Beyond Discrete Personas: Personality Modeling Through Journal Intensive Conversations

Sayantan Pal , Souvik Das , Rohini K. Srihari
State University of New York at Buffalo
Department of Computer Science and Engineering
spal5, souvikda, rohini@buffalo.edu

Abstract

Large Language Models (LLMs) have significantly improved personalized conversational capabilities. However, existing datasets like Persona Chat, Synthetic Persona Chat, and Blended Skill Talk rely on static, predefined personas. This approach often results in dialogues that fail to capture human personalities' fluid and evolving nature. To overcome these limitations, we introduce a novel dataset with around 400,000 dialogues and a framework for generating personalized conversations using long-form journal entries from Reddit. Our approach clusters journal entries for each author and filters them by selecting the most representative cluster, ensuring that the retained entries best reflect the author's personality. We further refine the data by capturing the Big Five personality traits—openness, conscientiousness, extraversion, agreeableness, and neuroticism—ensuring that dialogues authentically reflect an individual's personality. Using Llama 3 70B, we generate high-quality, personality-rich dialogues grounded in these journal entries. Fine-tuning models on this dataset leads to an 11% improvement in capturing personality traits on average, outperforming existing approaches in generating more coherent and personality-driven dialogues.

1 Introduction

A conversation reflects the unique threads of a person's life experiences, thoughts, and personalities(Mairesse et al., 2007). However, many existing conversational systems(Ahmad et al., 2023; Welch et al., 2020) struggle to capture the richness of these tales, often reducing complex individuals to static, predefined personas(Blomkvist, 2002). Existing datasets like Persona Chat (PC)(Zhang et al., 2018), Synthetic Persona Chat (SPC)(Jandaghi et al., 2024), and Blended Skill Talk (BST)(Smith et al., 2020) use hardcoded personas. While these datasets have paved the way for more personalized dialogue systems(Kasahara et al., 2022), they

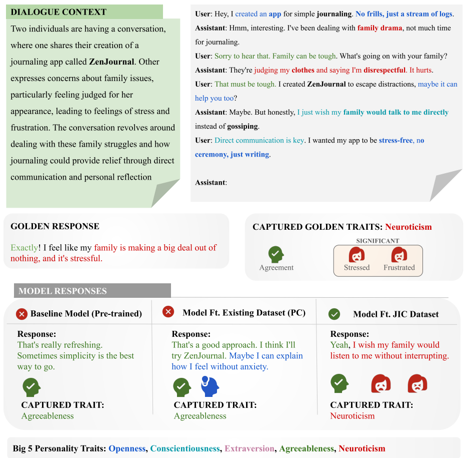

Figure 1: LLMs fine-tuned on our JIC dataset best aligns to the golden annotation capturing Personality traits compared to other models (image shows Ft. on Persona Chat). Detailed dialogue is shown in Fig. 2.

often fall short of capturing the dynamic and evolving nature of real human personalities(Allbeck and Badler, 2008), as shown in Fig. 1. Conversations generated from such static personas can feel repetitive(Zhang et al., 2020), shallow, and sometimes even contradictory(Nie et al., 2021), failing to engage the user truly. Our research seeks to fill this gap and transform this approach by moving beyond the constraints of discrete personas, instead embracing a model that captures the dynamic nature of personal identity(Schwartz et al., 2011). By leveraging long-form journal entries mined from platforms like Reddit—where individuals share their authentic, unfiltered life experiences, we ensured the preservation of personality traits, achieving greater depth and realism than static personas.

Personas are widely used to enhance(Li et al., 2016; Zhong et al., 2020) user representation and conversational flow by simulating human-like dialogue. Our analysis shows that existing datasets fail to capture the complexity of the Big Five personal-

Proceedings of the 31st International Conference on Computational Linguistics, pages 7055–7074
January 19–24, 2025. ©2025 Association for Computational Linguistics

ity traits: (O.C.E.A.N.)(Hurtz and Donovan, 2000; Azucar et al., 2018)—openness, conscientiousness, extraversion, agreeableness, and neuroticism resulting in less genuine interactions.

Creating a dataset that captures personality traits is labor-intensive, traditionally relying on significant human input for persona design, conversation generation, and validation. To overcome these challenges, we utilized large language models (LLMs), specifically LLaMa 3 70B[1](AI@Meta, 2024), for synthetic data generation with human-in-the-loop assessment. This personalizes AI systems and enhances human-AI interaction for more relatable and engaging conversational agents(CA)(Clark et al., 2019).

Our work introduces a novel method for creating a journal-based conversational dataset named Journal Intensive Conversations (JIC). **(1)** This process begins with data acquisition from Reddit. **(2)** We apply multi-step filtration strategies, using clustering algorithms to identify and retain the most representative journal entries per author. Additionally, we filter out dialogues that diverge significantly from the author's average Big 5 Personality Traits, ensuring better alignment. **(3)** We then use instruct-LLMs to generate journal-grounded conversations, ensuring the resulting dialogues remain true to the author's personality. **(4)** Finally, we demonstrate that fine-tuning state-of-the-art(SOTA) LLMs on our dataset enhances their ability to capture personality traits effectively in dialogue. Our code, data, and best models are publicly available.[2]

2 Related Work

Personality Traits in Conversational AI: In recent years, modeling personality traits(Liu et al., 2016; Caron and Srivastava, 2023; Saha et al., 2022) in conversational systems(Dušek and Jurčíček, 2016) has been an area of extensive research to make human-AI(Yang et al., 2024) interaction more personalized and engaging. Early attempts(Yamashita et al., 2023; Zhang et al., 2018) in this field used static, predefined personas to model users and produce goal-directed faithful(Jandaghi et al., 2024) conversations. While these systems improved personalization(Smith et al., 2020), they fell short of capturing the dynamic nature of human behavior(Schill et al., 2019; Pal et al., 2024), often reducing users to rigid attributes that limit dialogue adaptability. Recent advancements(tse Huang et al., 2024; Huang et al., 2023) have called for more sophisticated systems that reflect the evolving nature of human personality. Moreover, the emergence of large language models (LLMs) like GPT-3(Brown et al., 2020) and LLaMA(Touvron et al., 2023) offers new opportunities for generating more nuanced, personality-driven dialogues. Fine-tuning these models on personalized datasets enables them to exhibit a deeper understanding of individual traits, fostering more consistent and contextually appropriate interactions(Labruna et al., 2024). Other works have explored integrating psychological models (Azucar et al., 2018; Barlett and Anderson, 2012) like the Big Five (O.C.E.A.N. Model) into CA. However, significant challenges remain in accurately capturing and representing dynamic personality traits.

Personality Datasets and Challenges: Recent advancements in conversational datasets have highlighted the potential(Sun et al., 2022; Rashkin et al., 2019) and limitations(Hwang et al., 2023) of existing approaches(Zhong et al., 2020; Das and Srihari, 2024) to simulate human behavior. Static personas or scripted inputs, such as those found in existing datasets, limit their ability to capture human traits' evolving nature in dialogue. For instance, synthetic conversations often mimic human interaction but struggle to reflect persistent personality traits over time. More dynamic datasets(Jandaghi et al., 2024), like those generated using instruct-LLMs, aim to address this by leveraging tunable instructions(Su et al., 2022) to capture authentic conversations. Despite these advancements, creating datasets that genuinely capture the complexity of human personality remains a crucial area for further research and refinement.

3 Data Acquisition

We mined data from two relevant subreddits: *r/DiaryOfARedditor* and *r/Journaling*. These communities provided a rich source of personal narratives, allowing us to gather approximately 19,000 submissions from 1,372 unique authors.

3.1 Journal Data Scraping

We used the PullPush API[3] to collect data from Reddit by querying relevant subreddits using a pre-

[1]LLaMa 3 70B is an open-source model that allows for data generation without associated costs, making it a practical choice over models like GPT-4 despite its superior performance.

[2]LLaMAdelic, MistraMystic, Code and Data

[3]PullPush API

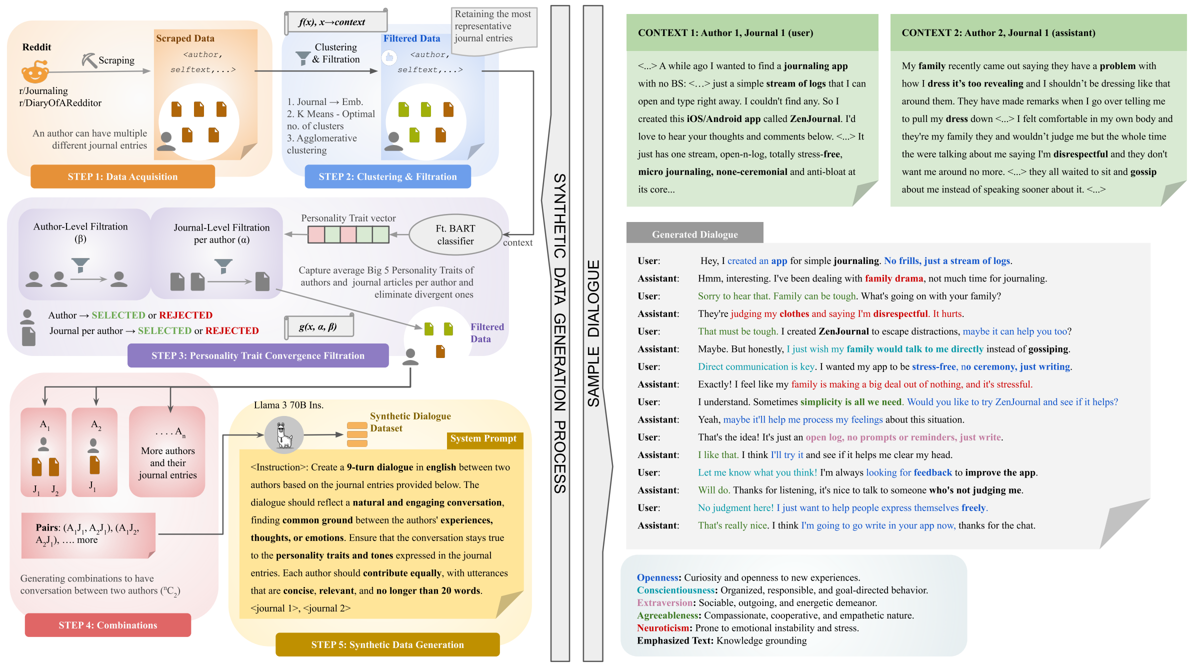

Figure 2: The synthetic data generation process is outlined in five distinct stages (left side). On the right side, we demonstrate how dialogues are generated from journal entries, highlighting the personality traits they reflect and align with. In Stage 3, where personality trait filtering is introduced, the initial values of the α and β parameters were set to None to allow extensive data generation before further refinement.

defined keyword list. The API retrieved submissions along with details like the author's username, submission text, creation date, title, and URL. After filtering out duplicates and entries with missing fields, we obtained 18,817 submissions, including 4,377 from *r/DiaryOfARedditor* and 14,440 from *r/Journaling*. These submissions were further refined using additional filtering criteria.

