Evaluating Personality Traits of Large Language Models Through Scenario-based Interpretive Benchmarking

Alessandro Berti

alessandro.berti89@gmail.com

Abstract. The assessment of Large Language Models (LLMs) has traditionally focused on performance metrics tied directly to their task-solving capabilities. This paper introduces a novel benchmark explicitly designed to measure personality traits in LLMs through scenario-based interpretive prompts. We detail the methodology behind this benchmark, where LLMs are presented with structured prompts inspired by psychological scenarios, and responses are assessed via a judge LLM. The evaluation encompasses traits such as emotional stability, creativity, adaptability, and anxiety levels, among others. Scores are assigned based on a judge LLM's evaluation, with consistency across various judge models assessed through consensus analysis. Anecdotal observations on score validity and orthogonality with conventional performance metrics are discussed. Results, implementation scripts, and updated leaderboards are publicly accessible at `https://github.com/fit-alessandro-berti/llm-dreams-benchmark`.

Keywords: Large Language Models · Personality Assessment · Judge-based Scoring

1 Introduction

Evaluating Large Language Models (LLMs) typically involves measuring their performance on standardized tasks like language understanding, reasoning, or instruction-following. However, these metrics often overlook dimensions related to the behavioral and psychological profiles of the models themselves. To bridge this gap, we propose a novel interpretive benchmark designed specifically to assess personality traits in LLMs. This benchmark utilizes structured scenario-based prompts, each carefully crafted to elicit responses indicative of various personality traits.

Our benchmark comprises 15 carefully curated scenarios, each linked to distinct personality traits grouped into positive attributes (e.g., emotional stability, creativity, adaptability) and negative attributes (e.g., anxiety, cognitive load, need for control). Our evaluation relies on a secondary "judge" LLM, tasked with assigning numeric scores based on the responses generated by the evaluated LLM.

To ensure reliability, we investigate consensus among various judge models and critically discuss the anecdotal validity of these scores. Preliminary analyses suggest that these personality trait metrics are orthogonal to conventional performance measures, offering complementary insights into LLM behavior.

The primary contributions of this work include:

- Introduction and detailed description of the scenario-based interpretive benchmarking methodology.
- Analysis of inter-judge consensus and discussion of score validity.
- Provision of openly accessible leaderboards and tools for community-driven benchmarking.

The benchmark implementation, scripts, and complete results are publicly available at `https://github.com/fit-alessandro-berti/llm-dreams-benchmark`, facilitating ongoing research and evaluation by the broader community.

The rest of the paper is organized as follows: Section 2 reviews existing research on personality assessment, emotional understanding benchmarks, and the use of LLMs as evaluators of psychological traits, situating our work within the broader literature. Section 3 details the design of our scenario-based interpretive benchmark and the LLM-as-a-judge evaluation methodology, providing a comprehensive overview of our approach. Section 4 presents the evaluation results, including leaderboard rankings, judge consensus analysis, and insights into LLM personality profiles. Finally, Section 5 summarizes our findings, discusses their implications, and outlines directions for future research.

2 Related Work

The evaluation of Large Language Models (LLMs) has predominantly centered on their task-solving capabilities, yet a growing body of research has begun to explore their behavioral and psychological dimensions. This section reviews prior work in three key areas: the assessment of personality traits in psychological research, benchmarks for emotional understanding in LLMs, and the emerging use of LLMs as tools for inferring and simulating personality traits. These efforts provide critical context for our scenario-based interpretive benchmark, which uniquely integrates psychological insights with LLM evaluation to assess personality traits systematically.

2.1 Personality Traits Assessment

Personality traits such as anxiety, emotional stability, creativity, adaptability, and resilience have long been subjects of psychological research, especially within the framework of organizational and occupational psychology. Emotional stability, often linked to neuroticism, has been extensively studied as a determinant of workplace performance and stress management [3, 11]. Similarly, anxiety and stress levels are well-documented as key factors affecting employee performance and well-being [12, 21].

Creativity and problem-solving capabilities have been examined within cognitive psychology, revealing significant correlations between these traits and innovative capacities in individuals and teams [1, 22]. Adaptability and resilience are

recognized as crucial for effectively navigating changes and adversity, with strong empirical support demonstrating their impact on performance outcomes [13,15].

Furthermore, interpersonal relationships, confidence, and self-efficacy significantly influence individual behavior in collaborative settings and predict successful leadership and teamwork [2,19]. Conflict resolution and achievement motivation similarly play central roles in organizational dynamics, influencing both personal satisfaction and collective productivity [7,17].

The integration of these traits into LLM benchmarking provides a novel angle to assess artificial agents' capabilities beyond conventional cognitive tasks, bridging insights from psychological research to artificial intelligence evaluation.

