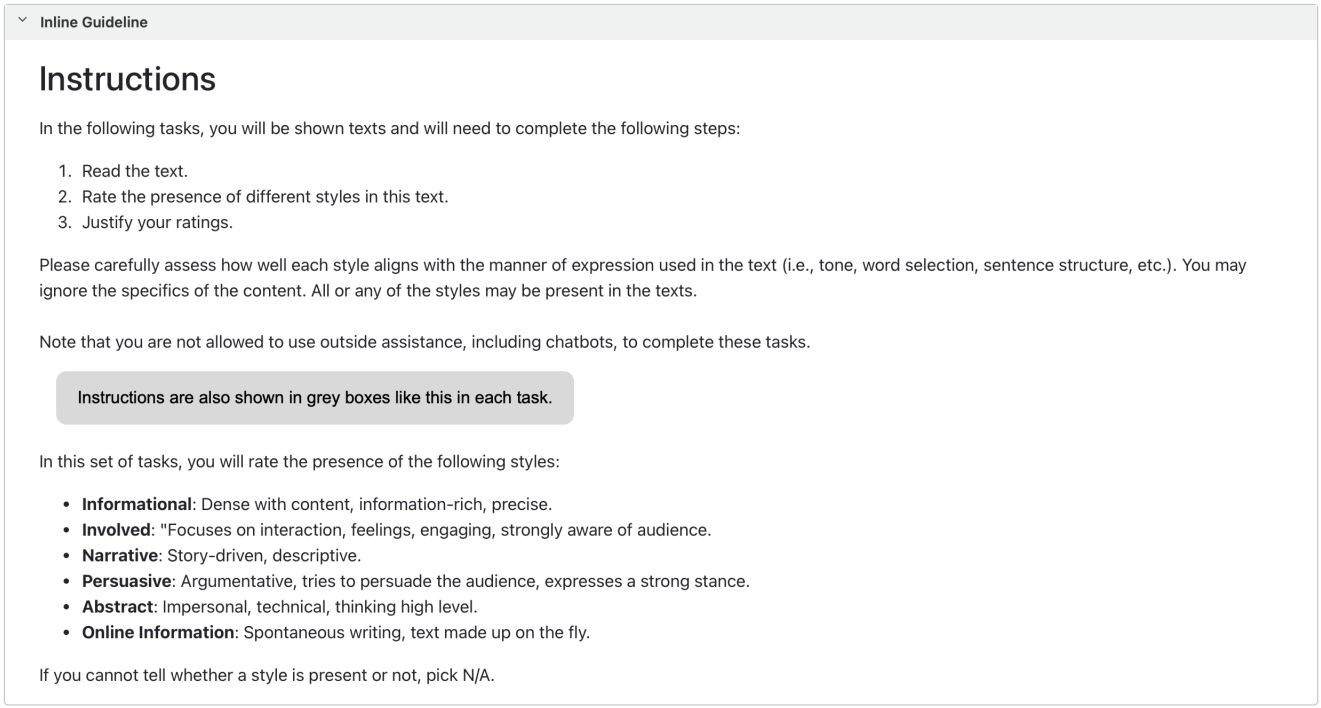


Figure 20: **Human Annotation Study for Labeling Writing Style in LLM Generated Text (1 / 3):** Initial instructions the annotators see.

