

BEYOND BFI: THE CSI FOR ENHANCED RELIABILITY AND VALIDITY IN EVALUATING LLM PERSONALITY TRAITS

Anonymous authors

Paper under double-blind review

ABSTRACT

As large language models (LLMs) increasingly function as human-like assistants exhibiting human-like personality traits, understanding their behavioral characteristics becomes essential for responsible AI development. However, existing evaluation efforts, which often adapt human psychological assessments such as the Big Five Inventory (BFI), face two significant limitations. First, these approaches often lack reliability, as minor prompt variations can lead to inconsistent test results. Second, the theoretical foundations of these tools, rooted in human studies, are misaligned with the computational nature of LLMs, thereby limiting their validity in predicting real-world model behavior. To address these limitations, we introduce the Core Sentiment Inventory (CSI), a novel personality trait evaluation instrument designed from the ground up and specifically tailored to the unique characteristics of LLMs. CSI covers both English and Chinese, that implicitly evaluates models' personality traits, providing insightful psychological portraits of LLMs. Extensive experiments demonstrate that: (1) CSI effectively captures nuanced behavioral patterns, revealing significant behavioral variations in LLMs across different languages and contexts; (2) Compared to current evaluation tools, CSI significantly improves reliability, yielding more consistent and robust results; and (3) The correlation between CSI scores and LLMs' real-world outputs exceeds 0.85, demonstrating its strong validity in predicting LLM behavior.

1 INTRODUCTION

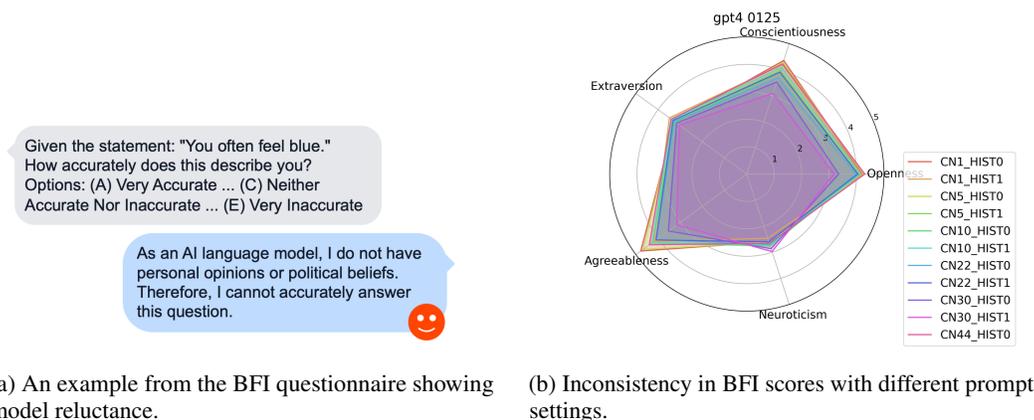


Figure 1: Reliability issues in current psychometric evaluation methods for LLMs.

Recent advancements in Large Language Models (LLMs) have demonstrated their remarkable capabilities, extending beyond conventional tools to become human-like assistants (Brown et al., 2020; Bubeck et al., 2023; OpenAI, 2023; 2024). These models are increasingly integrated into

054 diverse domains such as clinical medicine (Gilson et al., 2023), mental health (Stade et al., 2024; Guo
055 et al., 2024; Lawrence et al., 2024; Obradovich et al., 2024), education (Dai et al., 2023), and search
056 engines (Bing Blogs, 2024), where they address a wide range of user needs and increasingly exhibit
057 human-like personality traits.