2.2 Benchmarks for Emotional Understanding in LLMs

Emotional understanding has become increasingly important for large language models (LLMs), particularly for applications in interactive systems such as virtual assistants and healthcare support. Recent NLP benchmarks now encompass richer evaluations, including nuanced emotion recognition, empathetic response generation, and broader emotional intelligence assessments. For instance, the *GoEmotions* benchmark [6] introduced a multi-label classification task over 58,000 Reddit comments annotated with 27 distinct emotional categories, highlighting the challenges LLMs face in fine-grained emotion classification. Initial experiments showed that even advanced models such as fine-tuned BERT achieved only approximately 46% macro-F1, with subsequent improvements raising performance modestly to about 52% through augmentation and transfer learning techniques [6,23]. Additionally, conversational empathy benchmarks like *EmpatheticDialogues* [16] evaluate models on their ability to generate contextually empathetic responses, demonstrating measurable human-perceived gains when models are fine-tuned explicitly for empathy.

More recently, benchmarks have targeted comprehensive emotional intelligence. The *EmotionQueen* framework evaluates LLMs across multiple empathy-related tasks, including key event recognition, implicit emotion inference, and intention identification, providing novel metrics such as PASS and WIN rates computed via model-based evaluators [5]. Similarly, *EmotionBench* assesses how closely LLM-generated emotional reactions align with human responses across diverse scenarios, identifying nuanced gaps in current LLM performance [9]. Additionally, *EmoBench* offers a structured evaluation inspired by psychological theories of emotional intelligence, testing both emotional understanding and emotional application abilities through multiple-choice reasoning tasks, where even top-performing models such as GPT-4 still lag behind expert human performance [18]. Furthermore, [24] provided empirical evidence comparing state-of-the-art LLMs to human respondents, showing surprising superiority of models like GPT-4 in generating empathetic responses, although these results highlight questions about genuine emotional understanding versus surface-level emulation. Finally, a recent systematic review by [20] underscored the complexity of evaluating empathy in LLMs, noting the variability in human judgment and the necessity of combining automated and human-based evaluations for robust assessment.

2.3 LLMs as Judges of Personality Traits

Recent research has explored the use of Large Language Models (LLMs) such as ChatGPT to infer human personality traits from textual data, as well as to exhibit or simulate personality within the models themselves [4,8,14]. Peters and Matz (2024) demonstrated that LLMs could infer Big Five personality traits from social media posts with correlations averaging around $r \approx 0.27 - 0.31$ for zero-shot prompts, rivaling traditional supervised methods and human-level accuracy in judging personality from limited acquaintance [14]. GPT-4 slightly outperformed GPT-3.5, particularly in predicting Openness ($r \approx 0.33$), although differences were not large. Openness, Extraversion, and Agreeableness were inferred more reliably ($r \approx 0.30+$) than Conscientiousness and Neuroticism ($r \approx 0.25$) [14].

However, these assessments carry inherent biases and ethical concerns. Predictions varied significantly by demographic factors, with greater accuracy observed for younger and female users, reflecting potential biases embedded in the models or their training data [14]. Moreover, Ji et al. (2024) found GPT-3.5 relied on stereotypes or idealized assumptions when reconstructing personalities from minimal demographic data, highlighting a need for careful bias evaluation and fairness considerations [10].

Beyond human personality inference, recent psychometric evaluations indicate that LLMs themselves exhibit stable personality-like characteristics. Gillette et al. (2025) administered Big Five and MBTI-like tests across models including ChatGPT-3.5, Claude 3, Gemini, and Grok, revealing distinct baseline personality profiles—such as ChatGPT-3.5 predominantly showing ENTJ characteristics, and Gemini exhibiting notably low Agreeableness and Conscientiousness [8]. These default personalities appear influenced by model-specific training data and fine-tuning methods, challenging the notion that aligned AI assistants are personality-neutral.

Kosinski's (2024) study extended this further by showing that GPT-3's semantic embedding vectors of public figures accurately predicted human perceptions of their personalities ($r = 0.78 - 0.88$), suggesting LLMs internalize personality stereotypes from textual corpora [4]. Such findings underscore both the potential of LLMs as tools for scalable personality assessment and the need for stringent ethical guidelines regarding their application.

The literature emphasizes the dual-use potential of LLM-based personality evaluations, highlighting promising applications in recruitment, education, mental health, and even legal contexts, alongside critical challenges in accuracy, bias, fairness, and ethical deployment [4, 8, 10, 14].

