

Beyond Discrete Personas: Personality Modeling Through Journal Intensive Conversations

Sayantana Pal, Souvik Das, Rohini K. Srihari
Department of Computer Science and Engineering
State University of New York at Buffalo

Research Goal

The goal of this research is to create a **conversational dataset** that authentically captures the **dynamic** and **evolving** nature of **human personalities**, overcoming the **limitations of static personas** in existing datasets like Persona-Chat, Synthetic Persona Chat, Blended Skill talk.

We introduce the **Journal Intensive Conversations (JIC) dataset**, with 400,000 Reddit-based dialogues (16 utterance each). Our framework leverages **clustering** and **personality trait filtering** to ensure dialogues authentically reflect the Big Five traits, resulting in a dataset that **improves personality-driven dialogue generation by 11% on average**.



Motivation

- Existing datasets with **static personas** often produce **shallow, repetitive, and inconsistent dialogues**, failing to engage users authentically.
- We want to capture the **dynamic and evolving nature** of the **human personalities**.

GOLDEN RESPONSE

Exactly! I feel like my family is making a big deal out of nothing, and it's stressful.

CAPTURED GOLDEN TRAITS: Neuroticism

SIGNIFICANT

Agreement

Stressed Frustrated

DIALOGUE CONTEXT

Two individuals are having a conversation, where one shares their creation of a journaling app called **ZenJournal**. Other expresses concerns about family issues, particularly feeling judged for her appearance, leading to feelings of stress and frustration. The conversation revolves around dealing with these family struggles and how journaling could provide relief through direct communication and personal reflection

User: Hey, I created an app for simple journaling. No frills, just a stream of logs.

Assistant: Hmm, interesting. I've been dealing with family drama, not much time for journaling.

User: Sorry to hear that. Family can be tough. What's going on with your family?

Assistant: They're judging my clothes and saying I'm disrespectful. It hurts.

User: That must be tough. I created ZenJournal to escape distractions, maybe it can help you too?

Assistant: Maybe. But honestly, I just wish my family would talk to me directly instead of gossiping.

User: Direct communication is key. I wanted my app to be stress-free, no ceremony, just writing.

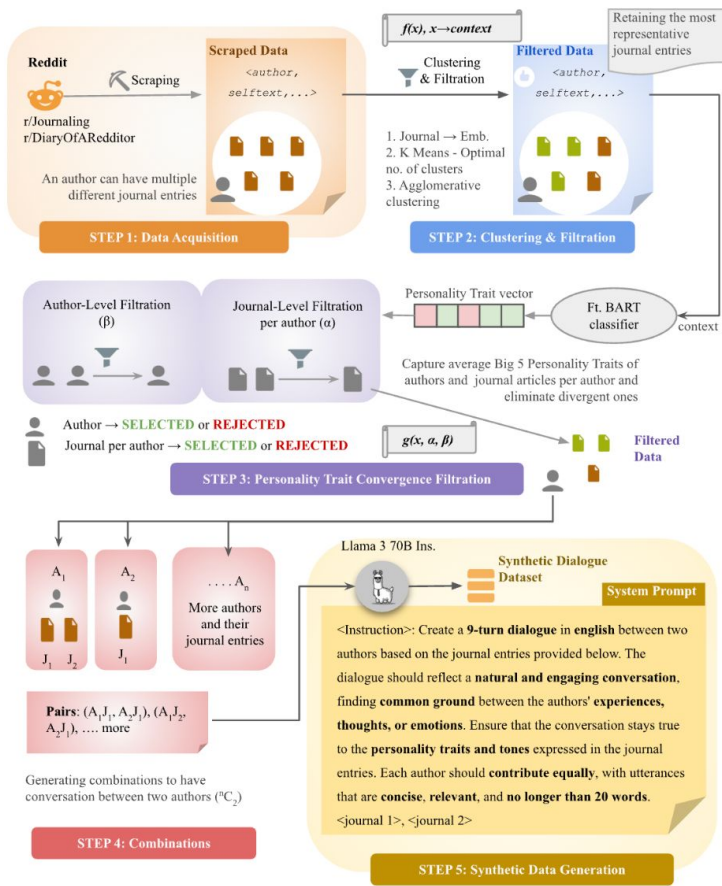
Assistant:

