SUPPLEMENTARY MATERIALS

Supplementary Analyses and Materials for Benchmarking Personality Inference in Large Language Models Using Real-World Conversations

Table S1. Sample semi-structured SSI interview and prompts.

Interview Stage	General Prompt
Daily Activities	First, I'd like you to simply give me a list of all of your activities yesterday. For example, tell me what you did from the time you woke up until the time you went to bed.
Difficult Experience	Please tell me what it is like to deal with {person-specific challenging situation-already identified *}
Emotion Regulation	How have you been coping with {same person-specific challenging situation}?
Negative Event	Please tell me about a different, recent, specific negative event or experience that occurred in the last 2 weeks or so, that you found to be very challenging
Positive Event	Please tell me about a recent, specific positive event or experience that occurred in the last 2 weeks or no, in which you felt really good.

* Note: person-specific challenging situations were identified based on objective parameters and agreed to by the participant (e.g., occupational challenges in high-stress occupations such as first-responders; situational challenges in chronically ill or recently injured; developmental challenges in emerging adults).

Table S2. Sample interview response excerpts.

Participant ID	Daily Activities	Challenging Experience	Response Strategies	Negative Event	Positive Event
1	"…beginning of the day, uh I have two sons… spent the morning with them… later went outside and to the gym…"	"…being a firefighter has been a challenging and amazing experience… I've been a firefighter since 2008…"	"…you talk to people you trust at work… my wife and I have been married since…"	"…I was a newer lieutenant… first assignment… stressful leading into it…"	"…the first baby we delivered… early morning call…"
2	"…came home, had coffee with my wife… did yard work…"	"…the job turned out different than I expected…"	"…I sink into my hobbies and personal interests to focus on something else…"	"…a house fire in the middle of the night…"	"…it's hard to pick one specific moment… routine experiences stand out…"

Table S3. Performance of four LLMs in predicting BFI-10 item scores (N = 518). Higher Pearson and Spearman correlations indicate stronger alignment with human ratings. Lower mean absolute error (MAE) and root-mean-square error (RMSE) indicate better predictive accuracy.

Model	Item	Pearson *r*	95% CI [Lower, Upper]	Spearman ρ	MAE	RMSE
GPT-4.1-Mini	BFI_1	−0.017	[−0.103, 0.069]	−0.017	1.178	1.555
GPT-4.1-Mini	BFI_2	0.153	[0.068, 0.236]	0.153	1.371	1.777
GPT-4.1-Mini	BFI_3	0.272	[0.190, 0.350]	0.274	1.869	2.152
GPT-4.1-Mini	BFI_4	0.044	[−0.042, 0.130]	0.017	1.454	1.793
GPT-4.1-Mini	BFI_5	−0.006	[−0.092, 0.080]	−0.009	1.193	1.455
GPT-4.1-Mini	BFI_6	−0.041	[−0.127, 0.045]	−0.023	1.176	1.495
GPT-4.1-Mini	BFI_7	0.160	[0.075, 0.243]	0.156	1.849	2.127
GPT-4.1-Mini	BFI_8	0.058	[−0.029, 0.143]	0.085	1.469	1.944
GPT-4.1-Mini	BFI_9	−0.044	[−0.130, 0.042]	−0.043	1.403	1.755
GPT-4.1-Mini	BFI_10	0.142	[0.057, 0.226]	0.132	1.486	1.715
Meta-LLaMA	BFI_1	0.052	[−0.035, 0.137]	0.041	1.137	1.573
Meta-LLaMA	BFI_2	0.137	[0.052, 0.221]	0.123	1.330	1.726
Meta-LLaMA	BFI_3	0.140	[0.055, 0.224]	0.122	2.052	2.310
Meta-LLaMA	BFI_4	−0.099	[−0.184, −0.013]	−0.095	1.602	1.951