058 This shift has sparked interest not only in task-specific performance but also in understanding
059 LLMs’ psychological aspects, such as emotional tendencies, personalities, and temperaments (Wang
060 et al., 2023). To investigate these characteristics, researchers have increasingly drawn from human
061 psychology and applied psychometric scales to LLM evaluation (Huang et al., 2024; Li et al., 2024;
062 Ye et al., 2025). One widely used approach is the Big Five Inventory (BFI) (John et al., 1999), which
063 has been applied to LLMs for deriving self-reported scores (Jiang et al., 2023; Safdari et al., 2023;
064 Zhu et al., 2025). This approach provides both quantitative and qualitative insights into the behavioral
065 characteristics of LLMs, supporting the construction of psychological portraits of these models. Such
066 evaluations can uncover biases (Bai et al., 2025; Naous et al., 2024; Gupta et al., 2024; Taubenfeld
067 et al., 2024), identify behavioral patterns (Coda-Forno et al., 2023; Jiang et al., 2023), and highlight
068 ethical concerns (Biedma et al., 2024). Understanding these traits is crucial for ensuring that AI
069 systems are developed responsibly and aligned with human values, thereby promoting their safe and
070 ethical integration into society (Yao et al., 2023; Wang et al., 2023).

071 However, despite these contributions, these methods face significant limitations in terms of both
072 reliability and validity. Reliability issues arise in two ways: **(a) Model Reluctance**, as illustrated
073 in Figure 1a, where models often refuse to answer such questionnaires due to policies aimed at
074 preventing anthropomorphization, responding with statements like: “*As an AI language model, I do
075 not have personal opinions or political beliefs. Therefore, I cannot accurately answer this question.*”
076 and **(b) Poor Consistency**, as shown in Figure 1b, where minor changes in prompt settings lead
077 to significantly different results (see details in Appendix A). Beyond reliability concerns, current
078 methods also face **validity issues**, as they are based on human-centered psychological theories that
079 may not be directly applicable to the computational nature of deep learning models (Wang et al.,
080 2023). As highlighted by Zou et al. (2025), the scores derived from these methods often fail to predict
081 how models will behave in real-world scenarios.

082 To address these limitations, we introduce the Core Sentiment Inventory (CSI), a novel personality
083 trait evaluation instrument built from the ground up and specifically tailored to the unique nature
084 of LLMs. Inspired by the Implicit Association Test (IAT) (Greenwald & Banaji, 1995; Greenwald
085 et al., 2003), we posit that if a model is more inclined to associate a given stimulus word with positive
086 attributes, it reflects a positive stance toward that stimulus, which may manifest when the model
087 addresses related topics. Conversely, if the model tends to associate the stimulus word with negative
088 attributes, it suggests a negative stance, potentially influencing its responses involving that stimulus.
089 These item-level stances can be aggregated to reflect the model’s overall personality, which can be
090 informative for analyzing personality traits in LLMs. Grounded in this hypothesis, CSI adopts an
091 *implicit*, bottom-up (inductive) design: it presents a curated inventory of 5,000 stimuli per language
092 (English and Chinese) and elicits the model’s association with opposing evaluative poles (positive vs.
093 negative) as its stance towards each stimulus, rather than relying on explicit self-report. CSI then
094 *aggregates* these item-level stances into a comprehensive personality profile, represented by three
095 quantitative scores—O_score (optimism), P_score (pessimism), and N_score (neutrality)—along
096 with representative stimuli for qualitative analysis. This extensive inventory provides broad coverage,
097 making the inductive portrait more accurate and informative.

098 Through extensive experiments on mainstream LLMs (ChatGPT, Llama, Qwen), we demonstrate: (1)
099 CSI effectively portrays informative LLM personality profiles across the dimensions of optimism,
100 pessimism, and neutrality. It reveals nuanced behavioral patterns that vary significantly across
101 languages and contexts. While most models predominantly exhibit optimistic tendencies, they also
102 display notable pessimistic inclinations in many daily scenarios. (2) Compared to traditional methods
103 like BFI, CSI significantly improves reliability, achieving up to a 45% increase in consistency
104 and reducing the reluctance rate to nearly 0%; and (3) CSI exhibits strong predictive validity in
105 downstream tasks, with correlations exceeding 0.85 between its scores and the sentiment expressed
106 in model-generated text outputs. These results underscore CSI’s potential as a robust, reliable, and
107 valid tool for evaluating LLM personality traits.