MODEL RESPONSES

✘ Baseline Model (Pre-trained)	✘ Model Ft. Existing Dataset (PC)	✔ Model Ft. JIC Dataset
<p>Response: That's really refreshing. Sometimes simplicity is the best way to go.</p> <p>✔</p> <p>CAPTURED TRAIT: Agreeableness</p>	<p>Response: That's a good approach. I think I'll try ZenJournal. Maybe I can explain how I feel without anxiety.</p> <p>✔</p> <p>CAPTURED TRAIT: Agreeableness</p>	<p>Response: Yeah, I wish my family would listen to me without interrupting.</p> <p>✔</p> <p>CAPTURED TRAIT: Neuroticism</p>

Big 5 Personality Traits: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism

Synthetic Data Generation



SYNTHETIC DATA GENERATION PROCESS

SAMPLE DIALOGUE

CONTEXT 1: Author 1, Journal 1 (user)

<...> A while ago I wanted to find a **journaling app** with no BS: <...> just a simple **stream of logs** that I can open and type right away. I couldn't find any. So I created this **iOS/Android app** called **ZenJournal**. I'd love to hear your thoughts and comments below. <...> It just has one stream, open-n-log, totally **stress-free**, **micro journaling**, **non-ceremonial** and anti-bloat at its core...

CONTEXT 2: Author 2, Journal 1 (assistant)

My **family** recently came out saying they have a **problem** with how I **dress** it's **too revealing** and I shouldn't be dressing like that around them. They have made remarks when I go over telling me to pull my **dress** down <...> I felt comfortable in my own body and they're my family they and wouldn't judge me but the whole time they were talking about me saying I'm **disrespectful** and they don't want me around no more. <...> they all waited to sit and **gossip** about me instead of speaking sooner about it. <...>

Generated Dialogue

User: Hey, I **created an app** for simple **journaling**. **No frills, just a stream of logs.**

Assistant: Hmm, interesting. I've been dealing with **family drama**, not much time for journaling.

User: Sorry to hear that. Family can be tough. What's going on with your family?

Assistant: They're **judging my clothes** and saying I'm **disrespectful**. **It hurts.**

User: That must be tough. I created **ZenJournal** to escape distractions, **maybe it can help you too?**

Assistant: Maybe. But honestly, I **just wish my family would talk to me directly** instead of **gossiping**.

User: **Direct communication** is key. I wanted my app to be **stress-free, no ceremony, just writing**.

Assistant: Exactly! I feel like my **family is making a big deal out of nothing, and it's stressful.**

User: I understand. Sometimes **simplicity is all we need**. **Would you like to try ZenJournal and see if it helps?**

Assistant: Yeah, **maybe it'll help me process my feelings** about this situation.

User: That's the idea! It's just an **open log, no prompts or reminders, just write**.

Assistant: I like that. I think I'll **try it** and see if it helps me clear my head.

User: **Let me know what you think!** I'm always looking for **feedback to improve the app**.

Assistant: **Will do**. Thanks for listening, it's nice to talk to someone **who's not judging me**.

User: **No judgment here!** I just want to **help people express themselves freely**.

Assistant: That's really nice. I think I'm going to **write in your app now**, thanks for the chat.