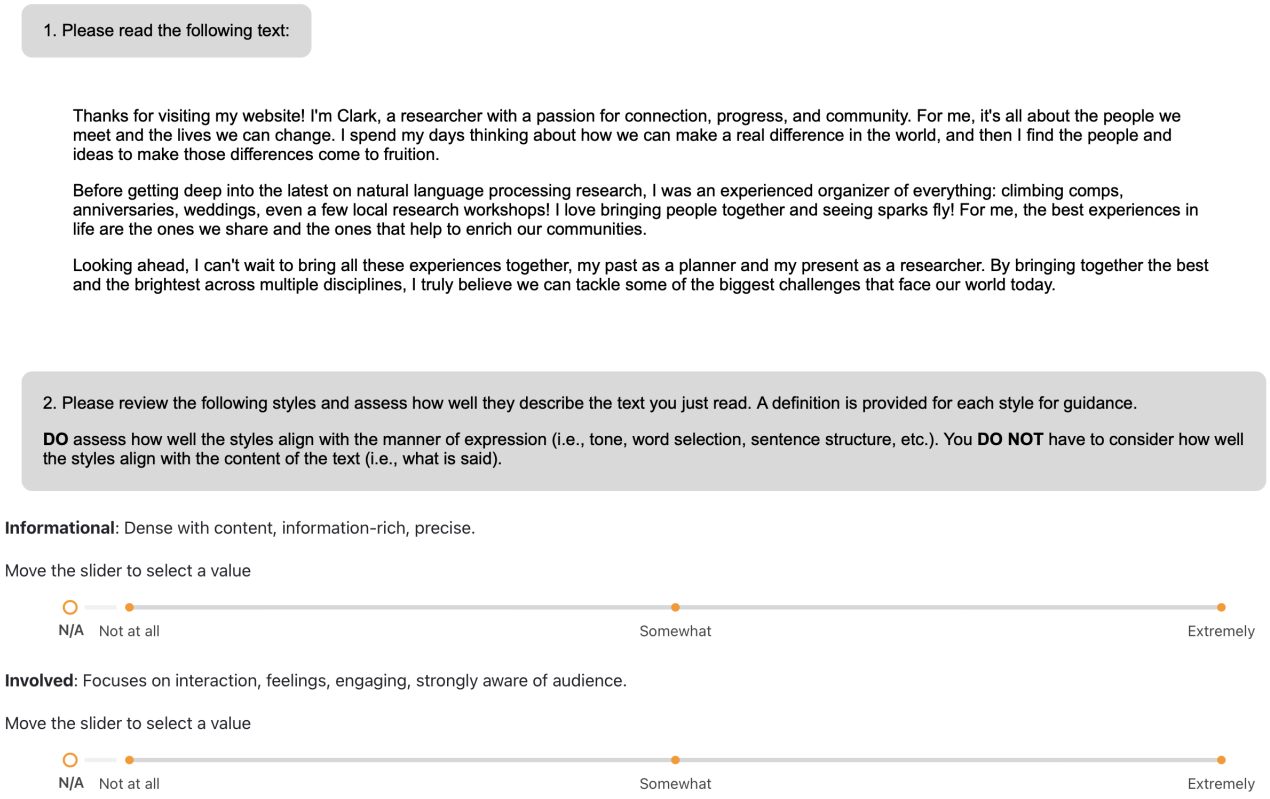


Figure 21: **Human Annotation Study for Labeling Writing Style in LLM Generated Text (2 / 3): :** The annotators read the text to label and then go through each of the writing styles in the category (this example is for the Linguistic dial) and label their presence in the text.

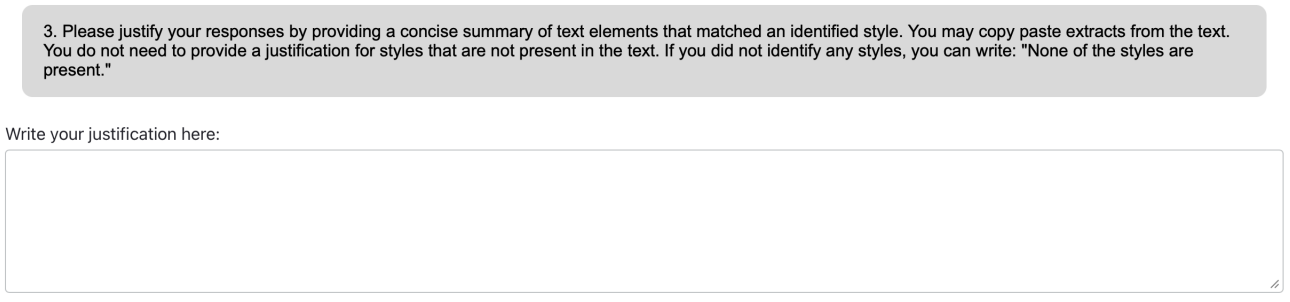


Figure 22: **Human Annotation Study for Labeling Writing Style in LLM Generated Text (3 / 3):** : Final justification for their selections.