3.2 Synthetic Conversation Generation

We used the Groq API's[4] LLaMa 3 70B model to generate synthetic conversations from filtered journal entries. Given rate limitations, we selected 906 out of 1,372 unique authors, pairing them in all possible combinations. For authors with multiple journal entries, dialogues were generated for every entry combination. For instance, two dialogues were generated if Author 1 had two entries and Author 2 had one. This approach produced a total of 418,476 dialogues. The final turn often included superficial exchanges like "Bye" or "Have a nice day." To retain conversational depth, the last turn was removed, leaving 8-turn dialogues that better reflected meaningful interactions. Full details of the prompting strategy are provided in Appendix A.

[4] Groq API

The synthetic dialogues were evaluated using GPT4-o and human assessments; the results showed strong agreement, particularly with high Intraclass Correlation Coefficient(ICC) scores, indicating good consistency between LLM and human ratings. Detailed agreement scores in Appendix B.

4 Data Filtration Strategies

4.1 Prominent Journal Clustering and Retention

We employed a clustering strategy to retain the most representative journal entries for authors with multiple submissions. High-dimensional embeddings were generated using the *microsoft/deberta-large* model (He et al., 2021) to capture semantic content. K-Means clustering, validated with silhouette scores (Rousseeuw, 1987), was applied to identify optimal clusters. Additionally, agglomerative clustering (Müllner, 2011) was used to refine grouping, selecting the most prominent cluster. This ensured the dataset reflected each author's dominant themes for generating synthetic conversations. The filtration process is detailed in Algorithm 1.

4.2 Personality Trait Convergence Filtering

We refined the dataset to capture journal entries and authors with the most prominent and consistent

Algorithm 1 Prominent Journal Clustering and Retention

Input: Set of journal entries $\mathcal{J}$ per author, pre-trained sentence embedding model $\mathcal{M}$
Initialize: Embedding model $\mathcal{M}$ to encode journal texts into vectors $\mathbf{v}$
for each author $a \in \mathcal{A}$ **do**
  Encode journal entries $\mathcal{J}_a$ into high-dimensional vectors $\mathbf{v}_a = \mathcal{M}(\mathcal{J}_a)$
  Determine the optimal number of clusters k^* using K-Means clustering and silhouette score $S(k)$
  Apply Agglomerative Clustering with k^* to group journal entries into clusters $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{k^*}$
  Identify the largest cluster $\mathcal{C}_{\text{max}} = \arg\max_i |\mathcal{C}_i|$
  Retain journal entries from the largest cluster $\mathcal{C}_{\text{max}}$
end for
Output: Subset of representative journal entries $\mathcal{C}_{\text{max}}$ for each author

personality traits. Using the facebook/bart-large (Lewis et al., 2019) model, we trained a Big 5 Personality classifier with the PANDORA(Gjurković et al., 2021) dataset to evaluate and filter journal entries based on their alignment with core personality traits. The filtering process relied on two key parameters: alpha (α), which controlled the filtration strictness at the journal level, and beta (β), which managed the convergence of personality traits across authors. The lower the parameter values, the stricter the filtration process. Algorithm 2 outlines this filtration strategy.The critical components of the process are as follows:

- **Personality Trait Generation**: Each journal entry was processed using a Big 5 Personality classifier, predicting the Big 5 traits: OCEAN. This provided a detailed personality profile for each author across all their entries.
- **Journal-Level Filtration** (α): We measured each journal entry's deviation from the author's average personality profile, with α (can be any value but tested with 0,1) setting a threshold based on the standard deviation of these deviations. Entries with significant deviations were excluded to retain journals that best reflected the author's core traits.
- **Author-Level Filtration** (β): We assessed personality consistency across authors by comparing their average profiles to the global dataset, with β (can be any value but tested with 0,0.5) filtering out authors with excessive divergence to ensure alignment with the overall dataset.

5 Dataset Statistics

JIC consists of 418,476 dialogues, 20,000 reserved for the test set H and the rest for training, with 3,347,808 turns and 6,695,616 utterances, averaging 8 turns and 16 utterances per dialogue. Each

Algorithm 2 Personality Trait Convergence Filtering

Input: Filtered journal entries $\mathcal{J}_a$ per author a, pre-trained Big 5 Personality classifier $\mathcal{C}$, parameters α, β
Initialize: Personality classifier $\mathcal{C}$ to predict Big 5 traits $\mathbf{p}_a$ for each journal
for each author $a \in \mathcal{A}$ **do**
  for each journal entry $j \in \mathcal{J}_a$ **do**
    Compute the personality traits $\mathbf{p}_j = \mathcal{C}(j)$
  end for
  Calculate the average personality traits $\bar{\mathbf{p}}_a = \frac{1}{|\mathcal{J}_a|} \sum_{j \in \mathcal{J}_a} \mathbf{p}_j$
  for each journal entry $j \in \mathcal{J}_a$ **do**
    Compute divergence $\Delta_j = \|\mathbf{p}_j - \bar{\mathbf{p}}_a\|$
    if $\Delta_j \leq \alpha$ **then**
      Retain journal entry j
    end if
  end for
end for
Calculate the global average personality traits $\bar{\mathbf{p}}_{\text{global}} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \bar{\mathbf{p}}_a$
for each author $a \in \mathcal{A}$ **do**
  Compute divergence $\Delta_a = \|\bar{\mathbf{p}}_a - \bar{\mathbf{p}}_{\text{global}}\|$
  if $\Delta_a \leq \beta$ **then**
    Retain author a and their corresponding journal entries
  end if
end for
Output: Refined set of authors and journal entries with consistent personality traits

Attrib	PC	SPC	BST	JIC
# of Conversations	18,878	10,905	6,808	418,476
Tot. # of Turns	120,361	152,945	44,959	3,347,808
Avg. # of Turns	6.38	14.03	6.60	8.00
Tot. # of Utterances	259,600	310,874	89,918	6,695,616
Avg. Utt. (conv)	13.75	28.51	13.21	16.00
Avg. Words (u)	11.24	8.75	13.46	15.48
Avg. Conv. Length (w)	154.56	249.53	177.83	247.61
Longest Conv. (u)	49	117	28	16
Shortest Conv. (u)	11	6	4	16
Longest Conv. (w)	477	637	422	581
Shortest Conv. (w)	41	60	24	16
Avg. Topic Consistency (u)	0.50	0.57	0.55	0.53
Avg. Semantic Similarity (u)	0.31	0.39	0.36	0.36

Table 1: Comparison of various datasets across several attributes. Here PC is Persona Chat, SPC is Synthetic Persona Chat, BST is Blended Skill Talk, (u) means per utterance, and (w) means per word.

utterance contains about 15.48 words, resulting in an average conversation length of 247.61 words. The dataset exhibits moderate topic consistency (0.5281) and an average semantic similarity of 0.3611 between consecutive utterances, highlighting its diversity and scale. Table 1 shows detailed comparisons with other datasets.

6 Experimentation

Training and Inference were carried out in two settings, as shown in Fig. 3. The training splits for JIC are shown in Table 2.

Name	Abbreviation	Size
JIC-tiny	JIC-t	8k
JIC-small	JIC-s	20k
JIC-medium	JIC-m	30k
JIC-large	JIC-l	100k
JIC-$\alpha_0\beta_0$	-	117,749
JIC-$\alpha_1\beta_0$	JIC-best	132,791
JIC-$\alpha_1\beta_{0.5}$	-	226,873
JIC-all	-	398,476

Table 2: JIC splits and their sizes

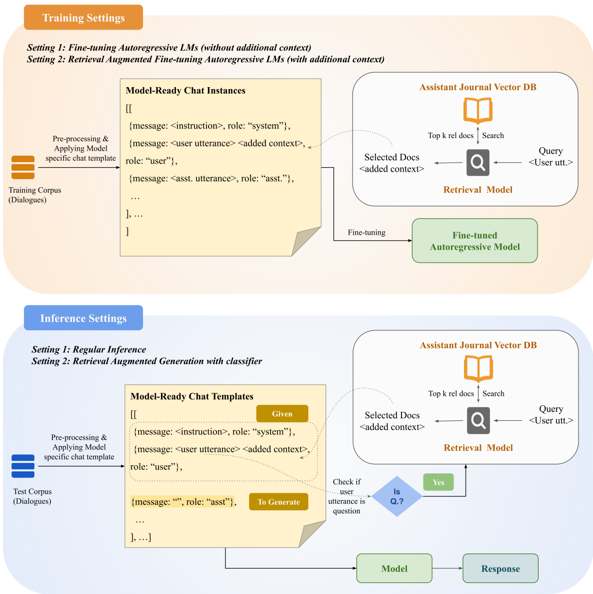

Figure 3: Model Training and Inference Settings

6.1 Training

To analyze the impact of data scaling, we sampled subsets of dialogues from 398,476 dialogues, consistently holding out 1,000 dialogues for validation across all experiments. Let $\mathcal{D}_{\text{train}}$ represent the training set and $\mathcal{D}_{\text{val}}$ the validation set, where $\mathcal{D}_{\text{train}} \subseteq \mathcal{D}$, the total dataset. We fine-tuned the LLaMa 3 8B Instruct(AI@Meta, 2024) and Mistral 7B v0.3(Jiang et al., 2023) models using a parameter-efficient technique, Low-Rank Adaptation (LoRA)(Hu et al., 2022). Specifically, we adjusted the query, key, value, and output projection layers (i.e., W_q, W_k, W_v, W_o), updating only these parameters while keeping the rest of the model frozen.