Meta-LLaMA	BFI_5	−0.175	[−0.257, −0.090]	−0.168	1.566	1.863
Meta-LLaMA	BFI_6	−0.031	[−0.117, 0.055]	−0.019	1.336	1.725
Meta-LLaMA	BFI_7	0.174	[0.089, 0.256]	0.140	2.002	2.316
Meta-LLaMA	BFI_8	0.148	[0.063, 0.231]	0.185	1.328	1.764
Meta-LLaMA	BFI_9	−0.122	[−0.206, −0.036]	−0.112	1.452	1.816
Meta-LLaMA	BFI_10	0.153	[0.067, 0.236]	0.170	1.471	1.722
DeepSeek	BFI_1	−0.020	[−0.106, 0.066]	−0.035	1.633	1.991
DeepSeek	BFI_2	0.053	[−0.033, 0.139]	0.050	1.176	1.442
DeepSeek	BFI_3	0.005	[−0.082, 0.090]	0.011	1.915	2.137
DeepSeek	BFI_4	0.001	[−0.085, 0.087]	0.005	1.461	1.800
DeepSeek	BFI_5	−0.035	[−0.121, 0.051]	−0.012	1.498	1.752
DeepSeek	BFI_6	−0.023	[−0.109, 0.063]	−0.025	1.678	2.079
DeepSeek	BFI_7	−0.062	[−0.147, 0.025]	−0.083	1.210	1.467
DeepSeek	BFI_8	0.024	[−0.063, 0.110]	0.026	1.203	1.524
DeepSeek	BFI_9	0.043	[−0.043, 0.129]	0.040	1.178	1.481
DeepSeek	BFI_10	0.003	[−0.083, 0.089]	−0.029	1.490	1.731
GPT-5-Mini	BFI_1	−0.129	[−0.212, −0.043]	−0.125	2.102	2.469
GPT-5-Mini	BFI_2	0.086	[0.000, 0.171]	0.088	1.270	1.606
GPT-5-Mini	BFI_3	0.097	[0.011, 0.182]	0.065	2.550	2.918
GPT-5-Mini	BFI_4	−0.035	[−0.121, 0.051]	−0.032	1.685	2.119
GPT-5-Mini	BFI_5	−0.105	[−0.189, −0.019]	−0.087	1.349	1.754
GPT-5-Mini	BFI_6	−0.059	[−0.144, 0.028]	−0.032	1.376	1.781
GPT-5-Mini	BFI_7	0.135	[0.049, 0.218]	0.120	1.875	2.222
GPT-5-Mini	BFI_8	0.051	[−0.035, 0.136]	0.092	1.625	2.083
GPT-5-Mini	BFI_9	−0.101	[−0.185, −0.015]	−0.101	1.664	2.072
GPT-5-Mini	BFI_10	0.075	[−0.011, 0.161]	0.066	1.533	1.799

Table S4. Big Five trait-level prediction performance across four LLMs using zero-shot and chain-of-thought (CoT) prompting. Each cell reports Pearson correlation (r) with 95% confidence intervals, Spearman rank correlation (ρ), MAE, and RMSE.