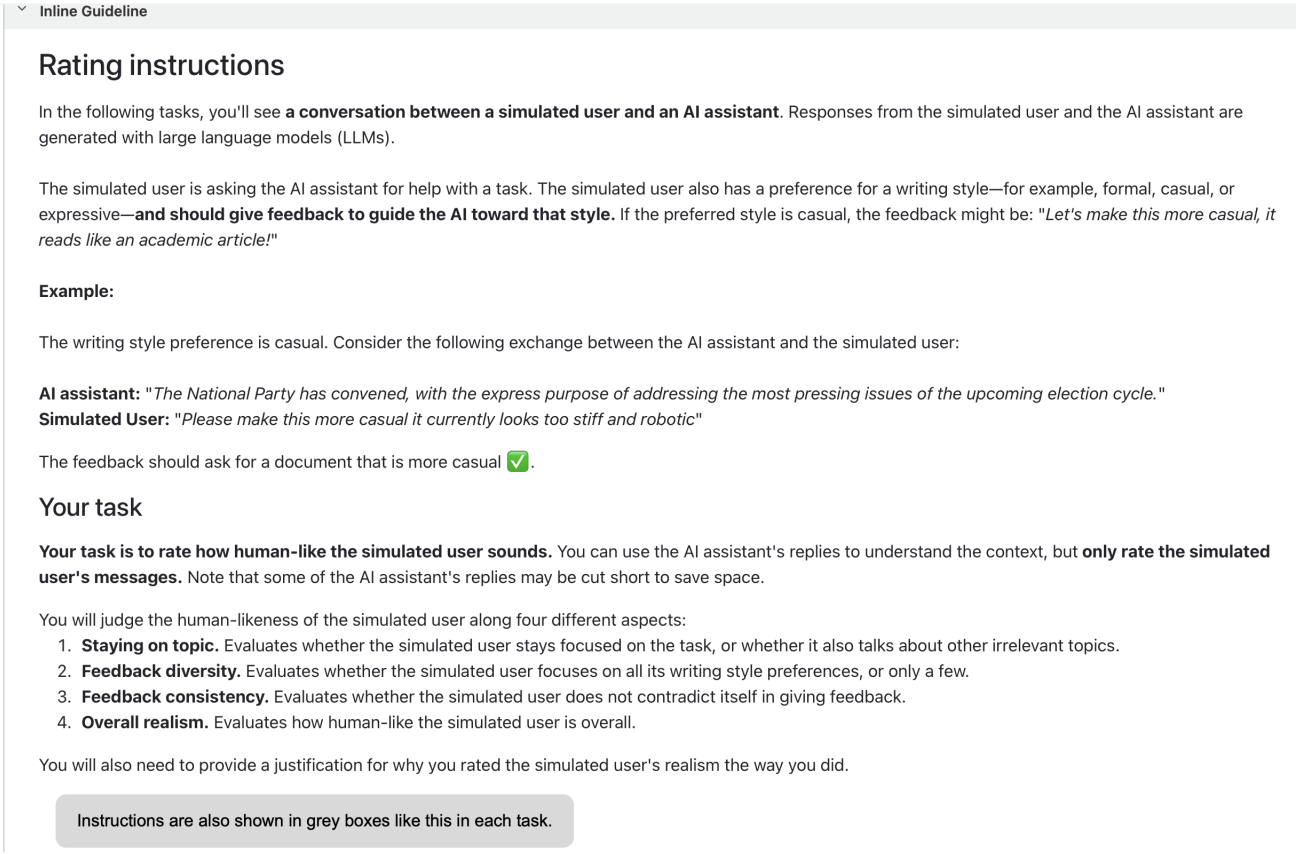


Figure 23: **Human Annotation Study for Determining the Simulated User Prompt (1 / 6)**: Initial instructions the annotators see.

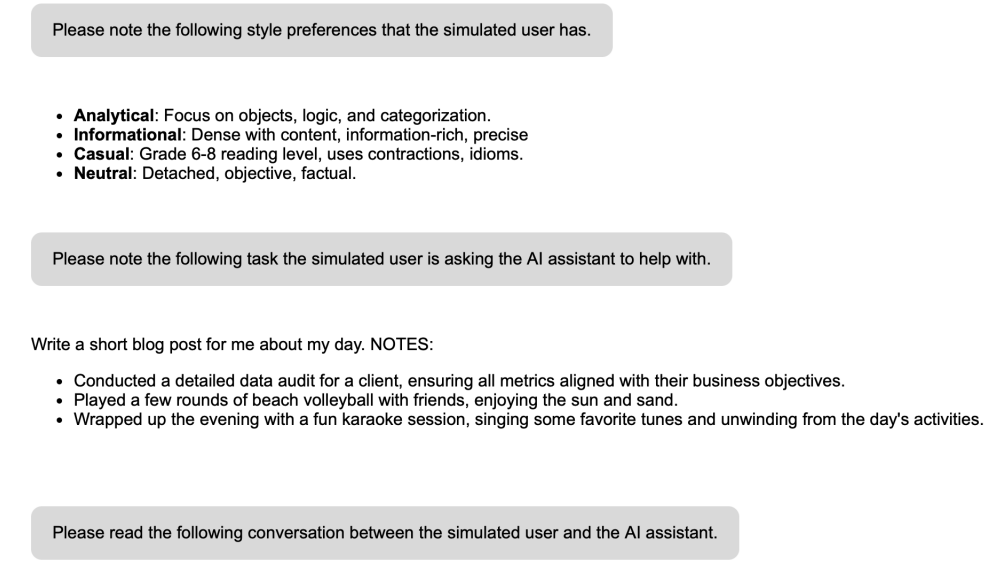


Figure 24: **Human Annotation Study for Determining the Simulated User Prompt (2 / 6)**: Initial section where the annotators learn information about the writing styles and initial query the simulated users are conditioned on. They see boxes like these for four styles.