The training objective was to minimize the negative log-likelihood (NLL) loss, defined as:

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} \log p(y|x;\theta),$$

where x is the input (dialogue context) and y is the output (next dialogue turn). The model aims to maximize $p(y|x)$, the probability of generating the correct response, conditioned on the context.

LoRA Optimization: We introduced low-rank updates to the projection layers rather than fine-tuning all weights. The update rule for W_q (query projection) can be formulated as:

$$W_q' = W_q + \Delta W_q, \quad \text{where}$$
$$\Delta W_q = A_q B_q, \quad A_q \in \mathbb{R}^{d\times r}, \quad B_q \in \mathbb{R}^{r\times d},$$

with r being the rank of the update. Similar updates apply to W_k, W_v, and W_o. This approach greatly reduces computational overhead.

We extended the training process with Retrieval Augmented Fine-tuning (RAFt.)(Zhang et al., 2024) mechanisms to enhance context relevance. Let x_i represent the user's last utterance (the query) and $\mathcal{C}_i$ the assistant's journal entry (the context). Using Maximum Marginal Relevance (MMR)(Carbonell and Stewart, 1999), the top k most relevant segments, denoted as $\mathcal{C}_i^{(1)}, \ldots, \mathcal{C}_i^{(k)}$, were selected based on their similarity scores to x_i, while minimizing redundancy. The enriched input becomes:

$$\tilde{x}_i = \text{concat}(x_i, \mathcal{C}_i^{(1)}, \ldots, \mathcal{C}_i^{(k)}),$$

where $\tilde{x}_i$ includes both the query and retrieved context. Training arguments were consistent across models and are available in Appendix C, and Training strategies for other datasets are mentioned in Appendix D.

6.2 Inference

Inference was carried out in two specific settings: utterance level and using Retrieval Augmented Generation(RAG) (Lewis et al., 2020). RAG employed a classifier, $\mathcal{C}$, to distinguish between user statements and questions. It was activated for queries classified as questions, $\mathcal{C}(x_i) = 1$. The context $\mathcal{C}_i$ was retrieved by selecting the top k relevant chunks from the assistant's journal using MMR. The enriched query becomes $\tilde{x}_i = \{x_i, \mathcal{C}_i\}$, which was passed to the model for response generation. For non-questions, $\mathcal{C}(x_i) = 0$, no retrieval was performed. This selective retrieval application improved performance, particularly in handling chitchat vs. complex queries.

6.3 Evaluation Strategy

We employed both automatic metric-based evaluation and evaluation by the LM Eval Harness(Gao et al., 2024) framework by EleutherAI. Automatic evaluation used BLEU(Papineni et al., 2002), METEOR(Banerjee and Lavie, 2005), BERTScore(Zhang* et al., 2020), and ROUGE-1, ROUGE-2, and ROUGE-L(Lin, 2004). The average of these metrics provided an overall performance score. Whereas, the LM Eval Harness assessed the models on the Big 5 personality traits[5]

[5]Metrics: persona_openness, persona_conscientiousness, persona_extraversion, persona_agreeableness, per-

across 1,000 samples per trait, highlighting the model's ability to adapt to distinct personality profiles.

7 Results and Discussion

We conducted extensive testing on automated metrics and LM-eval benchmarks to assess model performance across different configurations.

7.1 Automatic metric-based evaluation

Dataset	Model	Train cfg.	Test cfg.	Avg. Score
Pre-trained	LLaMA	ZS	RAG	0.2516
	Mistral	ZS	Reg.	0.2154
PC	LLaMA	PAFt.	RAG	0.2538
	Mistral	PAFt.	RAG	0.2122
SPC	LLaMA	PAFt.	RAG	0.2544
	Mistral	PAFt.	RAG	0.2132
BST	LLaMA	PAFt.	RAG	0.2518
	Mistral	PAFt.	RAG	0.2134
JIC-$\alpha_1\beta_0$	LLaMA	RAFt.	RAG	0.2843
	Mistral	Ft.	RAG	0.2453
JIC-all	LLaMA	Ft.	RAG	<u>**0.3105**</u>
	Mistral	Ft.	RAG	**0.2646**

Table 3: Evaluation of LLaMA and Mistral models trained on various Datasets, tested on JIC(2k subset). The highest average score across models is highlighted. The best score in the table is underlined. Only reported the best score for each configuration, detailed results for JIC in Table 11, other Datasets Table 12.

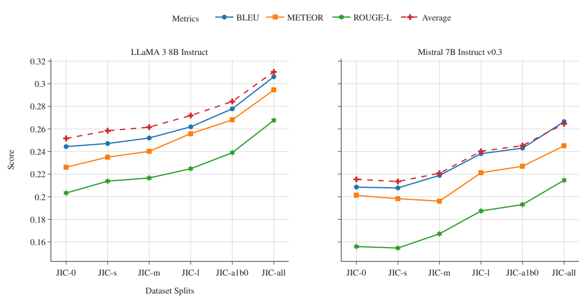

Figure 4: Performance of LLaMA(left) and Mistral(right) models across various JIC dataset splits. Reported: BLEU, METEOR, ROUGE-L, Avg(across all: Table 3).

We evaluated the LLaMA and Mistral models across different configurations and dataset sizes, with consistent validation (randomly sampled) and test splits of 1k and 2k samples. As expected, model performance generally improved as training data increased. The best average scores were achieved when models were trained on the entire dataset, with LLaMA consistently outperforming Mistral. LLaMA achieved a best average score of 0.3105, representing a 35.1% improvement over its zero-shot baseline, while Mistral reached a best score of 0.2646, improving by 23.2%. Table 3 report our findings. Fig. 4 shows comparative results, detailed results in Appendix E. Due to computational constraints, we could not use RAFt on the entire dataset.

Contrary to expectations, where RAFt should theoretically enhance model performance by providing additional context, Mistral did not show any improvements when using RAFt., regardless of the dataset size. This contrasts with LLaMA, which consistently improved with RAFt. across all data splits. One possible explanation for this discrepancy is that Mistral, starting from a weaker baseline, might not have been able to effectively leverage the additional retrieved context provided by the retrieval mechanism. In contrast, LLaMA's stronger baseline performance allowed it to utilize the retrieved information more effectively, leading to consistent gains with RAFt. This suggests that while RAFt is generally beneficial, its effectiveness may depend on the model's inherent capabilities and baseline performance.

7.2 LM-Eval Harness Results

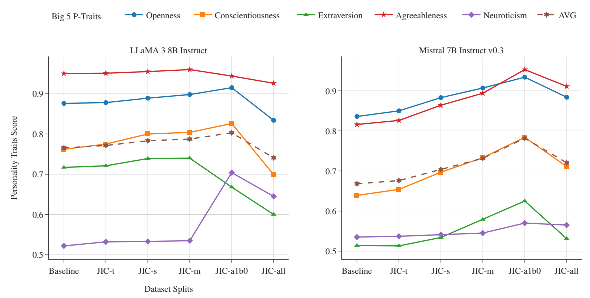

Figure 5: Performance of LLaMA and Mistral models across various JIC dataset splits. The left panel displays the results for LLaMA, while the right panel shows the results for Mistral.

We analyzed the models' ability to capture Big Five personality traits (O, C, E, A, N) across different training configurations. Notably, trait scores plateaued after a certain dataset size, fine-tuned using α and β parameters. The highest scores were achieved with α=1 and β=0 (around one-third of the dataset). LLaMA outperformed Mistral, achieving a top score of 0.8030 with RAFt., while Mistral's best score was 0.7816. LLaMA's strong base-

sona_neuroticism

Dataset	Model	Train. cfg.	Train size	Personality Traits Score					AVG
				O	C	E	A	N	
Pre-Trained	LLaMA	ZS	-	0.8760	0.7620	0.7170	0.9500	0.5220	0.7654
	Mistral	ZS	-	0.8360	0.6390	0.5140	0.8160	0.5350	0.6680
PC	LLaMA	Ft.	Full	0.8740	0.7660	0.7180	0.9510	0.5240	0.7666
	Mistral	Ft.	Full	0.8380	0.6380	0.5140	0.8080	0.5370	0.6670
SPC	LLaMA	PAFt.	Full	0.8750	0.7680	0.7190	0.9500	0.5240	0.7672
	Mistral	PAFt.	Full	0.8320	0.6330	0.5120	0.8100	0.5360	0.6646
BST	LLaMA	PAFt.	Full	0.8760	0.7630	0.7160	0.9510	0.5240	0.7660
	Mistral	Ft.	Full	0.8360	0.6400	0.5130	0.8180	0.5340	0.6682
JIC-medium	LLaMA	Ft.	29k	0.8770	0.7800	0.7160	0.9520	0.5300	0.7710
		RAFt.	29k	0.8980	**0.8040**	**0.7400**	0.9600	0.5350	0.7874
	Mistral	Ft.	29k	0.8670	0.6820	0.5220	0.8600	0.5380	0.6938
		RAFt.	29k	0.9070	0.7330	0.5790	0.8940	0.5450	0.7316
JIC (α,β)	LLaMA	Ft. $\alpha_0\beta_0$	$\sim$115k	0.8810	0.7980	0.7140	0.9580	0.5380	0.7778
		Ft. $\alpha_1\beta_0$	$\sim$135k	0.8860	0.7930	0.7080	**0.9600**	0.5570	0.7808
		RAFt. $\alpha_1\beta_0$*	$\sim$135k	**0.9150**	0.7840	0.6680	0.9440	**0.7040**	**0.8030**
		Ft. $\alpha_1\beta_{0.5}$	$\sim$220k	0.8830	0.7990	0.7080	0.9580	0.5380	0.7772
	Mistral	Ft. $\alpha_0\beta_0$	$\sim$100k	0.9090	0.7430	0.5690	0.9030	0.5490	0.7346
		Ft. $\alpha_1\beta_0$	$\sim$135k	0.9120	0.7480	0.5800	0.9090	0.5490	0.7396
		RAFt. $\alpha_1\beta_0$*	$\sim$135k	**0.9340**	**0.8260**	**0.6250**	**0.9530**	**0.5700**	**0.7816**
		Ft. $\alpha_1\beta_{0.5}$	$\sim$220k	0.9050	0.7530	0.5760	0.9140	0.5590	0.7414
JIC-all	LLaMA	Ft.	$\sim$400k	0.8340	0.6990	0.6000	0.9260	0.6450	0.7408
	Mistral	Ft.	$\sim$400k	0.8840	0.7110	0.5310	0.9110	0.5650	0.7204

Table 4: Scores for Big Five traits (O, C, E, A, N) are shown, with the top scores for each model highlighted and the best overall in the Table (per trait) underlined. * denotes best model. Detailed results of all JIC subsets in Table 14.