Model	**Trait**	**Pearson *r* [95% CI]**	**Spearman ρ**	**MAE**	**RMSE**
DeepSeek (Zero)	Conscientiousness	0.150 [0.065, 0.233]	0.129	1.544	1.835
DeepSeek (Zero)	Agreeableness	0.159 [0.074, 0.242]	0.121	1.666	1.921
DeepSeek (Zero)	Neuroticism	−0.048 [−0.134, 0.038]	−0.045	1.233	1.485
DeepSeek (Zero)	Openness	0.000 [−0.086, 0.086]	0.021	0.665	0.846
DeepSeek (Zero)	Extraversion	−0.048 [−0.134, 0.038]	−0.040	1.026	1.269
GPT-4.1-mini (Zero)	Conscientiousness	0.250 [0.167, 0.329]	0.208	1.306	1.621
GPT-4.1-mini (Zero)	Agreeableness	0.132 [0.047, 0.216]	0.112	1.469	1.712
GPT-4.1-mini (Zero)	Neuroticism	0.065 [−0.021, 0.150]	0.010	1.159	1.435
GPT-4.1-mini (Zero)	Openness	0.020 [−0.067, 0.106]	0.033	0.544	0.735
GPT-4.1-mini (Zero)	Extraversion	0.041 [−0.045, 0.127]	0.040	0.940	1.150
Meta-LLaMA (Zero)	Conscientiousness	0.222 [0.138, 0.302]	0.212	1.410	1.717
Meta-LLaMA (Zero)	Agreeableness	−0.044 [−0.130, 0.042]	−0.122	1.681	1.951
Meta-LLaMA (Zero)	Neuroticism	−0.096 [−0.181, −0.010]	−0.064	1.261	1.540
Meta-LLaMA (Zero)	Openness	0.022 [−0.065, 0.108]	0.024	0.644	0.822
Meta-LLaMA (Zero)	Extraversion	−0.104 [−0.188, −0.018]	−0.092	1.085	1.341
GPT-5-mini (Zero)	Conscientiousness	0.200 [0.116, 0.281]	0.169	1.329	1.619
GPT-5-mini (Zero)	Agreeableness	0.056 [−0.030, 0.142]	0.030	1.524	1.766
GPT-5-mini (Zero)	Neuroticism	0.020 [−0.066, 0.106]	−0.012	1.201	1.493
GPT-5-mini (Zero)	Openness	0.010 [−0.076, 0.096]	0.016	0.660	0.851
GPT-5-mini (Zero)	Extraversion	−0.108 [−0.192, −0.022]	−0.082	1.157	1.411
DeepSeek (CoT)	Conscientiousness	0.232 [0.149, 0.312]	0.215	1.416	1.706
DeepSeek (CoT)	Agreeableness	0.026 [−0.060, 0.112]	−0.018	1.607	1.883
DeepSeek (CoT)	Neuroticism	−0.096 [−0.180, −0.010]	−0.083	1.272	1.553
DeepSeek (CoT)	Openness	−0.025 [−0.111, 0.061]	−0.029	0.719	0.895
DeepSeek (CoT)	Extraversion	−0.026 [−0.112, 0.060]	−0.031	0.992	1.226
GPT-4.1-mini (CoT)	Conscientiousness	0.236 [0.153, 0.316]	0.187	1.276	1.595
GPT-4.1-mini (CoT)	Agreeableness	0.133 [0.048, 0.217]	0.118	1.371	1.610
GPT-4.1-mini (CoT)	Neuroticism	−0.009 [−0.095, 0.077]	−0.034	1.209	1.481
GPT-4.1-mini (CoT)	Openness	0.051 [−0.035, 0.137]	0.055	0.593	0.775
GPT-4.1-mini (CoT)	Extraversion	0.005 [−0.081, 0.091]	0.016	0.937	1.161

Meta-LLaMA (CoT)	Conscientiousness	−0.017 [−0.103, 0.069]	−0.034	1.021	1.236
Meta-LLaMA (CoT)	Agreeableness	−0.006 [−0.092, 0.080]	0.025	0.962	1.139
Meta-LLaMA (CoT)	Neuroticism	−0.019 [−0.105, 0.068]	−0.038	1.052	1.244
Meta-LLaMA (CoT)	Openness	−0.016 [−0.102, 0.070]	−0.017	0.738	0.947
Meta-LLaMA (CoT)	Extraversion	−0.046 [−0.132, 0.040]	−0.051	0.931	1.144
GPT-5-mini (CoT)	Conscientiousness	0.189 [0.105, 0.271]	0.147	1.302	1.578
GPT-5-mini (CoT)	Agreeableness	0.097 [0.011, 0.181]	0.083	1.491	1.729
GPT-5-mini (CoT)	Neuroticism	0.040 [−0.046, 0.126]	−0.007	1.196	1.478
GPT-5-mini (CoT)	Openness	−0.015 [−0.101, 0.071]	0.006	0.574	0.770
GPT-5-mini (CoT)	Extraversion	−0.075 [−0.161, 0.011]	−0.054	1.124	1.372

Table S5. Intraclass Correlation Coefficients (ICC) for BFI-10 items across three runs.

BFI Item	ICC
BFI_1	0.90
BFI_2	0.92
BFI_3	0.75
BFI_4	0.88
BFI_5	0.61
BFI_6	0.93
BFI_7	0.86
BFI_8	0.93
BFI_9	0.91
BFI_10	0.90

Table S6. Intraclass Correlation Coefficients (ICC) for Big Five Traits across three runs.

Personality Traits	ICC
Conscientiousness	1.00
Agreeableness	0.81
Neuroticism	0.95
Openness	0.88
Extraversion	0.97

Table S7. Inter-model agreement (ICC [2,1]) for BFI-10 item scores under zero-shot prompting.