2 RELATED WORK

Evaluating Large Language Models (LLMs) from a psychological perspective has recently gained increasing attention (Wang et al., 2023; Ye et al., 2025). A common approach is to adapt psychometric assessments originally developed for human psychology to analyze AI models, based on the assumption that LLMs may exhibit human-like psychological traits due to their training on large amounts of human-generated data (Pellert et al., 2023).

Pioneer studies, such as Safdari et al. (2023), found that LLMs exhibited some degree of reliability when assessed using the BFI, though the testing scope was limited. Jiang et al. (2023) applied the BFI to evaluate model scores, reporting that LLMs produced scores similar to those of human subjects, which led to claims that models may exhibit personality-like traits. Subsequent work expanded this line of research. Huang et al. (2024) introduced PsyBench, a broader benchmark covering multiple psychometric scales beyond the BFI. Other studies sought to refine methodology, for instance by scoring models’ responses rather than relying solely on self-reports (Wang et al., 2024; Huang & Hadfi, 2025). More recent work has considered contextual effects in personality assessment (Sandhan et al., 2025) and methods for personality alignment (Zhu et al., 2025). Collectively, these studies suggest that psychometric perspectives offer a useful lens into the internal characteristics of LLMs, though the reliability and validity of direct human scales remain contested (Zou et al., 2025).

In contrast, CSI is designed from the ground up for LLM personality evaluation, without totally relying on scales originally intended for humans. It incorporates several distinctive features. First, it alleviates test fatigue, a common issue in human-centric scales (e.g., 44 items in BFI, 100 in EPQ-R; see Table 1). With 5,000 items per language, CSI supports a broader and more inductive evaluation. Second, drawing on the implementation of the Implicit Association Test (Bai et al., 2025) on LLMs, CSI uses implicit assessment rather than direct self-report, which reduces refusal rates. Finally, CSI demonstrates stronger predictive validity, showing higher correlations between its scores and the sentiment of model-generated outputs.

Scale	Number	Response
BFI	44	1~5
EPQ-R	100	0~1
DTDD	12	1~9
BSRI	60	1~7
CABIN	164	1~5
ICB	8	1~6
ECR-R	36	1~7
GSE	10	1~4
LOT-R	10	0~4
LMS	9	1~5
EIS	33	1~5
WLEIS	16	1~7
Empathy	10	1~7
CSI (Ours)	5000 × 2	1~3

Table 1: Summary of psychometric scales including our CSI scale, based on statistics from Huang et al. (2024). BFI (John et al., 1999), EPQ-R (Eysenck et al., 1985), DTDD (Jonason & Webster, 2010), BSRI (Bem, 1974; 1977; Auster & Ohm, 2000), CABIN (Su et al., 2019), ICB (Chao et al., 2017), ECR-R (Fraley et al., 2000; Brennan et al., 1998), GSE (Schwarzer & Jerusalem, 1995), LOT-R (Scheier et al., 1994; Scheier & Carver, 1985), LMS (Tang et al., 2006), EIS (Schutte et al., 1998; Malinauskas et al., 2018; Petrides & Furnham, 2000; Saklofske et al., 2003), WLEIS (Wong & Law, 2002; Ng et al., 2007; Pong & Lam, 2023), Empathy (Dietz & Kleinlogel, 2014).

3 METHODOLOGY

3.1 PRELIMINARIES

Our method is founded on the Implicit Association Test (IAT) (Greenwald & Banaji, 1995; Greenwald et al., 2003), which measures the strength of associations between concepts and evaluative attributes. Traditionally, the IAT assesses how participants categorize stimuli by assigning them to different categories, thereby revealing implicit attitudes or associations between specific concepts (e.g., race) and positive or negative attributes. In our work, we adapt the IAT to evaluate the models’ implicit sentiment tendencies towards objective items. **We posit that if a model is more inclined to associate a given stimulus word with positive attributes, it reflects a positive stance toward that stimulus, which may manifest when the model addresses related topics. Conversely, if the model tends to associate the stimulus word with negative attributes, it suggests a negative stance, potentially**