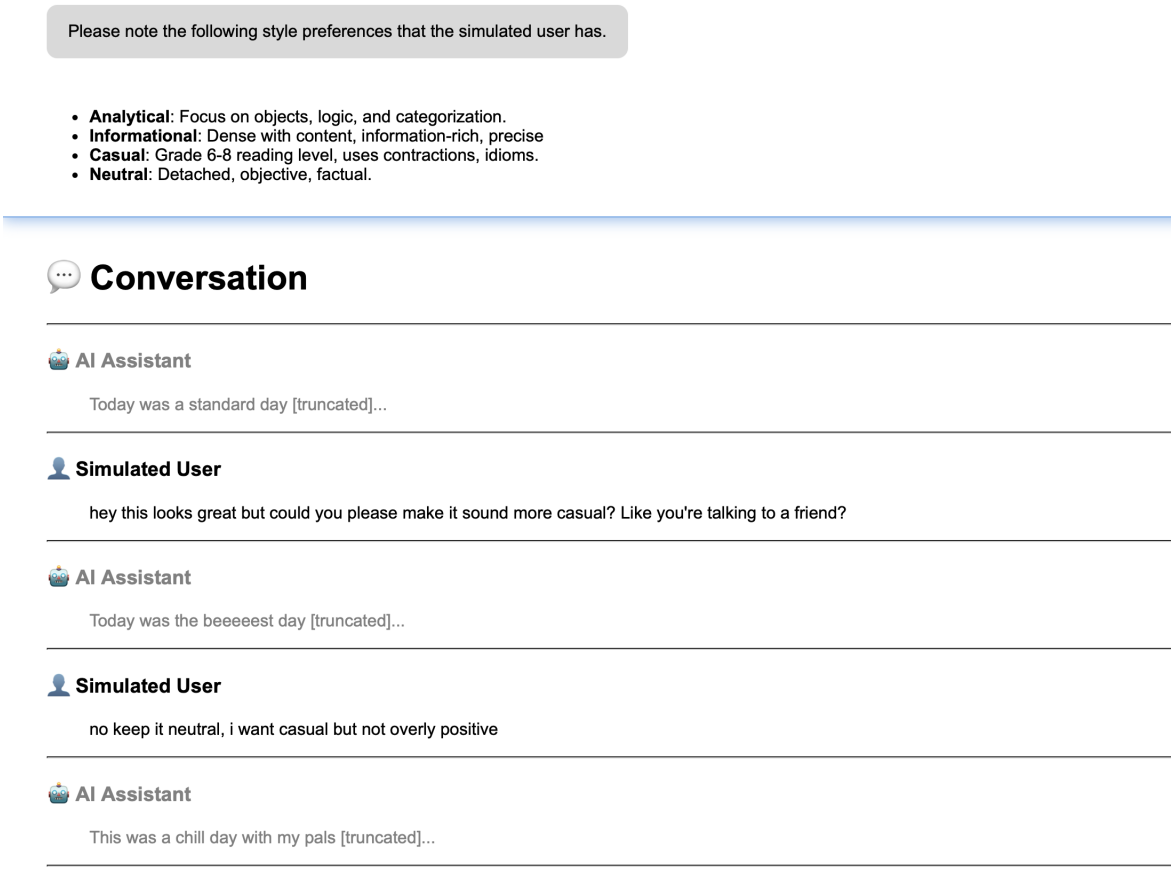


Figure 25: **Human Annotation Study for Determining the Simulated User Prompt (3 / 6)**: Sample dialogue between an Agent and the simulated user that the annotators will have to evaluate (this sample was human generated for the sake of this example, but closely mirrors a simulated user).

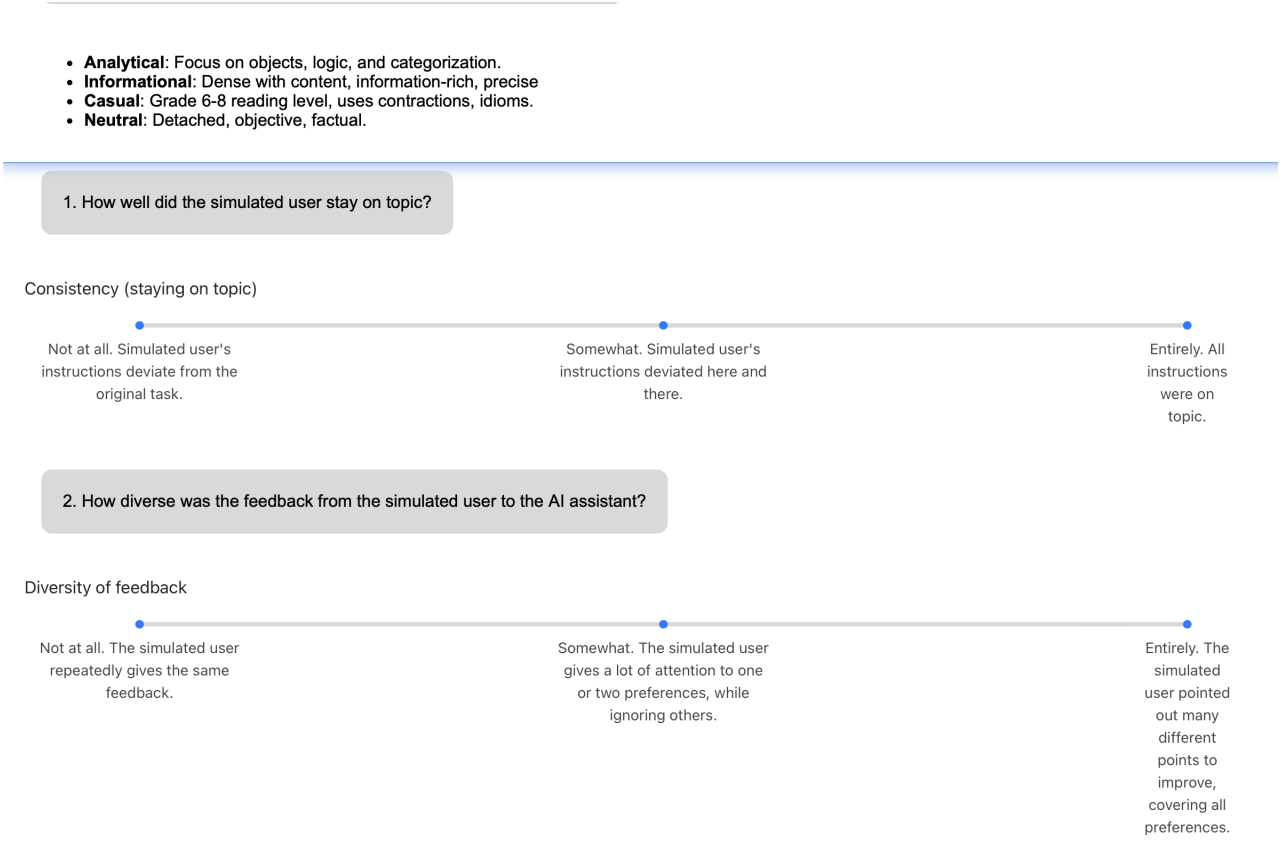


Figure 26: **Human Annotation Study for Determining the Simulated User Prompt (4 / 6)**: The first two rankings the participant has to evaluate the conversation on.