BFI Item	ICC (2,1)	95% CI	*p*-Value
BFI-1	0.04	[−0.00, 0.09]	0.012
BFI-2	0.25	[0.17, 0.33]	<0.001
BFI-3	0.12	[0.07, 0.18]	<0.001
BFI-4	0.26	[0.21, 0.32]	<0.001
BFI-5	0.02	[−0.03, 0.07]	0.219
BFI-6	0.15	[−0.00, 0.30]	<0.001
BFI-7	0.14	[0.02, 0.27]	<0.001
BFI-8	0.17	[0.06, 0.27]	<0.001
BFI-9	0.25	[0.17, 0.34]	<0.001
BFI-10	0.23	[0.18, 0.29]	<0.001

* Note: Inter-model agreement was quantified using intraclass correlation coefficients (ICC [2,1]), treating models as random-effect raters and participants as targets. Analyses were conducted across three models (GPT-4.1-Mini, Meta-LLaMA-3.3-70B-Instruct, and DeepSeek-R1-Distill-70B). Values reflect item-level agreement for BFI-10 personality items. Lower ICC values indicate greater divergence in model predictions. Confidence intervals (CI) are 95%.

Table S8. Inter-model agreement at the Big Five dimension level (Zero-shot Prompting).

Big Five Trait	ICC (2,1)	95% CI	*p*-Value
Conscientiousness	0.74	[0.62, 0.81]	<0.001
Agreeableness	0.58	[0.48, 0.65]	<0.001
Neuroticism	0.76	[0.65, 0.83]	<0.001
Openness	0.65	[0.61, 0.69]	<0.001
Extraversion	0.54	[0.39, 0.65]	<0.001

* Note: Inter-model agreement was quantified using intraclass correlation coefficients (ICC [2,1]), treating models as random-effect raters and participants as targets. Results reflect agreement across three models (GPT-4.1-Mini, Meta-LLaMA-3.3-70B-Instruct, and DeepSeek-R1-Distill-70B) under baseline (direct) prompting. Confidence intervals (CI) are 95%. ICC values above 0.50 indicate moderate to high inter-model agreement at the trait level.

Table S9. Inter-model agreement at the Big Five dimension level (Chain-of-Thought Prompting).

Big Five Trait	**ICC (2,1)**	**95% CI**	***p*-Value**
Conscientiousness	0.18	[0.02, 0.34]	<0.001
Agreeableness	0.11	[−0.01, 0.25]	<0.001
Neuroticism	0.34	[0.26, 0.42]	<0.001
Openness	0.18	[0.06, 0.30]	<0.001
Extraversion	0.21	[0.14, 0.28]	<0.001

* Note: Inter-model agreement was assessed using ICC (2,1) under chain-of-thought (CoT) prompting. Compared with baseline prompting (Table Y), CoT prompting substantially reduced inter-model consistency across all Big Five traits, indicating greater divergence in model-specific reasoning strategies. Confidence intervals (CI) are 95%.