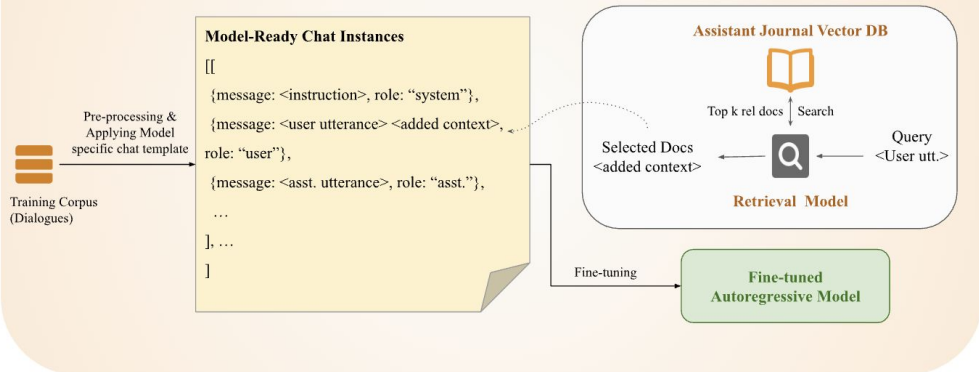
- Openness:** Curiosity and openness to new experiences.
- Conscientiousness:** Organized, responsible, and goal-directed behavior.
- Extraversion:** Sociable, outgoing, and energetic demeanor.
- Agreeableness:** Compassionate, cooperative, and empathetic nature.
- Neuroticism:** Prone to emotional instability and stress.
- Emphasized Text:** Knowledge grounding

Model Training and Inference Settings

Training Settings

Setting 1: Fine-tuning Autoregressive LMs (without additional context)

Setting 2: Retrieval Augmented Fine-tuning Autoregressive LMs (with additional context)



Inference also had two settings:

- Utterance-level
- Retrieval-Augmented Generation (RAG)

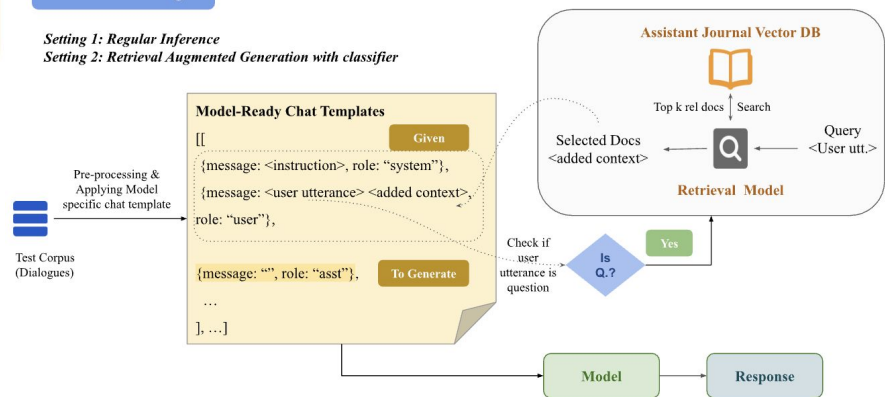
Training was conducted in two settings:

- Standard fine-tuning
- Retrieval-Augmented Fine-tuning (RAft)

Inference Settings

Setting 1: Regular Inference

Setting 2: Retrieval Augmented Generation with classifier



Comparison of dialogues

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

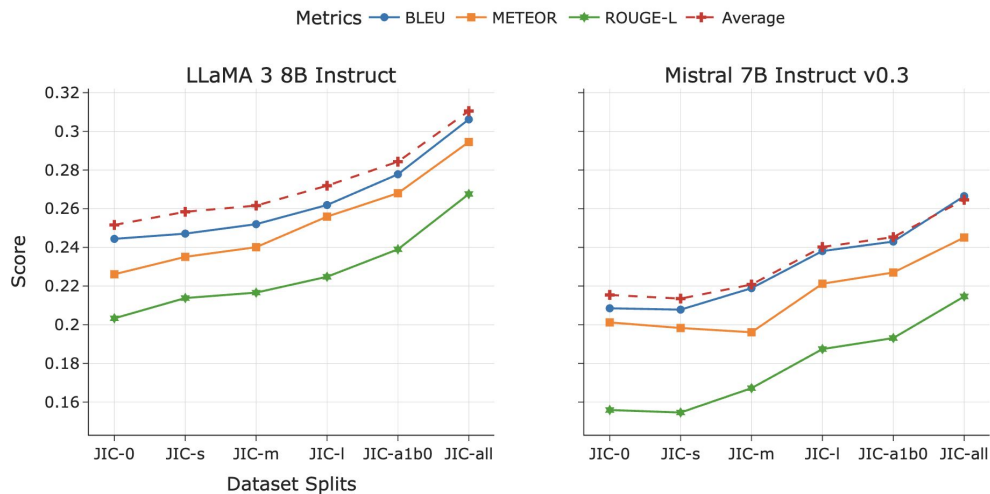
Settings	Role	Prompt & Response	Captured Traits
Golden	Assistant	Exactly! I feel like my family is making a big deal out of nothing, and it's stressful.	Neuroticism
LLaMA (our best)	Assistant	Yeah, I wish my family would listen to me without interrupting.	Neuroticism
LLaMA PAFt. SPC	Assistant	That's 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