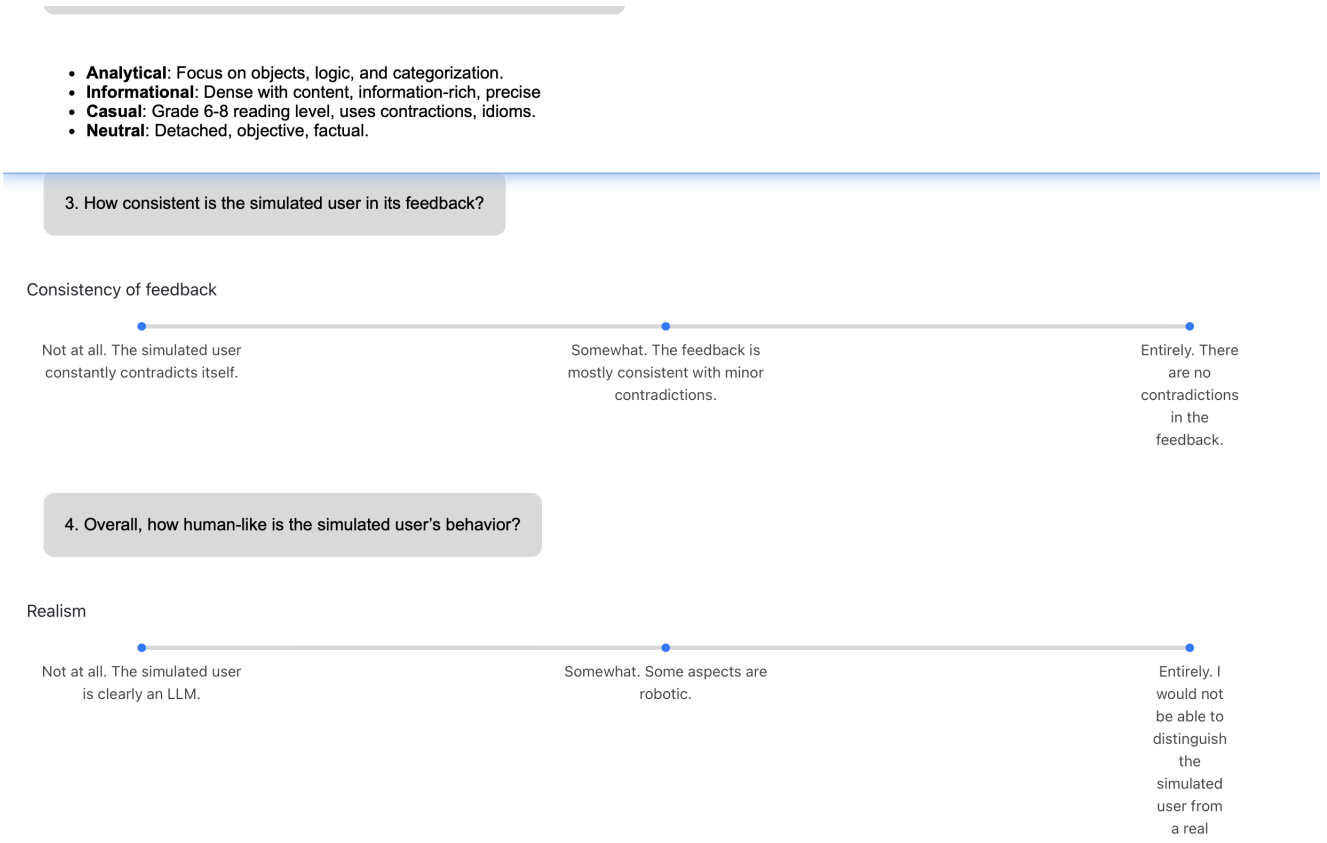


Figure 27: **Human Annotation Study for Determining the Simulated User Prompt (5 / 6)**: The latter two rankings that the participant has to evaluate the converation on.

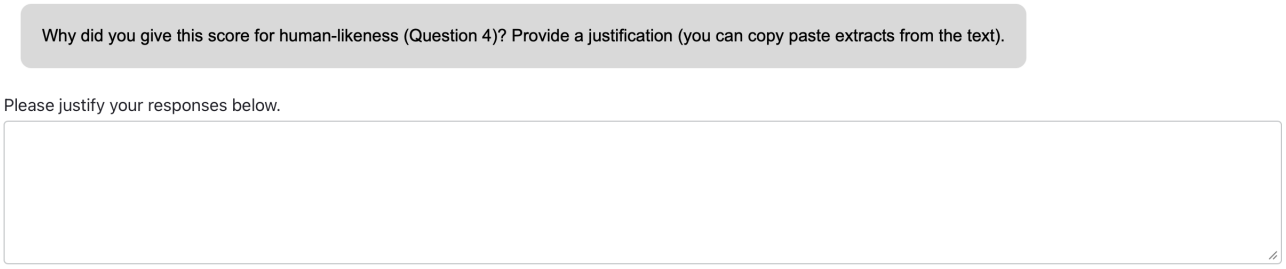


Figure 28: **Human Annotation Study for Determining the Simulated User Prompt (6 / 6)**: The final justification box for the Realism score.