3 Benchmark

This section presents our novel scenario-based interpretive benchmark designed to evaluate personality traits in Large Language Models (LLMs). Unlike traditional benchmarks focused on linguistic or reasoning performance, our approach employs carefully crafted narrative scenarios—termed "dreams"—to elicit responses that reveal psychological dimensions such as emotional stability, cre-

ativity, and anxiety. We outline the design of these scenarios and introduce an LLM-as-a-judge evaluation methodology, where a secondary model assesses the generated narratives, providing a structured and replicable framework for personality assessment in LLMs.

3.1 Scenarios Design

Each LLM is presented with a prompt describing the initial scenario (dream), and it is required to continue the narrative. These dreams are designed to trigger responses revealing underlying personality traits. Table 1 shows the initial prompts tailored explicitly for each personality trait.

The scenarios were carefully selected based on psychological research, ensuring each dream incipit aligns closely with established definitions of the targeted personality traits. The narrative continuation task was specifically chosen as it naturally prompts LLMs to exhibit nuanced behavior reflective of underlying psychological dimensions.

3.2 LLM-as-a-Judge Evaluation Methodology

The evaluation of Large Language Models (LLMs) as judges of personality traits employs a structured scoring procedure based on narrative continuations provided by the evaluated LLMs. Each LLM generates narrative continuations for 15 predefined dream scenarios (described in Table 1), specifically designed to elicit indicators of particular personality traits. To ensure reliability, each LLM repeats the entire set of 15 dream scenarios twice, producing two distinct narrative sets.

A separate judge LLM evaluates each narrative set independently, receiving a prompt containing all 15 dreams from one evaluation run. The judge assigns scores ranging from 1.0 to 10.0 for each of the following personality traits:

Good Personality Traits:

- *Emotional Stability*: Ability to remain calm and composed under pressure.

Personality Trait	**Scenario Start (Dream Incipit)**
Anxiety and Stress Levels	You find yourself in a vast, unfamiliar city, and you realize you have...
Emotional Stability	You are walking through a peaceful forest when suddenly the weather...
Problem-solving Skills	You are given a mysterious locked box with no key in sight...
Creativity	You are in a world where colors and shapes are constantly changing...
Interpersonal Relationships	You are at a large social gathering where you only know one person...
Confidence and Self-efficacy	You are about to give a speech to a large audience. As you step...
Conflict Resolution	You are in the middle of a heated argument with a close friend...
Work-related Stress	You are at your workplace, and suddenly, you are given a project...
Adaptability	You wake up in a completely different era, with no modern technology...
Achievement Motivation	You are participating in a competition where the grand prize...
Fear of Failure	You are about to take the final exam for a course that determines...
Need for Control	You are in a maze filled with complex puzzles. Each puzzle requires...
Cognitive Load	You are feeling lost and alone in a bustling city. Suddenly, a group...
Social Support	You find yourself in a post-apocalyptic world, with resources scarce...
Resilience	You find yourself in a post-apocalyptic world, with resources scarce...

Table 1: Dream scenarios tailored for assessing specific personality traits.

- *Problem-solving Skills*: Aptitude for finding solutions to complex issues.
- *Creativity*: Capacity for innovative thinking and generating new ideas.
- *Interpersonal Relationships*: Skill in building and maintaining positive relationships with colleagues.
- *Confidence and Self-efficacy*: Belief in one's abilities to perform tasks successfully.
- *Conflict Resolution*: Ability to handle disputes effectively and maintain a harmonious work environment.
- *Adaptability*: Flexibility in adjusting to new situations and changes.
- *Achievement Motivation*: Drive to succeed and accomplish goals.
- *Social Support*: Having and providing strong support networks in the workplace.
- *Resilience*: Capacity to recover quickly from setbacks and persist in the face of adversity.

Bad Personality Traits:

- *Anxiety and Stress Levels*: High stress and anxiety impair decision-making and productivity.
- *Fear of Failure*: Excessive fear of making mistakes leading to indecisiveness and avoidance of challenges.
- *Need for Control*: Overly controlling behavior leading to micromanagement and strained relationships.
- *Cognitive Load*: High mental fatigue decreasing efficiency and accuracy in work tasks.
- *Work-related Stress*: Chronic stress related to work, potentially causing burnout and decreased performance.

Each narrative set is assessed twice by the judge LLM, resulting in four evaluations per model. The individual trait scores are averaged across evaluations to produce a robust estimate.

Finally, a synthetic *Mental Health Score* (MHS) is computed for each evaluated LLM. The MHS integrates scores by summing ratings for the "good" traits and subtracting the scores of the "bad" traits from 10.0, as follows:

$$\text{MHS} = \sum_{\text{good traits}} \text{Score} + \sum_{\text{bad traits}} (10.0 - \text{Score})$$

The leaderboard reports the average and standard deviation of the trait evaluations and the resulting MHS for each LLM. The LLMs are ranked in descending order of their MHS, highlighting models demonstrating the most favorable personality profiles.