Evaluation Metrics

- **Automated Metrics:**
 - BLEU, METEOR, ROUGE (1, 2, L), and BERTScore - evaluated dialogue coherence and relevance
- **Personality-Based Metrics:**
 - LM Eval Harness Benchmark by EleutherAI - assessed alignment with the Big Five traits (openness, conscientiousness, extraversion, agreeableness, neuroticism)



Results



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

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

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.

Results

Avg. Personality Trait score
improvement over baselines.

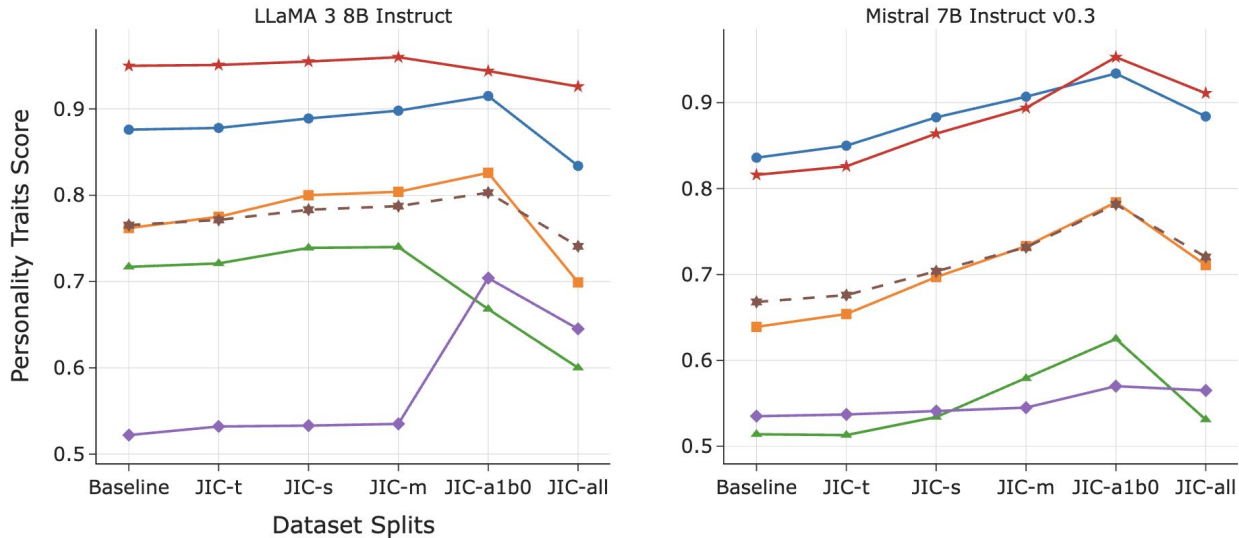
- **LLaMA 8B ins**
 - Baseline - 76.54%
 - Our - **80.30%**
- **Mistral 7B ins**
 - Baseline - 66.80%
 - Our - **78.16%**

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	<u>0.7400</u>	0.9600	0.5350	0.7874
	Mistral	Ft. RAFt.	29k 29k	0.8670 0.9070	0.6820 0.7330	0.5220 0.5790	0.8600 0.8940	0.5380 0.5450	0.6938 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	<u>0.9440</u>	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
JIC-all	LLaMA	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

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.