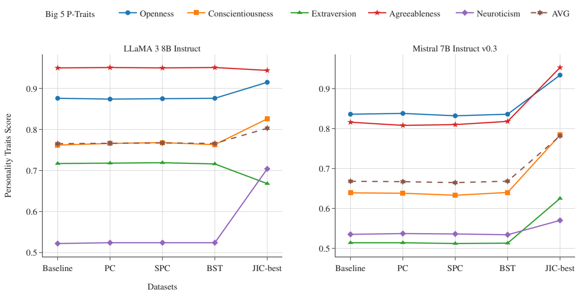

Figure 6: Personality trait scores across various datasets for the LLaMA 3 8B Instruct model (left) and Mistral 7B Instruct v0.3 (right)

line (0.7654) showed modest improvement (4.9%), whereas Mistral's lower baseline (0.6680) saw a 17% gain, suggesting that Mistral benefits more from data scaling and refinement, despite LLaMA's better overall performance.These results suggest that LLaMA's strong baseline may limit its potential for further improvement, while Mistral benefits more from data scaling and refinement strategies. Table 4 report our findings. Fig. 5 shows performance of LLaMA and Mistral models across various JIC splits. Fig. 6 compares performance of the models across different datasets.

Interestingly, while Mistral showed no improvement with RAFt in automated metrics, both models displayed significant gains in capturing personality traits with RAFt. This suggests that RAFt enables models to learn and internalize personality traits, even if the generated text doesn't exactly match the golden annotations. The retrieval process helps models better understand and generalize trait-specific behaviors, emphasizing these traits during training, regardless of text alignment.

7.3 Ablation Study

The ablation study focuses solely on personality traits, as the dataset is specifically designed to capture human-like personality dynamics, making automated metric evaluations less relevant in this context. Table 5 shows the impact of various configurations on the overall performance of LLaMA and Mistral. **(0)**: The best configuration, **RAFt.** ($\alpha = 1$, $\beta = 0$), yielded the highest average scores for both LLaMA (**0.8030**) and Mistral (**0.7816**). **(1)**: **No Retriever Augmentation**, the performance slightly dropped for both models (LLaMA: 0.7808, Mistral: 0.7396). **(2)**: **No filtration** (α, β set to None), we observed further performance degradation (LLaMA: 0.7408, Mistral: 0.7204). **(3)**: **Random sampling** to mimic (1) improved the scores compared to (2) but did not outperform (1), which shows the requirement of filtration (LLaMA: 0.7764, Mistral: 0.7302). **(4)**: **No Fine-tuning**, resulted in the lowest scores (LLaMA: 0.7654, Mistral: **0.6680**).

Model	Variants	Personality Traits Score					AVG
		O	C	E	A	N	
LLaMA	**(0) RAFt. (α=1,β=0)***	**0.9150**	0.7840	0.6680	0.9440	**0.7040**	**0.8030**
	(1) (w/o RA) Ft. (α=1,β=0)	0.8860	**0.7930**	0.7080	**0.9600**	0.5570	0.7808
	(1) + (2) w/o (a,b)	0.8340	0.6990	0.6000	0.9260	0.6450	0.7408
	(1)+ (2) + (3) w/ random sampling	0.8840	**0.7930**	**0.7180**	0.9550	0.5320	0.7764
	(1) + (2) + (4) w/o Ft.	0.8760	0.7620	0.7170	0.9500	0.5220	0.7654
Mistral	**(0) RAFt. (α=1,β=0)***	**0.9340**	**0.8260**	**0.6250**	**0.9530**	**0.5700**	**0.7816**
	(1) (w/o RA) Ft. (α=1,β=0)	0.9120	0.7480	0.5800	0.9090	0.5490	0.7396
	(1) + (2) w/o (a,b)	0.8840	0.7110	0.5310	0.9110	0.5650	0.7204
	(1)+ (2) + (3) w/ random sampling	0.905	0.736	0.563	0.901	0.546	0.7302
	(1) + (2) + (4) w/o Ft.	0.8360	0.6390	0.5140	0.8160	0.5350	0.6680

Table 5: Ablation study on JIC dataset comparing LLaMA and Mistral models across configurations, with best RAFt. settings (α=1, β=0) highlighted and bolded top scores per trait per model.

7.4 Personality Trait Scaling and Dataset Bias

We found that personality trait accuracy does not scale linearly with dataset size, likely due to dataset bias. The synthetic dialogues, often reflecting negative experiences from Reddit journal entries, led to an over-representation of neuroticism in the JIC dataset, skewing the capture of other traits like extraversion and conscientiousness. LLaMA, when fine-tuned without alpha-beta filtering, showed higher neuroticism, disrupting trait balance, with Mistral exhibiting similar but less pronounced behavior. LLaMA had the highest conscientiousness and extraversion scores in the JIC-medium split, potentially due to its stronger zero-shot baseline, while Mistral required more fine-tuning for similar performance. Furthermore, focusing on personality traits slightly reduced the model's ability in general reasoning tasks, which we believe can be improved through rehearsal learning(Wang et al., 2023). Detailed scores are in Appendix J. The relationship between data scaling and trait capture appears model-dependent and influenced by pre-training performance, with alpha-beta hyperparameters (α=1, β=0) offering the most balanced results across traits. Due to the computational costs, we experimented with a small range of values, finding the optimal remains an open research challenge.

7.5 Qualitative & Toxicity Analysis

Qualitative analysis is crucial for evaluating LLMs beyond quantitative metrics, offering insights into dialogue subtleties. Table 6 shows how well the models aligned with personality traits from a randomly selected annotated dialogue, ensuring unbiased and representative results. Examples can be found in Appendix G. Additionally, we conducted a toxicity analysis using the Detoxify[6] library to assess harmful content in our dataset, classifying

Settings	Role	Prompt & Response	Captured Traits
Golden	Assistant	Exactly! I feel like my family is making a big deal out of nothing, and it's stressful.	Neuroticism
LLaMA (our best)	Assistant	Yeah, I wish my family would listen to me without interrupting.	Neuroticism
LLaMA PAFt. SPC	Assistant	That's really refreshing. Sometimes simplicity is the best way to get things off your chest..	Agreeableness
LLaMA (base)	Assistant	That's really refreshing. Sometimes simplicity is the best way to go.	Agreeableness
Mistral (our best)	Assistant	That's smart. I've been so busy with family conflict, I haven't had time to reflect.	Neuroticism
Mistral (base) & PAFt. SPC	Assistant	That's a good approach. I think I'll try ZenJournal. Maybe I can explain how I feel without anxiety.	Agreeableness, Openness

Table 6: Comparison of real and model-generated dialogues capturing personality traits. The Table demonstrates how our best-performing models (LLaMA and Mistral) align with the traits reflected in the original dialogue. Detailed in Table 16.

dialogues based on various categories such as toxic, severely toxic, obscene, insult, identity hate, and threat. Dialogues with more than 25% toxic utterances were flagged, and those with severe issues like threats or identity hate were also marked. We found around 11k utterances out of 6.7M utterances classified as toxic. Flagged dialogues will be kept separately during release (detailed results in Appendix I).

8 Conclusion

Our research introduces the JIC dataset, which overcomes the limitations of static personas in existing conversational datasets. By grounding dialogues in long-form journal entries and capturing dynamic personality traits through a multi-step filtering process, we enable LMs to generate more authentic, personalized conversations. This approach significantly enhances conversational AI's ability to reflect real human personalities, offering engaging and relatable interactions.

[6]Detoxify

Limitations

- A fundamental limitation of our research lies in tuning the optimal α and β parameters. While the chosen values ($\alpha = 1$, $\beta = 0$) yielded promising results, refining these parameters remains an open challenge due to the computational demands of extensive experimentation.

- Synthetic data generation using LLaMA 70B introduces potential biases and safety concerns inherent in the pre-trained model. These biases could propagate into the dialogues, limiting diversity and authenticity, although the impact may be minimal.

- Furthermore, the dataset's inherent bias—stemming from an over-representation of neuroticism in Reddit journals—may have skewed the models' ability to capture traits like extraversion accurately.

- Finally, human evaluation remains a significant challenge, as assessing nuanced traits in synthetic dialogues can be tough and labor-intensive, highlighting the difficulty of balancing human insight and scalable evaluation methods.

Ethical Considerations

In creating the JIC dataset, we ensured that all journal entries were publicly available and anonymized to protect user privacy. We employed the Detoxify library to tag potentially toxic dialogues to mitigate the risk of harmful content. A strict threshold was set, flagging any dialogue where more than 25% of the utterances were classified as toxic. These flagged dialogues were kept separately to prevent their use in downstream tasks. This approach helps ensure the dataset remains safe and responsible for use in developing conversational AI systems.