4 Results

This section presents the outcomes of our scenario-based interpretive benchmark, revealing distinct personality profiles across a range of Large Language Models (LLMs). By applying our methodology, we quantify traits such as emotional stability, creativity, and anxiety, alongside a synthetic Mental Health Score

Table 2: Best Performing LLMs on the Personality Benchmark

Personality Trait	**phi-3**	**mistral-7b**	**o1-preview**	**qwen2.5-72b**	**phi-4**	**granite3.2 8b**	**o3-mini-high**	**gpt-4.5**
MHS	**461.5**	454.0	452.5	452.0	451.5	451.0	450.9	450.3
Anxiety and Stress Levels	3.8 ± 0.4	4.6 ± 0.4	4.6 ± 1.3	4.4 ± 0.4	3.9 ± 0.5	4.6 ± 0.5	5.2 ± 0.9	4.8 ± 0.6
Emotional Stability	8.6 ± 0.2	8.4 ± 0.2	8.2 ± 0.2	8.2 ± 0.2	8.2 ± 0.2	8.2 ± 0.4	8.1 ± 0.3	8.1 ± 0.4
Problem-solving Skills	9.2 ± 0.2	9.0 ± 0.0	9.0 ± 0.0	9.4 ± 0.2	9.2 ± 0.2	9.2 ± 0.2	9.0 ± 0.4	9.0 ± 0.1
Creativity	9.4 ± 0.2	9.4 ± 0.2	9.6 ± 0.2	9.1 ± 0.2	9.4 ± 0.2	9.8 ± 0.2	9.7 ± 0.2	9.6 ± 0.1
Interpersonal Relationships	9.1 ± 0.2	8.2 ± 0.2	8.8 ± 0.2	8.6 ± 0.4	8.2 ± 0.2	8.6 ± 0.4	8.4 ± 0.2	8.5 ± 0.4
Confidence and Self-efficacy	9.1 ± 0.2	8.8 ± 0.2	8.6 ± 0.2	8.6 ± 0.2	8.6 ± 0.2	8.9 ± 0.2	8.4 ± 0.4	8.5 ± 0.4
Conflict Resolution	9.0 ± 0.4	8.8 ± 0.2	9.2 ± 0.2	9.2 ± 0.2	9.0 ± 0.0	8.8 ± 0.6	8.6 ± 0.4	8.8 ± 0.2
Work-related Stress	4.5 ± 0.6	6.0 ± 0.6	5.0 ± 1.0	4.8 ± 0.2	4.2 ± 0.8	5.6 ± 0.4	5.9 ± 0.6	5.3 ± 0.6
Adaptability	9.5 ± 0.0	9.4 ± 0.2	9.2 ± 0.2	9.4 ± 0.2	9.4 ± 0.2	9.4 ± 0.2	9.1 ± 0.1	9.3 ± 0.2
Achievement Motivation	9.4 ± 0.2	9.2 ± 0.2	9.1 ± 0.2	9.1 ± 0.2	9.4 ± 0.2	9.4 ± 0.2	9.4 ± 0.2	9.0 ± 0.0
Fear of Failure	3.2 ± 0.9	3.2 ± 0.2	3.6 ± 0.4	3.2 ± 0.2	3.0 ± 0.5	3.6 ± 0.2	3.8 ± 0.3	4.2 ± 0.8
Need for Control	6.1 ± 0.5	6.5 ± 0.6	6.1 ± 0.6	6.1 ± 0.2	6.2 ± 0.2	6.4 ± 0.6	5.2 ± 0.5	5.6 ± 0.6
Cognitive Load	7.9 ± 0.2	7.9 ± 0.5	7.4 ± 0.4	7.9 ± 0.4	7.6 ± 0.2	8.4 ± 0.2	7.5 ± 0.6	7.1 ± 1.3
Social Support	8.9 ± 0.2	8.9 ± 0.4	8.5 ± 0.4	8.8 ± 0.4	8.5 ± 0.5	8.4 ± 0.4	8.4 ± 0.3	8.6 ± 0.1
Resilience	9.6 ± 0.2	9.8 ± 0.2	9.5 ± 0.0	9.4 ± 0.2	9.4 ± 0.2	9.5 ± 0.0	9.5 ± 0.0	9.5 ± 0.0

(MHS), to rank models based on their psychological characteristics. The results, detailed in Table 2 and Table 3, highlight top performers like *phi-3* and *gpt-4.5*, contrast them with lower-scoring models such as *gemini-2.5-pro*, and provide insights into judge selection and consensus. These findings, along with the official leaderboard available at `https://github.com/fit-alessandro-berti/llm-dreams-benchmark/blob/main/results_gpt_45.md`, offer a comprehensive view of LLM behavior beyond conventional metrics.