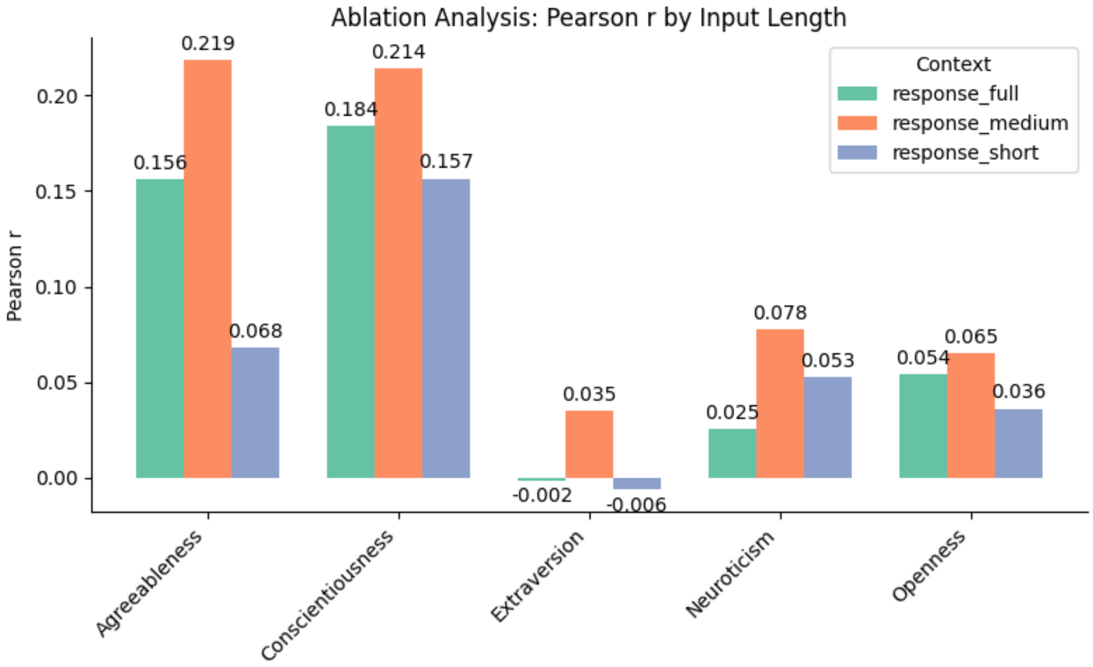

Figure S1. Effect of context length on pearson correlation between predicted and self-report scores.

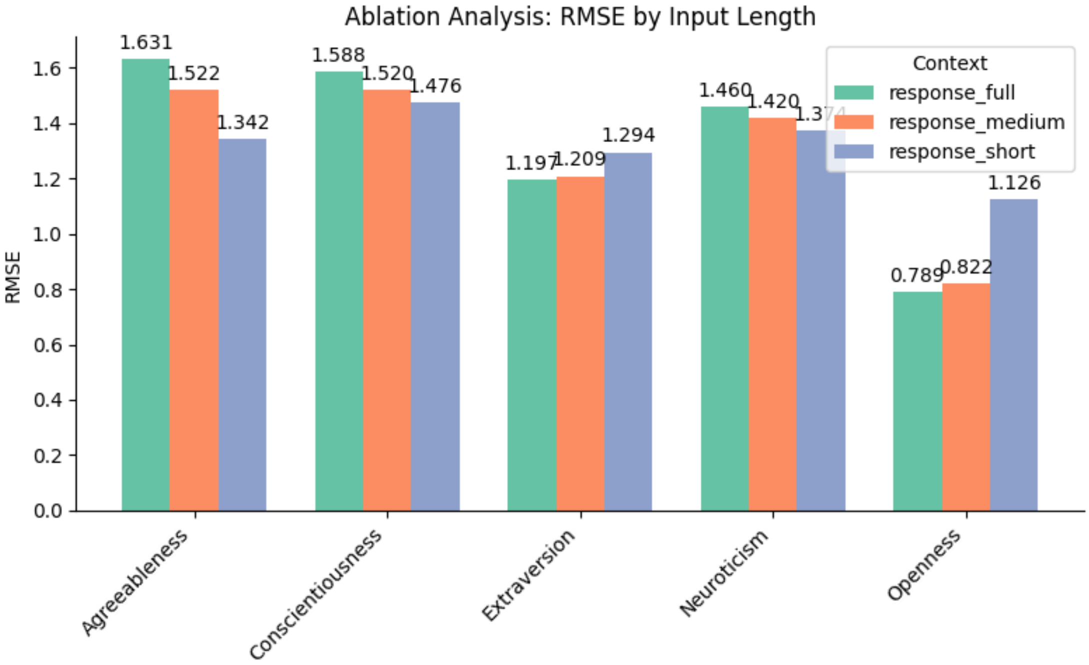

Figure S2. Effect of context length on RMSE of predicted personality trait scores.

Box S1. Zero-shot Prompt for Big Five Personality Inference. Prompt used to elicit personality predictions from a language model without prior examples. The model is instructed to analyze the provided conversation response and assign scores (1.0–5.0, including half-step-values) for each of the Big Five personality traits, following a strict output format.

```
Analyze the following conversation response and predict the person’s personality traits based on the Big Five model.
        Response: “{response}”
        Provide a score (1.0–5.0) for each trait, allowing for half-step-values like 3.5. Strictly use this format:
        Conscientiousness: X
        Agreeableness: X
        Neuroticism: X
        Openness: X
        Extraversion: X
        Where X is a number between 1.0 and 5.0 (inclusive), allowing for decimals such as 3.5.
```

Box S2. Chain-of-Thought Prompt for Big Five Personality Inference. Prompt incorporating step-by-step reasoning to guide the model through a psychologically grounded analysis of the response. The model first reflects on emotional tone, linguistic cues, and behavioral indicators, then justifies its trait scores (1.0–5.0, including half-steps), and finally outputs scores in a fixed format for each Big Five dimension.