Results

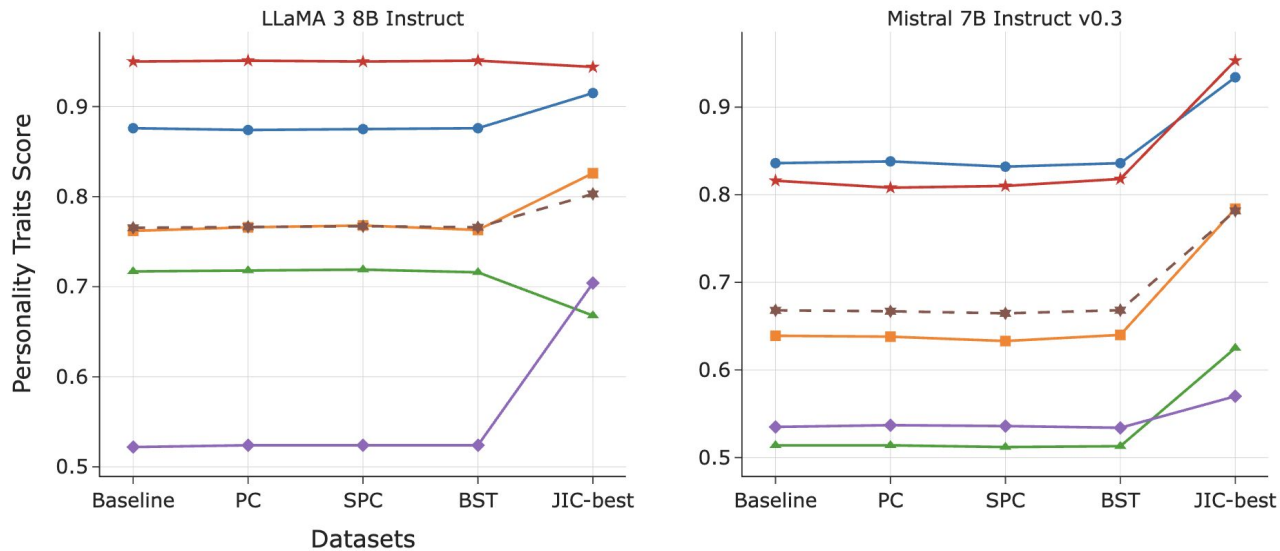
Big 5 P-Traits — Openness — Conscientiousness — Extraversion — Agreeableness — Neuroticism — AVG



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

Results

Big 5 P-Traits — Openness — Conscientiousness — Extraversion — Agreeableness — Neuroticism — AVG



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

Discussion

- **Non-linear Scaling of Personality Trait Accuracy:** Personality trait accuracy **does not scale linearly** with dataset size, primarily **due to biases in the dataset**, such as the **over-representation of neuroticism** in Reddit journals.
- **Trait-Specific Model Behavior:** Fine-tuning without strict filtering (α , β) resulted in a higher emphasis on neuroticism, disrupting trait balance, while **filtering parameters ($\alpha = 1$, $\beta = 0$) offered the most balanced results across traits**.
- **Impact on General Reasoning:** Focusing on personality traits **slightly reduced** the model's general reasoning capabilities, indicating a potential trade-off that might be addressed through advanced techniques like **rehearsal learning**.

Conclusion & Limitations

- **Conclusion:** The JIC dataset captures **dynamic human personalities**, addressing limitations of static personas in conversational AI. Fine-tuning LLMs on this dataset **improved personality-driven dialogue generation by 11%, enabling more engaging interactions.**
- **Limitations**
 - **Bias from Reddit data**, particularly the **over-representation of neuroticism**, affects personality trait balance.
 - **Computationally intensive tuning of α and β parameters** for personality filtering.
 - Synthetic data risks **propagating biases from pre-trained models**, limiting diversity.

Thanks!

Read the paper [here](#)

Scan the QR to read the paper