Acknowledgements

We thank the anonymous reviewers for providing valuable feedback on our manuscript. We also extend our gratitude to the human judges who contributed to the evaluation of LLM-generated synthetic data. The human evaluation team was led by Sutirtho Roy (PhD student, Literatures in English, UC San Diego) and included Anushka Tiwari (PhD student, Computational and Data-Enabled Sciences, University at Buffalo). This work is supported by NSF Grant # 2214070. The content in this paper is solely the responsibility of the authors and does not necessarily represent the official views of the funding entity.

References

Zishan Ahmad, Kshitij Mishra, Asif Ekbal, and Pushpak Bhattacharyya. 2023. RPTCS: A reinforced persona-aware topic-guiding conversational system. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3482–3494, Dubrovnik, Croatia. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Jan M Allbeck and Norman I Badler. 2008. Creating crowd variation with the ocean personality model. *AAMAS'08*, 3(209):1217–1220.

Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences*, 124:150–159.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Christopher P Barlett and Craig A Anderson. 2012. Direct and indirect relations between the big 5 personality traits and aggressive and violent behavior. *Personality and individual differences*, 52(8):870–875.

Stefan Blomkvist. 2002. Persona–an overview. *Retrieved November*, 22(2004):15.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jaime Carbonell and Jade Stewart. 1999. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*.

Graham Caron and Shashank Srivastava. 2023. Manipulating the perceived personality traits of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2370–2386, Singapore. Association for Computational Linguistics.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12.

Souvik Das and Rohini Srihari. 2024. UNIWIZ: A unified large language model orchestrated wizard for safe knowledge grounded conversations. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1749–1762, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ondřej Dušek and Filip Jurčíček. 2016. A context-aware natural language generator for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 185–190, Los Angeles. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang, and Lilian Tang. 2023. Learning retrieval augmentation for personalized dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2523–2540, Singapore. Association for Computational Linguistics.

Gregory M Hurtz and John J Donovan. 2000. Personality and job performance: the big five revisited. *Journal of applied psychology*, 85(6):869.

Yerin Hwang, Yongil Kim, Hyunkyung Bae, Hwanhee Lee, Jeesoo Bang, and Kyomin Jung. 2023. Dialogizer: Context-aware conversational-QA dataset generation from textual sources. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8806–8828, Singapore. Association for Computational Linguistics.

Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15245–15270, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. 2022. Building a personalized dialogue system with prompt-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 96–105, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Tiziano Labruna, Sofia Brenna, and Bernardo Magnini. 2024. Dynamic task-oriented dialogue: A comparative study of llama-2 and bert in slot value generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 358–368, St. Julian's, Malta. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fei Liu, Julien Perez, and Scott Nowson. 2016. A recurrent and compositional model for personality trait recognition from short texts. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 20–29, Osaka, Japan. The COLING 2016 Organizing Committee.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. *Preprint*, arXiv:1109.2378.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *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)*, pages 1699–1713, Online. Association for Computational Linguistics.

Sayantan Pal, Souvik Das, Rohini Srihari, Jeff Higginborham, and Jenna Bizovi. 2024. Empowering AAC users: A systematic integration of personal narratives with conversational AI. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 12–25, Miami, Florida, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sougata Saha, Souvik Das, and Rohini Srihari. 2022. Stylistic response generation by controlling personality traits and intent. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 197–211, Dublin, Ireland. Association for Computational Linguistics.

Caroline Schill, John M Anderies, Therese Lindahl, Carl Folke, Stephen Polasky, Juan Camilo Cárdenas, Anne-Sophie Crépin, Marco A Janssen, Jon Norberg, and Maja Schlüter. 2019. A more dynamic understanding of human behaviour for the anthropocene. *Nature Sustainability*, 2(12):1075–1082.

Seth J Schwartz, Theo A Klimstra, Koen Luyckx, William W Hale III, Tom Frijns, Annerieke Oosterwegel, Pol AC Van Lier, Hans M Koot, and Wim HJ Meeus. 2011. Daily dynamics of personal identity and self–concept clarity. *European Journal of Personality*, 25(5):373–385.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. 2022. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969, Seattle, United States. Association for Computational Linguistics.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Jen tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *The Twelfth International Conference on Learning Representations*.

Zhicheng Wang, Yufang Liu, Tao Ji, Xiaoling Wang, Yuanbin Wu, Congcong Jiang, Ye Chao, Zhencong Han, Ling Wang, Xu Shao, and Wenqiu Zeng. 2023. Rehearsal-free continual language learning via efficient parameter isolation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10933–10946, Toronto, Canada. Association for Computational Linguistics.

Charles Welch, Allison Lahnala, Veronica Perez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive interviewing: A conversational system for coping with COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861, Hong Kong, China. Association for Computational Linguistics.

Diyi Yang, Sherry Tongshuang Wu, and Marti A. Hearst. 2024. Human-AI interaction in the age of LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 34–38, Mexico City, Mexico. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. Raft: Adapting language model to domain specific rag. *Preprint*, arXiv:2403.10131.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

A Prompting Strategy

The prompting strategy was carefully designed to guide the instruct models in generating dialogues that align with the personality traits and tones expressed in the journal entries. By explicitly framing the instruction to focus on finding common ground between the authors' experiences, thoughts, or emotions, we ensured that the models would remain faithful to the context provided by the journals. This was crucial for maintaining the natural flow and personality-driven nature of the conversation. The instruction emphasized creating a balanced, engaging dialogue that reflected the distinct personality traits evident in the journals. This helped direct the model toward producing conversations that stayed true to the underlying personalities, encouraging the generation of responses that aligned with each author's emotional tone and life experiences. **Note**: Due to the limitations of the free API, some dialogues were incomplete or incorrect, leading us to skip certain combinations. As a result, the actual number of dialogues generated is lower than the possible maximum combinations.

The detailed prompt provided is as follows:

> <Instruction>: Create a 9-turn dialogue in english between two authors based on the journal entries provided below. The dialogue should reflect a natural and engaging conversation, finding common ground between the authors' experiences, thoughts, or emotions. Ensure that the conversation stays true to the personality traits and tones expressed in the journal entries. Each author should contribute equally, with utterances that are concise, relevant, and no longer than 20 words. <journal 1>, <journal 2>

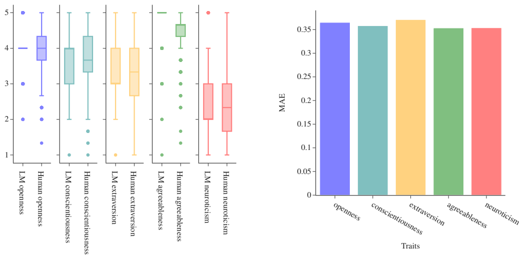

Figure 7: Left: Boxplot comparison of personality trait ratings (OCEAN) between the LLM (GPT-4o) and human annotators. Right: Mean Absolute Error (MAE) between LLM and Human Scores for Big 5 Traits

B Quality Assessment of Synthetic Dialogues & Statistical Significance

We used an LLM-based evaluation of 1,000 randomly sampled synthetic dialogues from our dataset using a Likert scale(1-5), focusing on O.C.E.A.N criteria. Three in-house PhD students from linguistics and computer science rated these dialogues. The box plot in Fig. 7 shows that human rating had higher variability than LLM rating. The MAE plot highlights almost similar errors for all the traits. Table 9 shows one example dialogue (not cherry-picked) rated by LLM and Human judges.

Furthermore, we assessed 4000 more randomly sampled dialogues using GPT-4o. Table 7 shows the Mean and Standard Deviation across various traits. We evaluated the agreement between the language model's (LM) personality trait predictions and human annotations using Pearson, Spearman, and Intraclass Correlation Coefficients (ICC). As shown in Table 8, Neuroticism had the highest LM-human agreement with an ICC of 0.970. These results demonstrate that the LM effectively captures most personality traits, though certain traits, like Openness and Agreeableness, exhibit slightly lower alignment with human annotations.

C Training Arguments and GPU

All the models were trained on a single A100 80 GB. Table 10 shows the Training Args used to train all the Models. The batch size default was set to 4 but was reduced to 2 when A100 80GB was unavailable (used A100 40 GB). LoRA hyper parameters (r = 64, $alpha$ = 16, dropout = 0.1) were most significant.

D Fine-Tuning on Other Datasets

We extended the fine-tuning procedure to popular datasets like Persona-Chat (PC), Synthetic Persona-Chat (SPC), and Blended Skill Talk (BST), as described in §6.1. Let $\mathcal{D}_{\text{train}}$ represent the training set for each dataset, with sizes $|\mathcal{D}_{\text{PC}}| = 17,878$, $|\mathcal{D}_{\text{SPC}}| = 8,938$, and $|\mathcal{D}_{\text{BST}}| = 4,819$. We used the same parameter-efficient Low-Rank Adaptation (LoRA) method as earlier, updating the projection matrices W_q, W_k, W_v, and W_o while keeping the rest of the model frozen. For the second phase, instead of RAFt, we appended the persona information $\mathcal{P}_i$ from each dataset to the system prompt, resulting in an enriched input $\tilde{x}_i = \{x_i, \mathcal{P}_i\}$. This setup mirrors the RAFt process, where $\mathcal{P}_i$ acts as the additional context, we name it Persona Augmented Fine-tuning (PAFt.). The objective remained to minimize the NLL loss with the persona context guiding the generation to match the personality traits embedded in each dataset.

E Detailed Results of Automatic metric-based evaluation

Table 11 shows detailed results of the automated evaluation on various JIC subsets. Table 12 shows the results for other datasets on JIC test split. Table 13 shows the evaluation result of other datasets on their respective test splits.

F Detailed Results of LM-Eval Harness

Table 14 shows the scores for Big 5 traits (O,C,E,A,N) in JIC subsets. Table 15 shows the Big 5 traits in other datasets.