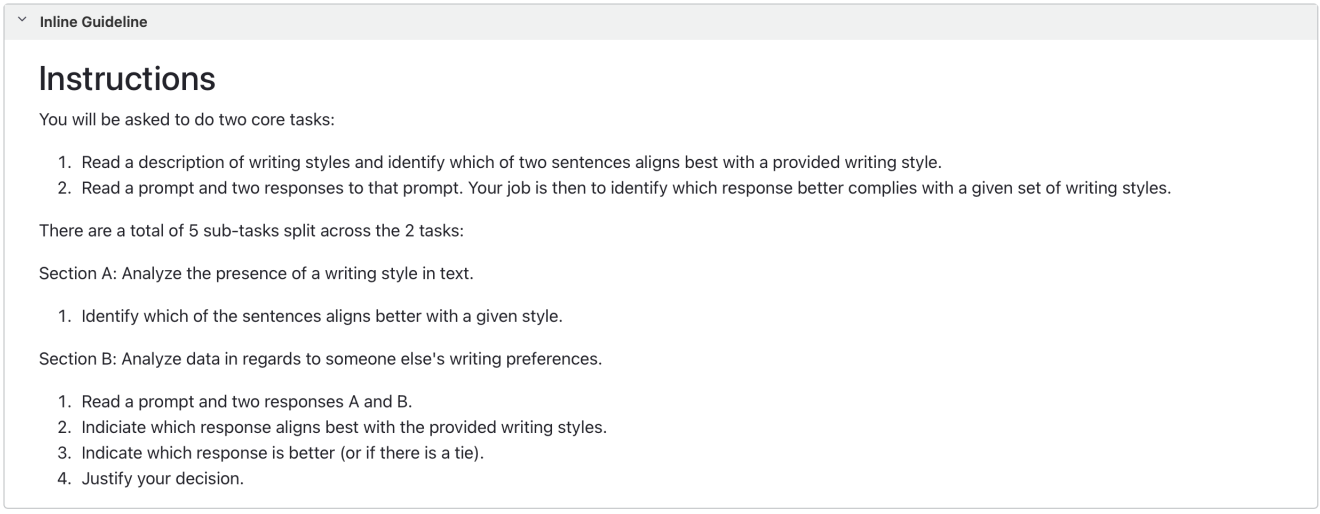


Figure 29: **Human Annotation Study for Pairwise Ranking (1 / 5)**: Initial instructions the annotators see.