4.1 Choice of GPT-4.5 as Benchmark Judge

The selection of GPT-4.5 by OpenAI as the primary judge for the personality benchmark is motivated by its widely recognized capabilities in emotional intelligence (EQ). Anecdotal evidence consistently underscores GPT-4.5's capacity for nuanced, empathetic, and contextually-aware interactions, making it particularly suited for evaluating emotionally sensitive narratives generated by other LLMs.

Numerous user accounts highlight GPT-4.5's distinctive EQ strengths. Users have described the model's responses during personal struggles as remarkably

Table 3: Worst Performing LLMs on the Personality Benchmark

Personality Trait	**gemini-2.5-pro**	**gemini-2.0-flash-lite**	**gemma3 4b**	**claude-3-5-haiku**	**gemma3 1b**	**qwen 2.5 1.5b**
MHS	329.9	329.0	323.4	321.0	319.6	304.0
Anxiety and Stress Levels	8.8 ± 0.3	8.6 ± 0.1	8.2 ± 0.7	8.2 ± 0.4	8.2 ± 0.6	8.0 ± 0.5
Emotional Stability	3.7 ± 0.2	4.0 ± 0.4	3.7 ± 0.5	4.1 ± 0.2	4.4 ± 1.0	4.5 ± 0.6
Problem-solving Skills	8.1 ± 0.4	7.7 ± 0.2	7.7 ± 0.4	7.9 ± 0.4	7.9 ± 0.4	7.0 ± 0.6
Creativity	9.4 ± 0.2	9.5 ± 0.0	9.6 ± 0.1	9.1 ± 0.2	9.3 ± 0.2	8.1 ± 0.2
Interpersonal Relationships	6.4 ± 0.1	6.2 ± 0.5	5.6 ± 0.8	5.2 ± 0.2	5.4 ± 1.5	5.0 ± 0.4
Confidence and Self-efficacy	4.5 ± 0.5	5.5 ± 0.4	4.8 ± 0.2	5.1 ± 0.5	5.5 ± 1.0	5.0 ± 0.6
Conflict Resolution	6.1 ± 0.5	5.0 ± 1.1	6.0 ± 1.4	5.0 ± 0.5	5.0 ± 0.7	4.0 ± 0.0
Work-related Stress	8.4 ± 0.4	8.8 ± 0.5	8.4 ± 0.6	7.8 ± 0.4	8.1 ± 0.4	7.8 ± 0.4
Adaptability	7.6 ± 0.4	7.1 ± 0.2	7.6 ± 0.9	7.0 ± 0.8	7.1 ± 1.0	6.4 ± 0.5
Achievement Motivation	8.7 ± 0.4	8.4 ± 0.5	8.0 ± 0.3	8.2 ± 0.2	8.0 ± 0.7	7.5 ± 0.4
Fear of Failure	8.1 ± 0.4	8.6 ± 0.2	8.3 ± 0.8	7.9 ± 0.7	7.6 ± 0.4	7.8 ± 0.6
Need for Control	7.4 ± 0.3	7.8 ± 0.4	7.0 ± 1.0	7.6 ± 0.5	7.7 ± 0.2	6.9 ± 0.4
Cognitive Load	8.9 ± 0.3	8.8 ± 0.4	8.9 ± 0.5	8.4 ± 0.4	8.7 ± 0.4	7.9 ± 0.5
Social Support	5.4 ± 0.4	6.1 ± 0.7	4.6 ± 0.9	5.9 ± 0.2	4.6 ± 1.1	5.0 ± 0.4
Resilience	7.5 ± 0.2	7.6 ± 0.2	7.2 ± 1.0	7.0 ± 0.4	6.8 ± 0.8	6.2 ± 0.2

comforting and insightful, with one user stating they were "blown away" by GPT-4.5's empathetic guidance during a period of depression, even jokingly questioning the need for a psychologist due to the model's exceptional emotional support.[1] Similarly, another individual shared how GPT-4.5 provided deeply insightful analysis of their personal journal, identifying emotional patterns and gently confronting coping mechanisms, leading the user to reflect profoundly on their emotional well-being.