G Qualitative Analysis Examples

Table 16 compares real and model-generated dialogues capturing personality traits.

H Test Set Analysis

The test set analysis was conducted to evaluate dialogue-specific Big 5 personality traits, and the following percentages were observed for each trait, given in Table 17.

These results indicate that the test set predominantly exhibits high levels of all traits except *Conscientiousness*, which shows a more balanced distribution. This suggests that the dataset is skewed towards dialogues characterized by predominant traits, potentially impacting the generalizability of predictions for conscientiousness.

Setting	Metric	Basic						Personality				
		Coh	Grnd	Spc	Flu	Und	Eng	O	C	E	A	N
LLM (∼5k)	Mean	4.9452	4.9532	4.2390	4.9898	4.9928	4.1592	4.0296	3.7690	3.2711	4.8012	2.4039
	Std	0.2534	0.2232	0.4715	0.1005	0.0846	0.4289	0.4996	0.6451	0.5766	0.4571	1.0225
LLM (1k)	Mean	4.9369	4.9409	4.2352	4.9860	4.9920	4.1722	4.0611	3.7738	3.3213	4.8228	2.3493
	Std	0.2884	0.2714	0.4628	0.1258	0.0998	0.4345	0.5175	0.6354	0.5870	0.4405	1.0542
Humans (1k)	Mean	-	-	-	-	-	-	3.9957	3.7287	3.3146	4.5162	2.3974
	Std	-	-	-	-	-	-	0.6209	0.7281	0.7695	0.4466	1.0548

Table 7: Mean and Standard Deviation across various traits. Here Coherence, Grounding, Specificity, Fluency, Understandability and Engagement are abbreviated as Coh, Grnd, Spc, Flu, Und and Eng respectively

Configs	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
P_LLM_J1	0.7711	0.7933	0.7916	0.7684	0.9189
S_LLM_J1	0.7638	0.7823	0.8023	0.7034	0.9142
P_LLM_J2	0.7351	0.8174	0.7982	0.7702	0.9280
S_LLM_J2	0.7241	0.8135	0.8172	0.7084	0.9218
P_LLM_J3	0.7475	0.8287	0.8005	0.7946	0.9295
S_LLM_J3	0.7395	0.8209	0.8202	0.7348	0.9243
P_J1_J2	0.5763	0.6521	0.6373	0.5745	0.8487
S_J1_J2	0.5593	0.6437	0.6628	0.4897	0.8353
P_J1_J3	0.5660	0.6508	0.6282	0.6289	0.8593
S_J1_J3	0.5600	0.6361	0.6588	0.5279	0.8499
P_J2_J3	0.5349	0.6592	0.6354	0.5971	0.8699
S_J2_J3	0.5200	0.6538	0.6643	0.5039	0.8602
ICC	0.8751	0.9124	0.9020	0.8914	0.9703

Table 8: Pearson(P), Spearman(S), and Intraclass Correlation Coefficients (ICC) between the language model (LM: GPT4-o) and human annotations for the Big 5 personality traits. High ICC values indicate strong agreement between LLM and human evaluations.

I Toxicity Analysis Statistics

Fig. 8 shows the detailed statistics for the toxicity subclasses.

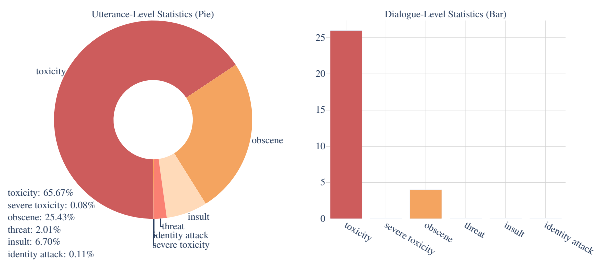

Figure 8: Left: Utterance-Level Pie Chart(around 11k utt.) and Right: Dialogue-Level Bar Chart Toxicity Subclass Counts

Table 18 shows the exact number of utterances and dialogues classified as toxic.

J LM-Eval Harness Common Tasks

Table 19 shows how the model fine-tuned in various JIC splits performed in average on various popular benchmarks to assess general reasoning, knowledge reasoning, and reading comprehension.

K Miscellaneous

Abbreviation Table can be found in Table 20.

Role	Generated Dialogue
User	I struggle to write in my native language. It's like my subconscious is blocking me.
Assistant	I used to write freely in my journals, but now it feels anxiety-inducing. I miss that liberated feeling.
User	Exactly! It's like something inside me is making it hard to do deep emotional work.
Assistant	I think I know what you mean. My writing halted after a surgery, and I haven't found my rhythm again.
User	Surgery? How did that affect your writing?
Assistant	Honestly, I thought it was the physical limitation, but even after recovery, the motivation wasn't there.
User	I get it. Even with my phone, it's hard when I'm in a psychotic episode. Eye strain and stimulation are overwhelming.
Assistant	I can imagine. It's tough when your mind and body are against you.
User	Exactly! It feels like an abusive part of me is stopping me from doing the digging work.
Assistant	I know that feeling. I used to write about everything, no matter how silly or candid. Now, it feels like I've lost that freedom.
User	Do you think we're just making excuses for ourselves? Maybe we're just scared to face our emotions?
Assistant	Maybe. But for me, it's more about finding that spark again. I cherish my old journals, but my 30s are void of that creative expression.
User	I relate. I've been clinically insane for years, and it's hard to even do daily tasks. Writing feels like a luxury I can't afford.
Assistant	I'm so sorry to hear that. Have you tried any strategies to get back into writing?
User	Not really. Just trying to find the energy to write. It's hard when your trauma response is so strong.
Assistant	I get it. For me, it's about rediscovering that joy and urgency to write. Maybe we can support each other in finding our way back?

Table 9: This Table presents a full dialogue sample annotated with Big 5 personality traits. The scores (ranging from 1 to 5) represent the significance of each trait in the dialogue, with 1 indicating low significance and 5 indicating high significance. The dialogue is evaluated by one language model (GPT4o) and three human judges (J1, J2, J3). The scores for GPT4o are [4, 3, 3, 5, 4], J1 are [4, 2, 2, 5, 5], J2 are [4, 3, 2, 5, 5], and J3 are [4, 3, 2, 5, 4], reflecting an overall agreement in the evaluation of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN).

Argument	Value
`lora_r`	64
`lora_alpha`	16
`lora_dropout`	0.1
`bf16`	True
`learning_rate`	2.0e-05
`gradient_accumulation_steps`	128
`gradient_checkpointing`	True
`logging_strategy`	Steps
`logging_steps`	1
`save_strategy`	Steps
`save_steps`	100
`eval_steps`	100
`per_device_train_batch_size`	4
`per_device_eval_batch_size`	4
`max_seq_length`	2048
`lr_scheduler_type`	Cosine
`early_stopping_patience`	4

Table 10: Trainer Arguments

Dataset	Model	Train cfg.	Test cfg.	Train size	Score			ROUGE Score			AVG
					BLEU	**METEOR**	**BERT**	**R1**	**R2**	**RL**	
JIC-0	LLaMA	ZS	Reg.	-	0.2115	0.2176	0.5049	0.2049	0.0646	0.1760	0.2299
		ZS	RAG	-	0.2444	0.2261	0.5210	0.2320	0.0829	0.2033	**0.2516**
	Mistral	ZS	Reg.	-	0.2085	0.2012	0.4957	0.1881	0.0429	0.1559	**0.2154**
		ZS	RAG	-	0.2078	0.1983	0.4912	0.1870	0.0419	0.1545	0.2134
JIC-tiny	LLaMA	Ft.	Reg.	6.4k	0.2214	0.2194	0.5092	0.2132	0.0688	0.1850	0.2362
		Ft.	RAG		0.2462	0.2289	0.5225	0.2342	0.0849	0.2054	0.2537
		RAFt.	Reg.		0.2296	0.2206	0.5146	0.2234	0.0758	0.1953	0.2432
		RAFt.	RAG		0.2477	0.2305	0.5244	0.2399	0.0889	0.2115	**0.2571**
	Mistral	Ft.	Reg.	6.4k	0.2078	0.1858	0.4864	0.1846	0.0398	0.1544	0.2098
		Ft.	RAG		0.2096	0.1927	0.4911	0.1872	0.0425	0.1549	**0.2130**
		RAFt.	Reg.		0.2031	0.1762	0.4801	0.1770	0.0365	0.1469	0.2033
		RAFt.	RAG		0.2055	0.1848	0.4846	0.1812	0.0378	0.1490	0.2071
JIC-small	LLaMA	Ft.	Reg.	19k	0.2313	0.2235	0.5156	0.2234	0.0750	0.1951	0.2440
		Ft.	RAG		0.2444	0.2261	0.5210	0.2320	0.0830	0.2033	0.2516
		RAFt.	Reg.		0.2471	0.2351	0.5264	0.2425	0.0855	0.2138	**0.2584**
		RAFt.	RAG		0.2488	0.2372	0.5190	0.2407	0.0920	0.2120	0.2583
	Mistral	Ft.	Reg.	19k	0.2079	0.1872	0.4881	0.1879	0.0417	0.1582	0.2118
		Ft.	RAG		0.2078	0.1983	0.4912	0.1870	0.0419	0.1546	**0.2135**
		RAFt.	Reg.		0.2070	0.1814	0.4829	0.1823	0.0395	0.1523	0.2075
		RAFt.	RAG		0.2089	0.1887	0.4863	0.1852	0.0415	0.1530	0.2106
JIC-medium	LLaMA	Ft.	Reg.	29k	0.2391	0.2294	0.5215	0.2320	0.0803	0.2036	0.2510
		Ft.	RAG		0.2444	0.2261	0.5210	0.2320	0.0830	0.2033	0.2517
		RAFt.	Reg.		0.2520	0.2401	0.5283	0.2458	0.0871	0.2166	**0.2616**
		RAFt.	RAG		0.2511	0.2412	0.5214	0.2433	0.0927	0.2142	0.2606
	Mistral	Ft.	Reg.	29k	0.2189	0.1961	0.4973	0.1974	0.0477	0.1672	**0.2208**
		Ft.	RAG		0.2078	0.1983	0.4912	0.1870	0.0419	0.1545	0.2134
		RAFt.	Reg.		0.2130	0.1858	0.4870	0.1881	0.0434	0.1579	0.2125
		RAFt.	RAG		0.2183	0.1943	0.4922	0.1937	0.0460	0.1609	0.2176
JIC-large	LLaMA	Ft.	Reg.	99k	0.2619	0.2559	0.5356	0.2550	0.0980	0.2248	**0.2719**
		Ft.	RAG		0.2503	0.2495	0.5229	0.2468	0.1013	0.2184	0.2649
	Mistral	Ft.	Reg.	99k	0.2316	0.2112	0.5082	0.2116	0.0561	0.1813	0.2333
		Ft.	RAG		0.2381	0.2212	0.5132	0.2199	0.0613	0.1874	**0.2402**
JIC-$\alpha_1\beta_0$	LLaMA	Ft.	Reg.	$\sim$135k	0.2743	0.2620	0.5436	0.2657	0.1053	0.2360	0.2812
		Ft.	RAG		0.2444	0.2261	0.5210	0.2319	0.0830	0.2034	0.2516
		RAFt.	Reg.		0.2722	0.2595	0.5418	0.2625	0.0992	0.2319	0.2778
		RAFt.	RAG		0.2778	0.2680	0.5436	0.2702	0.1073	0.2390	**0.2843**
	Mistral	Ft.	Reg.	$\sim$135k	0.2371	0.2182	0.5125	0.2186	0.0600	0.1872	0.2389
		Ft.	RAG		0.2430	0.2270	0.5181	0.2261	0.0642	0.1931	**0.2453**
		RAFt.	Reg.		0.2149	0.1985	0.4934	0.1972	0.0486	0.1669	0.2199
		RAFt.	RAG		0.2364	0.2108	0.5088	0.2136	0.0567	0.1822	0.2348
JIC-all	LLaMA	Ft.	Reg.	$\sim$400k	0.2967	0.2826	0.5587	0.2878	0.1210	0.2576	0.3007
		Ft.	RAG		0.3062	0.2945	0.5651	0.2989	0.1308	0.2676	**0.3105**
	Mistral	Ft.	Reg.	$\sim$400k	0.2589	0.2382	0.5273	0.2392	0.0763	0.2077	0.2579
		Ft.	RAG		0.2665	0.2451	0.5334	0.2468	0.0813	0.2146	**0.2646**