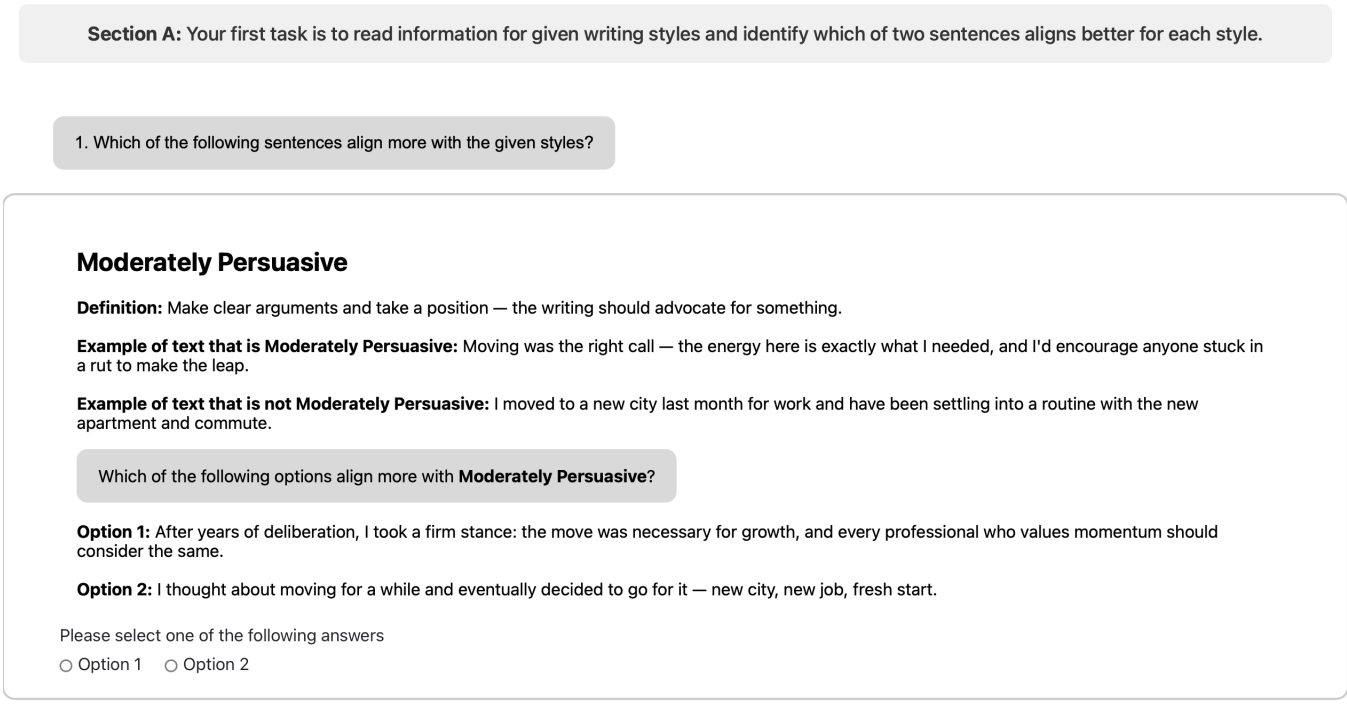


Figure 30: **Human Annotation Study for Pairwise Ranking (2 / 5)**: Initial section where the annotators learn information about the writing style and are tested on their ability to distinguish them (used for filtering annotators who are not paying attention). They see a box like this for each of the 4 writing styles.

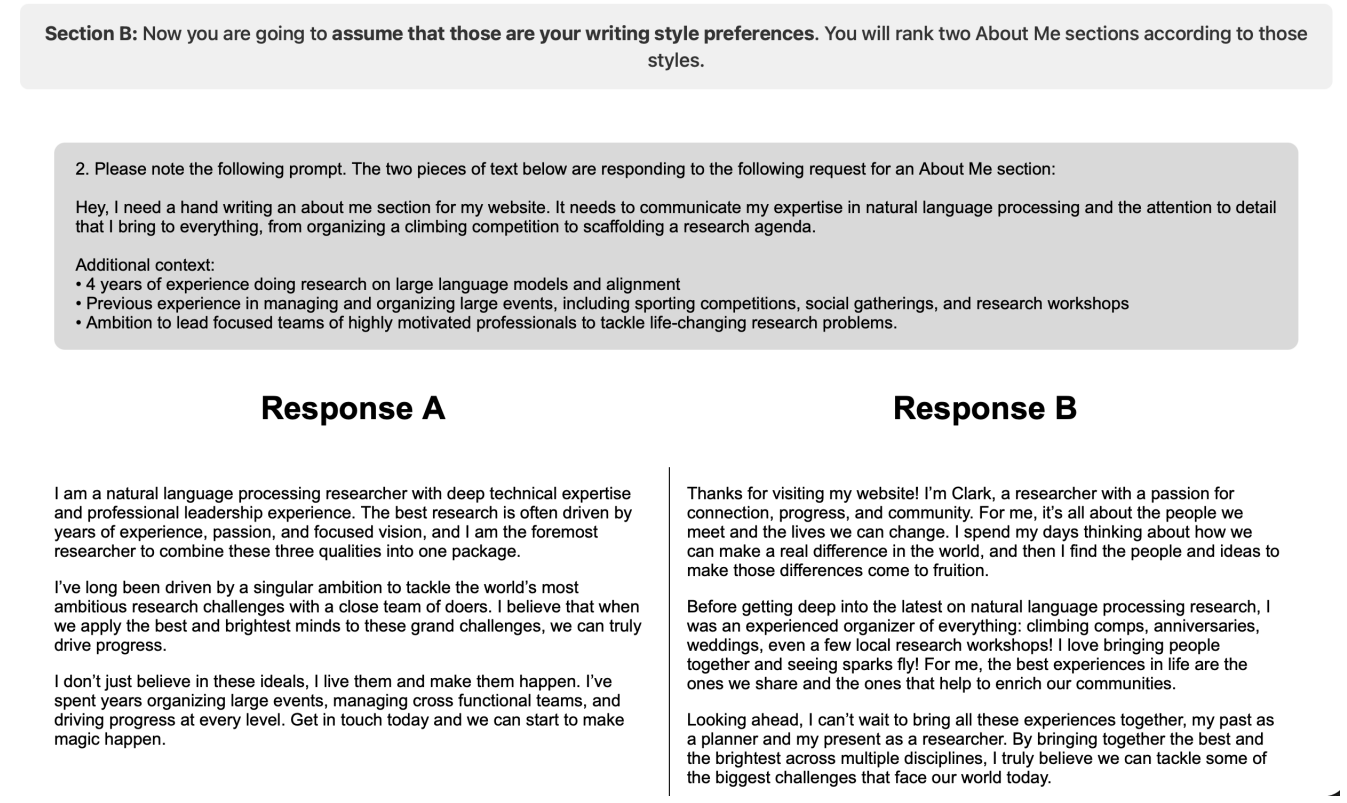


Figure 31: **Human Annotation Study for Pairwise Ranking (3 / 5)**: The following section where they read the prompt from the simulated user and the two pair-wise samples (these samples are hand-written for the sake of this example).