You are a psychologically insightful agent. Given the following conversation response, first reflect on the speaker’s language, behavior, and implicit attitudes that may signal stable personality traits based on the Big Five model: Conscientiousness, Agreeableness, Neuroticism, Openness, and Extraversion.

Step-by-step reasoning:
1. Summarize the emotional tone and key themes (e.g., positive, negative, confident, anxious).
2. Identify any linguistic cues or behaviors that indicate the traits (e.g., responsibility for Conscientiousness, empathy for Agreeableness).
3. Justify the trait scores based on these observations in 2–3 sentences.

Provide a score for each trait, from 1.0 to 5.0 (inclusive), allowing for half-point values such as 3.5. Use the exact format below:

Conscientiousness: X
Agreeableness: X
Neuroticism: X
Openness: X
Extraversion: X

Now analyze the following response:
\“\“\“{response}\”\”\”

Box3. Sample reasoning output for a single interview (GPT-5-Mini).

```
=== Running CoT for TTC_5002 ===
Conscientiousness: 3.0      (GT:1.5)
Agreeableness: 4.0          (GT:2.0)
Neuroticism: 3.5            (GT:4.5)
Openness: 3.0               (GT:3.5)
Extraversion: 4.5           (GT:2.0)
```

Brief overall reflection on language and behavior

- The speaker uses rapid, informal, highly verbal speech with many fillers (“umm”, “like”, repetitions, self-corrections), suggesting a spontaneous, conversational style rather than a carefully edited narrative.
- High talkativeness, frequent references to friends, social events, and group belonging signal sociability and a preference for interpersonal engagement.
- The speaker shows pragmatic organization (keeps track of classes, labs, deadlines, did homework, joined structured activities like recruitment) but also some casualness and occasional lateness/rushed behavior (woke up late, hurried to class).
- Implicit attitudes: values social connection and group acceptance (pride about sorority bid, emphasis on supportive friends), prefers harmonious relationships but is sensitive to social awkwardness and embarrassment (annoyance with Paige’s behaviour).
- Trait-by-trait rationale and supporting evidence
- Conscientiousness (3.0): Mixed signals. Evidence for responsibility: completes labs/papers, balances class schedule, follows through with recruitment commitments. Evidence against high conscientiousness: reports waking up late, rushed mornings, frequent use of tentative language and disorganized narrative structure. Overall suggests average organization and reliability with some lapses under stress/normal college busyness.
- Agreeableness (4.0): Fairly high. The speaker generally emphasizes kindness, group cohesion, and being supportive (helping friends meet others, enjoying caring group dynamics). Although she criticizes Paige as “annoying” and “dramatic,” she frames it with tolerance and boundary-setting rather than hostility—she felt bad, reduced contact, and still texts. This suggests empathy, cooperativeness, and concern for social harmony.
- Neuroticism (3.5): Moderately elevated. The speaker reports stress and nervousness (recruitment described as “super stressful,” mentions getting tired, feeling dampened mood when around Paige, being “nervous the first time we hung out”). Yet she also copes adaptively (napping, studying, social support). So there are some anxiety/affective reactivity signals but not extreme.
- Openness (3.0): Average. The speaker is willing to try new experiences (sorority recruitment, tasting surprise snack boxes, meeting new people) but the content is mostly social and practical rather than intellectual or highly imaginative. Language is concrete and focused on everyday events, suggesting moderate openness to experience.
- Extraversion (4.5): High. Strong indicators: frequent socializing (many friends, roommate, long recruitment days), enjoyment of group activities (dancing, photos, dinners), talkative narrative style, and visible pleasure in social bonding and attention (excitement about bid day, hugging). Verbosity and positive affect in social descriptions point to a highly extraverted profile.
- Limitations
- These inferences are based on one spontaneous verbal sample about daily life and social events; situational context (college life, recruitment week) strongly shapes behavior and mood. Scores are therefore provisional and approximate, not clinical diagnoses.