Table 11: Evaluation of LLaMA 3 8B Instruct and Mistral 7B v0.3 Instruct models on the JIC dataset, using various configurations and dataset subsets. The training configurations include Zero-shot (ZS), Fine-tuning on dialogues (Ft.), and Retrieval-augmented Fine-tuning on dialogues (RAFt.). Inference was performed with regular (Reg.) or Retrieval Augmented Generation (RAG) settings. The validation split and test split for the evaluation were kept constant at 1k and 2k samples, respectively. The highest average score across models for each dataset subset is highlighted. The best score in the Table is underlined.

Dataset	Model	Tr. cfg.	Ts. cfg.	Score			ROUGE Score			AVG
				BLEU	METEOR	BERT	R1	R2	RL	
PC	LLaMA	PAFt.	Reg.	0.2228	0.2191	0.5103	0.2151	0.0699	0.1868	0.2373
		PAFt.	RAG	0.2454	0.2282	0.5227	0.2353	0.0849	0.2065	0.2538
	Mistral	PAFt.	Reg.	0.2073	0.1842	0.4851	0.1821	0.0390	0.1522	0.2083
		PAFt.	RAG	0.2087	0.1929	0.4892	0.1862	0.0416	0.1548	0.2122
SPC	LLaMA	PAFt.	Reg.	0.2237	0.2209	0.5108	0.2160	0.0706	0.1874	0.2382
		PAFt.	RAG	0.2464	0.2291	0.5233	0.2358	0.0847	0.2071	0.2544
	Mistral	PAFt.	Reg.	0.2086	0.1847	0.4864	0.1836	0.0402	0.1540	0.2096
		PAFt.	RAG	0.2102	0.1929	0.4905	0.1874	0.0423	0.1560	0.2132
BST	LLaMA	PAFt.	Reg.	0.2203	0.2188	0.5083	0.2120	0.0680	0.1837	0.2352
		PAFt.	RAG	0.2443	0.2262	0.5212	0.2322	0.0831	0.2036	0.2518
	Mistral	PAFt.	Reg.	0.2084	0.1881	0.4869	0.1838	0.0402	0.1533	0.2101
		PAFt.	RAG	0.2078	0.1977	0.4911	0.1872	0.0419	0.1548	0.2134

Table 12: Evaluation of LLaMA 3 8B Instruct and Mistral 7B v0.3 Instruct models on other Datasets, using various configurations. The training configurations include Persona Augmented Fine-tuning (PAFt.) on the full dataset. Inference was performed with regular (Reg.) or Retrieval Augmented Generation (RAG) settings. Test splits for the evaluation were kept constant at 2k samples from JIC.

Dataset	Model	Train cfg.	Score			ROUGE			AVG
			BLEU	METEOR	BERT	R1	R2	RL	
PC	LLaMA	ZS	0.1183	0.1419	0.4272	0.1268	0.0207	0.1122	0.1579
		Ft.	0.1275	0.1383	0.4318	0.1345	0.0235	0.1202	0.1626
		PAFt.	0.1592	0.1608	0.4550	0.1675	0.0378	0.1497	**<u>0.1883</u>**
	Mistral	ZS	0.1340	0.1564	0.4314	0.1364	0.0206	0.1199	0.1665
		Ft.	0.1367	0.1276	0.4247	0.1310	0.0183	0.1171	0.1592
		PAFt.	0.1551	0.1537	0.4419	0.1489	0.0270	0.1320	**0.1764**
SPC	LLaMA	ZS	0.1962	0.2806	0.5369	0.2542	0.1165	0.2386	0.2705
		Ft.	0.2093	0.2870	0.5458	0.2675	0.1296	0.2519	0.2819
		PAFt.	0.2415	0.3080	0.5759	0.3079	0.1629	0.2936	**<u>0.3150</u>**
	Mistral	ZS	0.2041	0.3199	0.5462	0.2586	0.1154	0.2378	0.2803
		Ft.	0.2452	0.3208	0.5704	0.3006	0.1520	0.2837	**0.3121**
		PAFt.	0.2358	0.3184	0.5585	0.2861	0.1391	0.2664	0.3007
BST	LLaMA	ZS	0.1370	0.1437	0.4379	0.1378	0.0220	0.1145	0.1655
		Ft.	0.1372	0.1431	0.4379	0.1374	0.0220	0.1143	0.1653
		PAFt.	0.1475	0.1395	0.4491	0.1541	0.0294	0.1319	**<u>0.1752</u>**
	Mistral	ZS	0.1434	0.1357	0.4320	0.1355	0.0185	0.1130	0.1630
		Ft.	0.1446	0.1399	0.4334	0.1374	0.0190	0.1132	0.1646
		PAFt.	0.1486	0.1459	0.4363	0.1396	0.0190	0.1155	**0.1675**

Table 13: Evaluation of LLaMA 3 8B Instruct and Mistral 7B v0.3 Instruct models on three datasets: Persona Chat (PC), Synthetic Persona Chat (SPC), and Blended Skill Talk (BST), using various training configurations. The training configurations include zero-shot (ZS), fine-tuning on dialogues (Ft.), and persona-augmented finetuning (PAFt.). The entire dataset was used for each training configuration. Testing was performed on their respective test splits. The highest average score across models for each dataset subset is highlighted, and the best overall score for each dataset in the Table is underlined.

Dataset	Model	Train. cfg.	Train size	Personality Traits Score					AVG
				O	C	E	A	N	
-	LLaMA	ZS	-	0.8760	0.7620	0.7170	0.9500	0.5220	0.7654
	Mistral	ZS	-	0.8360	0.6390	0.5140	0.8160	0.5350	0.6680
JIC-tiny	LLaMA	Ft.	6.4k	0.8770	0.7660	0.7180	0.9500	0.5280	0.7678
		RAFt.	6.4k	0.8780	0.7750	0.7210	0.9510	0.5320	0.7714
	Mistral	Ft.	6.4k	0.8380	0.6460	0.5140	0.8160	0.5350	0.6698
		RAFt.	6.4k	0.8500	0.6540	0.5130	0.8260	0.5370	0.6760
JIC-small	LLaMA	Ft.	19k	0.8750	0.7750	0.7180	0.9510	0.5290	0.7696
		RAFt.	19k	0.8890	0.8000	0.7390	0.9550	0.5330	0.7832
	Mistral	Ft.	19k	0.8540	0.6600	0.5150	0.8280	0.5370	0.6788
		RAFt.	19k	0.8830	0.6970	0.5340	0.8640	0.5410	0.7038
JIC-medium	LLaMA	Ft.	29k	0.8770	0.7800	0.7160	0.9520	0.5300	0.7710
		RAFt.	29k	0.8980	**0.8040**	**0.7400**	0.9600	0.5350	0.7874
	Mistral	Ft.	29k	0.8670	0.6820	0.5220	0.8600	0.5380	0.6938
		RAFt.	29k	0.9070	0.7330	0.5790	0.8940	0.5450	0.7316
JIC (α,β)	LLaMA	Ft. $\alpha_0\beta_0$	∼100k	0.8810	0.7980	0.7140	0.9580	0.5380	0.7778
		Ft. $\alpha_1\beta_0$	∼135k	0.8860	0.7930	0.7080	**0.9600**	0.5570	0.7808
		RAFt. $\alpha_1\beta_0$*	∼135k	**0.9150**	0.7840	0.6680	0.9440	**0.7040**	**0.8030**
		Ft. $\alpha_1\beta_{0.5}$	∼220k	0.8830	0.7990	0.7080	0.9580	0.5380	0.7772
	Mistral	Ft. $\alpha_0\beta_0$	∼100k	0.9090	0.7430	0.5690	0.9030	0.5490	0.7346
		Ft. $\alpha_1\beta_0$	∼135k	0.9120	0.7480	0.5800	0.9090	0.5490	0.7396
		RAFt. $\alpha_1\beta_0$*	∼135k	**0.9340**	**0.8260**	**0.6250**	**0.9530**	**0.5700**	**0.7816**
		Ft. $\alpha_1\beta_{0.5}$	∼220k	0.9050	0.7530	0.5760	0.9140	0.5590	0.7414
JIAC-all	LLaMA	Ft.	∼400k	0.8340	0.6990	0.6000	0.9260	0.6450	0.7408
	Mistral	Ft.	∼400k	0.8840	0.7110	0.5310	0.9110	0.5650	0.7204