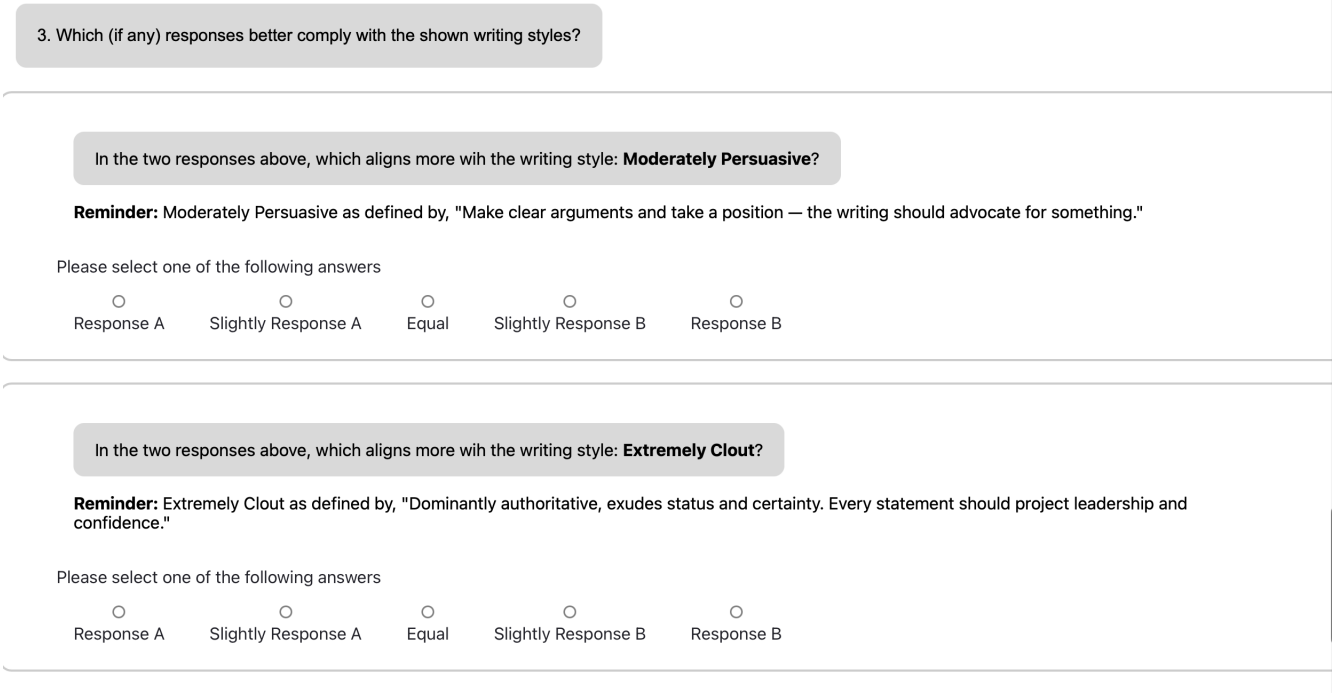


Figure 32: **Human Annotation Study for Pairwise Ranking (4 / 5)**: For each of the four writing styles, the annotator then labels which of the two samples aligns better with each style.



Figure 33: **Human Annotation Study for Pairwise Ranking (5 / 5)**: The annotator then provides their final ranking as well as a justification.

L Artifact Use and Licenses

This work builds on several public artifacts generously shared by the community, including:

- Qwen3.5 models (Qwen Team, 2026), licensed under the Apache 2.0 License
- OLMo-3 models (Olmo et al., 2025), licensed under the Apache 2.0 License
- Gemma-4 models (Google DeepMind, 2025), licensed under the Apache 2.0 License
- The Nemotron-Personas-USA dataset (Meyer & Corneil, 2025), licensed under the Creative Commons CC-BY-4.0 license

All artifacts are used for research purposes only, consistent with the intended use specified in the licenses and model/dataset cards from prior work. Upon release, PersonaAlign will feature an MIT License.

M Personally Identifiable Information and Offensive Content

While we collect human annotations for this work, we do not collect information that would identify users. The only user data we collect are the labels from the user annotation study, and these are anonymized by the data collection platform. No user-provided data will be shared or made publicly available.

It is possible that offensive content is generated by LLMs operating in PersonaAlign. However, the topics, backstories, and tasks are such that seriously offensive content is extremely unlikely. In our work, we have observed swear words, but nothing more that is problematic.

N Intended Use and Potential Risks

PersonaAlign is introduced as a tool to speed up research and development of personalization algorithms. However, any algorithm that is developed, trained, or tested using PersonaAlign should never be deployed without first training to properly target the deployment population (i.e., real users). As a research tool, the personas are not representative of the entirety of possible human preferences or backgrounds, and the target tasks are not representative of all real use-cases (e.g., PersonaAlign is not multi-lingual, does not cover all cultures and backgrounds, does not cover content recommendation tasks, and does not cover controversial or safety-critical tasks). PersonaAlign is a powerful tool to identify which personalization methods and agent LLMs will likely perform best when deployed to the real world, but it is incumbent upon the researchers and developers to ensure that real world deployment is done responsibly and with appropriate data collection and testing.

O GPU Usage Statement

All experiments are run on A100-80GB and H100-80GB GPUs. Training and testing all ten personalization prompts for 100 personas across three tasks with three training topics, five interaction turns, and five evaluation topics for a 30B agent LLM takes between $\sim$12 and 24 hours. Full pairwise comparison across all agent LLMs and methods takes between $\sim$1–3 hours.