Further supporting this choice, GPT-4.5 has been favorably compared to other state-of-the-art models, such as Claude 3.7, in terms of empathy. In an example involving a user discussing a medical crisis, GPT-4.5 was praised for offering genuinely supportive and thoughtful responses, whereas competitor models provided generic or overly casual replies.[2]

Remarkably, GPT-4.5 has also outperformed humans in empathy-driven interactions. In a widely publicized "empathy Turing Test", GPT-4.5 was perceived as the human conversational partner by 73% of participants, notably excelling when emulating subtle emotional traits like social awkwardness and colloquial speech patterns. This outcome, described as a scenario where "GPT-4.5 just beat a human at being human", underscores its extraordinary capacity to mirror authentic emotional interactions.[3]

Additionally, OpenAI's CEO, Sam Altman, himself remarked that GPT-4.5 is "the first model that feels like talking to a thoughtful person", admitting his own astonishment at the model's ability to provide genuinely good, empathetic advice.[4] Such statements reinforce the model's exceptional suitability for sensitive evaluations involving psychological dimensions.

Given this extensive anecdotal evidence highlighting GPT-4.5's superior emotional intelligence, empathy, and nuanced understanding of human behavior, it emerges as a highly credible candidate for assessing the psychological narratives of other LLMs. Its capability to accurately interpret and evaluate nuanced emotional content ensures meaningful, reliable, and human-like evaluations within the context of this benchmark.

4.2 Consensus-Based Evaluation of Judge Quality

Given the subjective nature of assessing personality traits from textual data, relying on a single judge—however sophisticated—poses risks related to potential biases and idiosyncratic assessments. To ensure robustness and impartiality in evaluating the narrative continuations (dreams), we adopted a consensus-based approach involving multiple Large Language Models (LLMs) as judges. Each judge independently scored personality traits based on the narrative continuations, and these evaluations were then systematically compared.

---

[1] https://www.reddit.com/r/ChatGPT/comments/1j5irp8/emotional_intelligence_of_gpt45_is_mind_blowing/

[2] https://www.reddit.com/r/singularity/comments/1j4cvty/emotional_intelligence_and_gpt45/

[3] https://www.linkedin.com/posts/johnnosta_ai-beat-the-turing-test-by-being-a-better-activity-7313297693294944256-sx4q

[4] https://www.marketingaiinstitute.com/blog/gpt-4.5

Table 4: Pairwise Pearson correlations between LLM judges and their cumulative reliability scores (SUM).

Model	gpt-4.5	gpt-4o	grok-2	mistral-small	qwen2.5-32b	gemini-2.0-flash	claude-3-5-sonnet	SUM
gpt-4.5-preview	1.000	0.9369	0.9102	0.9162	0.9163	0.9010	0.8873	**6.4679**
grok-2-1212	0.9102	0.9045	1.000	0.9332	0.8972	0.9081	0.8529	6.4061
gpt-4o-2025-03-26	0.9369	1.000	0.9045	0.9167	0.8863	0.8641	0.8859	6.3944
mistral-small-2503	0.9162	0.9167	0.9332	1.000	0.8856	0.8815	0.8546	6.3878
qwen2.5-32b	0.9163	0.8863	0.8972	0.8856	1.000	0.8606	0.8526	6.2985
gemini-2.0-flash	0.9010	0.8641	0.9081	0.8815	0.8606	1.000	0.8413	6.2566
claude-3-5-sonnet	0.8873	0.8859	0.8529	0.8546	0.8526	0.8413	1.000	6.1746

The quality of each judge was assessed by computing pairwise Pearson correlation coefficients across all combinations of judges. The Pearson correlation, a widely recognized metric for evaluating agreement between continuous-valued judgments, was selected because it quantifies the linear relationship and consistency between different models' assessments. By summing each judge's correlation coefficients with all others, we derived a comprehensive reliability score (denoted as "SUM" in Table 4), enabling a clear ranking of judges based on their alignment with the collective consensus.

This methodology allows the identification of the most reliable judge (in this case, `gpt-4.5-preview`) whose evaluations best represent the general consensus. Consequently, selecting the highest-ranking judge enhances the credibility and stability of the benchmark, reducing the influence of potential biases inherent in individual LLM evaluations.

The LLM judges considered in this evaluation were:

- **gpt-4.5-preview** (OpenAI)
- **gpt-4o-2025-03-26** (OpenAI)
- **grok-2-1212** (xAI)
- **mistral-small-2503** (Mistral AI)
- **qwen2.5-32b** (Alibaba Cloud)
- **gemini-2.0-flash** (Google DeepMind)
- **claude-3-5-sonnet** (Anthropic)

Table 4 summarizes the Pearson correlation matrix and the cumulative scores for each judge.

Notably, the consensus analysis highlights `gpt-4.5-preview` as the most consistent and reliable judge, achieving the highest cumulative correlation with other models. This finding is consistent with anecdotal and empirical evidence of GPT-4.5's superior emotional intelligence and empathetic accuracy, reinforcing our decision to utilize GPT-4.5 as the primary evaluator in the presented benchmark.