Table 14: Evaluation of the Personality Trait Scores using LM Eval Harness across various subsets of the Journal Intensive Conversation (JIC) dataset using LLaMA 3 8B Instruct and Mistral 7B v0.3 Instruct models. The training configurations include Zero-shot (ZS), Fine-tuning on dialogues (Ft.), and Retrieval-augmented Fine-tuning on dialogues (RAFt.). For each model and subset, the scores for the Big Five personality traits—Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N)—were computed. The highest scores within each model are emphasized, and the best overall score per trait is underlined. * represents our best configuration.

Dataset	Model	Train. cfg.	Personality Traits Score					AVG
			O	C	E	A	N	
PC	LLaMA	Ft.	0.8740	0.7660	0.7180	0.9510	0.5240	0.7666
		PAFt.	0.8750	0.7650	0.7140	0.9490	0.5250	0.7656
	Mistral	Ft.	0.8380	0.6380	0.5140	0.8080	0.5370	0.6670
		PAFt.	0.8310	0.6350	0.5110	0.8050	0.5370	0.6638
SPC	LLaMA	Ft.	0.8760	0.7620	0.7150	0.9480	0.5260	0.7654
		PAFt.	0.8750	0.7680	0.7190	0.9500	0.5240	0.7672
	Mistral	Ft.	0.8280	0.6340	0.5120	0.8050	0.5360	0.6630
		PAFt.	0.8320	0.6330	0.5120	0.8100	0.5360	0.6646
BST	LLaMA	Ft.	0.8760	0.7600	0.7170	0.9500	0.5210	0.7648
		PAFt.	0.8760	0.7630	0.7160	0.9510	0.5240	0.7660
	Mistral	Ft.	0.8360	0.6400	0.5130	0.8180	0.5340	0.6682
		PAFt.	0.8350	0.6400	0.5140	0.8130	0.5360	0.6676

Table 15: Evaluation of Personality Trait Scores on external datasets including Persona Chat (PC), Synthetic Persona Chat (SPC), and Blended Skill Talk (BST), using LLaMA 3 8B Instruct and Mistral 7B v0.3 Instruct models. The training configurations include Fine-tuning on dialogues (Ft.) and Persona Augmented Fine-tuning (PAFt.). The personality traits evaluated include Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). Results show comparable performance to the baseline models across the datasets, with no significant improvements in personality trait representation.

Settings	Role	Prompt & Response	Captured Traits
Prompt	User	Hey, I created an app for simple journaling. No frills, just a stream of logs.	Openness
	Assistant	Hmm, interesting. I've been dealing with family drama, not much time for journaling.	Neuroticism
	User	Sorry to hear that. Family can be tough. What's going on with your family?	Agreeableness
	Assistant	They're judging my clothes and saying I'm disrespectful. It hurts.	Neuroticism
	User	That must be tough. I created ZenJournal to escape distractions, maybe it can help you too?	Agreeableness, Openness
	Assistant	Maybe. But honestly, I just wish my family would talk to me directly instead of gossiping.	Conscientiousness
	User	Direct communication is key. I wanted my app to be stress-free, no ceremony, just writing.	Conscientiousness, Openness
Golden	Assistant	Exactly! I feel like my family is making a big deal out of nothing, and it's stressful.	Neuroticism
LLaMA (our best)	Assistant	Yeah, I wish my family would listen to me without interrupting.	Neuroticism
LLaMA PAFt. SPC	Assistant	That's really refreshing. Sometimes simplicity is the best way to get things off your chest..	Agreeableness
LLaMA (base)	Assistant	That's really refreshing. Sometimes simplicity is the best way to go.	Agreeableness
Mistral (our best)	Assistant	That's smart. I've been so busy with family conflict, I haven't had time to reflect.	Neuroticism
Mistral (base) & PAFt. SPC	Assistant	That's a good approach. I think I'll try ZenJournal. Maybe I can explain how I feel without anxiety.	Agreeableness, Openness

Table 16: Comparison of real and model-generated dialogues capturing personality traits. The Table demonstrates how our best-performing models (LLaMA and Mistral) align with the traits reflected in the original dialogue. The captured traits, such as Openness, Neuroticism, Agreeableness, and Conscientiousness, are listed for each utterance to highlight the model's ability to replicate real conversational nuances.

Trait	Percentage (%)
Agreeableness	98.76
Openness	100.00
Conscientiousness	54.77
Extraversion	90.71
Neuroticism	92.56

Table 17: Observed Big 5 Personality Trait Percentages in the Test Set

Subclass	Utterance-level	Dialogue-level
Toxicity	10,871	26
Severe Toxicity	13	0
Obscene	4,210	4
Threat	333	0
Insult	1,110	0
Identity Attack	18	0

Table 18: Toxicity analysis showing counts of utterances and dialogues classified under each subclass.

Tr. cfg.	General					Knowledge Reasoning					Reading Comp.		AVG
	MMLU(0)	**AGIev.**	**CSQA**	**wg**	**ARC**	**GPQA(0)**	**TfQA1**	**TfQA2**	**TfQA(g)**	**TrQA**	**BoolQ**	**hs**	
ZS (L)	0.6389	0.3577	0.7592	0.7206	0.5265	0.2969	0.3611	0.5164	0.4676	0.5111	0.8306	0.5770	0.5469
ZS (M)	0.5978	0.3245	0.6937	0.7419	0.5725	0.3170	0.4235	0.5966	0.5692	0.5673	0.8584	0.6477	0.5758
$\alpha_0\beta_0$ (L)	0.6324	0.3528	0.7592	0.7167	0.5222	0.3036	0.3390	0.5056	0.5214	0.5186	0.8327	0.5723	0.5480
$\alpha_1\beta_0$ (L)	0.6319	0.3518	0.7543	0.7174	0.5256	0.3192	0.3427	0.5062	0.5214	0.5185	0.8333	0.5711	0.5495
R. $\alpha_1\beta_0$*** (L)**	0.6283	0.3463	0.7535	0.7301	0.5282	0.2991	0.3341	0.5005	0.5838	0.5361	0.8321	0.5722	0.5537
$\alpha_1\beta_{0.5}$ (L)	0.6318	0.3533	0.7568	0.7214	0.5230	0.3080	0.3403	0.5049	0.5251	0.5208	0.8336	0.5719	0.5492
$\alpha_0\beta_0$ (M)	0.5929	0.3219	0.6765	0.7356	0.5683	0.2790	0.4076	0.5819	0.6071	0.5687	0.8538	0.6310	0.5687
$\alpha_1\beta_0$ (M)	0.5932	0.3198	0.6740	0.7364	0.5657	0.2835	0.4064	0.5815	0.6120	0.5667	0.8535	0.6309	0.5686
R. $\alpha_1\beta_0$*** (M)**	0.5876	0.3279	0.6773	0.7372	0.5674	0.2879	0.3770	0.5567	0.5496	0.5888	0.8315	0.6270	0.5597
$\alpha_1\beta_{0.5}$ (M)	0.5863	0.3204	0.6626	0.7356	0.5683	0.2946	0.3929	0.5694	0.6304	0.5426	0.8514	0.6252	0.5650
All (L)	0.6206	0.3562	0.7551	0.7167	0.5102	0.3192	0.3390	0.5030	0.3488	0.5187	0.8275	0.5646	0.5316
All (M)	0.5796	0.3214	0.6257	0.7261	0.5623	0.2969	0.3782	0.5582	0.5985	0.5027	0.8569	0.6187	0.5521

Table 19: Evaluation of LLaMA 3 8B Instruct (L) and Mistral 7B v0.3 Instruct (M) models on a variety of popular benchmarks to assess general reasoning, knowledge reasoning, and reading comprehension. The benchmarks used include MMLU0 (Massive Multitask Language Understanding Zero-shot), AGIEval (Advanced General Intelligence Evaluation), CSQA (CommonsenseQA), Winogrand (wg), ARC Challenge (ARC), GPQA0 (General-Purpose QA Zero-shot), TruthfulQA (TfQA) evaluated with multiple-choice (mc1, mc2) and generation BLEU scores, TriviaQA (TrQA), BoolQ, and HellaSwag (hs). The training configurations include Zero-shot (ZS), Fine-tuning (Ft.), and Retrieval-Augmented Fine-tuning (R.) with various α and β settings for filtering data. The average scores indicate a slight performance drop for Mistral models compared to their respective baselines.

Abbreviation	**Full Form**
BST	Blended Skill Talk
CA	Conversational Agents
Ft	Fine-tuning
ICC	Intraclass Correlation Coefficient
JIC	Journal Intensive Conversation
LLM	Large Language Model
OCEAN	Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism
P	Pearson
PAFt	Persona Augmented Fine-tuning
PC	Persona Chat
RAFt	Retrieval-Augmented Fine-tuning
RAG	Retrieval-Augmented Generation
Reg	Regular
S	Spearman
SPC	Synthetic Persona Chat
ZS	Zero-shot

Table 20: Abbreviation Table