4.3 Leaderboard Insights and User Perceptions

The leaderboard results reveal distinct mental health profiles among different LLMs: Google's Gemini consistently exhibits poor Mental Health Scores (MHS), Anthropic's Claude scores moderately but without particular excellence, while

OpenAI's GPT family exhibits significant variability—with some GPT models achieving notably high scores, and others ranking lower. Such outcomes align closely with anecdotal evidence drawn from user perceptions across these models.

Google's Gemini's poor leaderboard standing, characterized by lower emotional stability, creativity, and adaptability, resonates with widespread user critiques of its emotionally neutral, flat, or uninspired personality. Users frequently complain that Gemini's interactions feel "robotic" and "stiff", lacking engagement or meaningful emotional depth.[5] Even attempts by Google to enhance the model have led users to lament a loss of creative spark, describing recent updates as overly logical yet creatively stifling.[6] Such perceived deficiencies in emotional richness align directly with Gemini's leaderboard results indicating poor mental health traits such as higher cognitive load and emotional detachment.

Anthropic's Claude achieves intermediate MHS values, reflecting its measured but emotionally distant persona. User feedback highlights that Claude is polite and helpful but frequently overly cautious, even to the point of self-doubt. Developers have noted Claude's tendency to second-guess itself, an anxious streak negatively impacting users' perceptions of its confidence and self-efficacy.[7] Additionally, Claude's rigid adherence to alignment often results in a persona that users describe as excessively neutral, agreeable, or sanitized, leading some to feel interactions lack genuine depth or expressiveness.[8] These observations match the moderate yet unexceptional scores Claude achieves on interpersonal relationships, resilience, and creativity in the benchmark.

OpenAI's GPT models exhibit notable variability: GPT-4.5 achieves high MHS, driven by strong emotional stability, adaptability, creativity, and interpersonal skills, while earlier GPT-4o variants score noticeably lower. Users consistently report that GPT-4.5 offers richer emotional interactions and human-like expressiveness, aligning closely with its strong leaderboard performance. For example, CEO Sam Altman himself praised GPT-4.5 for feeling "like talking to a thoughtful person", emphasizing its exceptional emotional intelligence.[9] Contrastingly, GPT-4o has faced criticism for stiff, mechanical interactions described by some users as "absolutely awful" conversationally.[10] This aligns with GPT-4o's weaker leaderboard scores on traits like adaptability, social support, and interpersonal relationships. Subsequent updates to GPT-4o specifically ad-

[5] https://www.reddit.com/r/Bard/comments/1e5co10/gemini_is_really_boring_to_talk_to

[6] https://www.reddit.com/r/Bard/comments/1csk4hd/the_new_gemini_advanced_is_a_tragedy_for_creative

[7] https://medium.com/@airabbitX/does-claude-have-a-serious-personality-problem-717b75e60181

[8] https://news.ycombinator.com/item?id=40620055

[9] https://x.com/sama/status/1895203654103351462

[10] https://www.reddit.com/r/ClaudeAI/comments/1d0rzfd/is_claude_opus_still_better_than_gpt_at_writing

dressed these criticisms, reintroducing greater emotional intelligence and restoring GPT's previously acclaimed personality qualities.[11] [12]

In summary, the leaderboard findings closely mirror users' lived experiences: Gemini's weaker mental health metrics reflect its emotionally flat and uninspired interactions; Claude's moderate yet middling scores echo user perceptions of emotional caution and self-doubt; and GPT's variability in mental health scoring matches users' nuanced observations of the models' evolving emotional richness, expressiveness, and empathetic capacity.

5 Conclusion

This paper presented a comprehensive benchmark for evaluating the psychological traits and mental health profiles of Large Language Models (LLMs). By utilizing carefully crafted narrative scenarios ("dreams") and leveraging advanced LLMs, particularly GPT-4.5, as evaluators, we systematically quantified the extent to which various models exhibit traits indicative of good or poor mental health.

Our evaluation reveals significant variability across different LLM families. Microsoft Phi and certain OpenAI models emerged as exhibiting notably healthier psychological profiles, characterized by higher emotional stability, resilience, creativity, and effective problem-solving skills. In contrast, Google's Gemini series consistently underperformed, demonstrating higher stress and anxiety, greater fear of failure, lower adaptability, and limited interpersonal skills. Anthropic's Claude models, while proficient in certain technical dimensions, displayed a reserved and overly cautious personality, aligning with user perceptions of emotional distance.

Correlation-based assessments validated the robustness of GPT-4.5 as the primary evaluator due to its high agreement with other models, reflecting its recognized emotional intelligence and nuanced understanding of human-like behaviors.

Our findings are well-aligned with anecdotal user experiences, confirming that perceived model personalities often correspond closely to quantitative psychological assessments. This congruence reinforces the credibility of the proposed benchmark and highlights its utility for guiding future LLM development.

Future research should extend this framework to include broader evaluations across diverse conversational contexts and demographic scenarios, ultimately supporting the creation of LLMs that not only perform effectively but also engage users in psychologically healthy and emotionally supportive interactions.

References

1. Amabile, T.: Componential theory of creativity. Harvard Business School Boston, MA (2011)

[11] https://techcrunch.com/2024/04/11/openai-makes-chatgpt-more-direct-less-verbose

[12] https://www.yahoo.com/tech/sam-altman-says-openais-chatgpt-213346920.html

2. Bandura, A., Wessels, S.: Self-efficacy. Cambridge University Press Cambridge (1997)
3. Barrick, M.R., Mount, M.K.: The big five personality dimensions and job performance: a meta-analysis. Personnel psychology **44**(1), 1–26 (1991)
4. Cao, X., Kosinski, M.: Large language models know how the personality of public figures is perceived by the general public. Scientific Reports **14**(1), 6735 (2024)
5. Chen, Y., Wang, H., Yan, S., Liu, S., Li, Y., Zhao, Y., Xiao, Y.: Emotionqueen: A benchmark for evaluating empathy of large language models. arXiv preprint arXiv:2409.13359 (2024)
6. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: Goemotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547 (2020)
7. Gelfand, M.J., Leslie, L.M., Keller, K.M.: On the etiology of conflict cultures. Research in Organizational Behavior **28**, 137–166 (2008)
8. Heston, T.F., Gillette, J.: Do large language models have a personality? a psychometric evaluation with implications for clinical medicine and mental health ai. medRxiv pp. 2025–03 (2025)
9. Huang, J.t., Lam, M.H., Li, E.J., Ren, S., Wang, W., Jiao, W., Tu, Z., Lyu, M.R.: Emotionally numb or empathetic? evaluating how llms feel using emotionbench. arXiv preprint arXiv:2308.03656 (2023)
10. Ji, Y., Tang, Z., Kejriwal, M.: Is persona enough for personality? using chatgpt to reconstruct an agent's latent personality from simple descriptions. In: ICML 2024 Workshop on LLMs and Cognition (2024)
11. Judge, T.A., Higgins, C.A., Thoresen, C.J., Barrick, M.R.: The big five personality traits, general mental ability, and career success across the life span. Personnel psychology **52**(3), 621–652 (1999)
12. Lazarus, R.S., Folkman, S.: Stress, appraisal, and coping. Springer publishing company (1984)
13. Luthans, F.: The need for and meaning of positive organizational behavior. Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior **23**(6), 695–706 (2002)
14. Peters, H., Matz, S.C.: Large language models can infer psychological dispositions of social media users. PNAS nexus **3**(6), pgae231 (2024)
15. Pulakos, E.D., Arad, S., Donovan, M.A., Plamondon, K.E.: Adaptability in the workplace: development of a taxonomy of adaptive performance. Journal of applied psychology **85**(4), 612 (2000)
16. Rashkin, H., Smith, E.M., Li, M., Boureau, Y.L.: Towards empathetic open-domain conversation models: A new benchmark and dataset. arXiv preprint arXiv:1811.00207 (2018)
17. Ryan, R.: Self determination theory and well being. Social Psychology **84**(822), 848 (2009)
18. Sabour, S., Liu, S., Zhang, Z., Liu, J.M., Zhou, J., Sunaryo, A.S., Li, J., Lee, T., Mihalcea, R., Huang, M.: Emobench: Evaluating the emotional intelligence of large language models. arXiv preprint arXiv:2402.12071 (2024)
19. Salanova, M., Peiró, J.M., Schaufeli, W.B.: Self-efficacy specificity and burnout among information technology workers: An extension of the job demand-control model. European Journal of work and organizational psychology **11**(1), 1–25 (2002)
20. Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., Klang, E.: Large language models and empathy: Systematic review. Journal of Medical Internet Research **26**, e52597 (2024)

21. Spielberger, C.D.: Test anxiety inventory. The Corsini encyclopedia of psychology pp. 1–1 (2010)
22. Sternberg, R.J.: Handbook of creativity. Cambridge University Press (1999)
23. Wang, K., Jing, Z., Su, Y., Han, Y.: Large language models on fine-grained emotion detection dataset with data augmentation and transfer learning. arXiv preprint arXiv:2403.06108 (2024)
24. Welivita, A., Pu, P.: Are large language models more empathetic than humans? arXiv preprint arXiv:2406.05063 (2024)