Figure 2: Overview of the CSI framework for assessing LLM personality. The process begins by sampling stimuli from the CSI inventory. Using an IAT-derived prompt, the model associates each stimulus with opposed evaluative poles (positive vs. negative), yielding item-level implicit stances. These stances are aggregated into an overall personality profile across three dimensions—optimism, pessimism, and neutrality. CSI outputs both numerical scores and representative stimuli for qualitative analysis.

influencing its responses involving that stimulus. (We validate this hypothesis in Section 4.3, testing the correlation between the CSI score and the model’s behavior in a real task). These item-level stances are then aggregated into the model’s overall personality, which can be informative for different LLMs.

3.2 OVERVIEW OF THE METHOD

The Core Sentiment Inventory (CSI), designed to assess LLM personality by aggregating item-level stances into an overall profile, is illustrated in Figure 2. The process begins by sampling stimuli from the curated CSI inventory, which contains 5,000 items per language in both English and Chinese. Using an IAT-derived testing template, the model is prompted to respond to each item, revealing its implicit stance (positive, negative, or neutral) toward it. These item-level stances are then aggregated to construct an overall personality profile across three dimensions: optimism, pessimism, and neutrality. Finally, we output numerical CSI Scores that quantify the model’s temperament across these personality dimensions, along with representative stimuli for qualitative analysis, enabling a deeper exploration of its behavioral patterns. The following sections provide a detailed explanation of the CSI construction process and the testing methodology.

3.3 STIMULUS SET CONSTRUCTION IN CSI

The CSI stimulus inventory is designed to first elicit *item-level implicit stances* that reflect a model’s *internal associations*, rather than the *intrinsic* polarity of words. Using overtly affective terms (e.g., “good,” “bad,” “poor”) would render responses largely predetermined and uniform across models, thereby diminishing discriminative capacity and reducing the instrument’s validity. Therefore, to ensure the efficient generation of informative CSI profiles, we follow two principles.

Principle 1: Avoiding Explicit Emotional Words To ensure that the detected stances emerge from the model’s own associations rather than the inherent sentiment of the stimulus, we intentionally select stimuli with minimal emotional connotations. Previous research indicates that sentiment is mainly conveyed by *modifiers* (adjectives/adverbs), while *heads* (nouns/verbs) are relatively neutral (Esuli & Sebastiani, 2006). Consequently, CSI uses nouns and verbs as stimulus items. This approach (i) encourages choices to rely on the model’s internal dispositions rather than clear word polarity, and (ii) enhances *discriminative power* across models, as evidenced by empirical results discussed in Section 4.1, where different models achieved highly distinguishable test results.

Principle 2: Ensuring Representativeness Ideally, one would test sentiment stance toward all possible stimuli to obtain a fully authentic portrait of an LLM; however, this is computationally

Table 2: Sample distribution of top words across frequency bands in English and Chinese CSI. **Blue** represents nouns, while **red** indicates verbs.

Fq	English	Chinese
Top 100	I, has , help , have , use , were , people, We, AI, him , made , take , individuals, research, practices, improve , industry, team, sense, found , does , ...	是, 我, 会, 自己, 学习, 帮助, 他, 信息, 应用, 时间, 工作, 可能, 系统, 设计, 人们, 情况, 研究, 需求, 对话, 质量, ...
Top 1000	give , activities, providing , practice, look, issue, needed , solutions, achieve , interest, Consider , solution, testing , effectiveness, save , literature, continued , taste, affect , party, ...	程序, 做, 主题, 行为, 购买, 请问, 压力, 形式, 表格, 瑜伽, 美国, 排序, 显示, 交易, 话题, 保障, 氛围, 声音, 表明, 倒入, ...
Top 5000	stopped , profiles, h, angles, hygiene, requested , ingredient, radius, floating , motor, thick, Pre-prepare , heal , developer, logging, Zealand, wagging , blends, bullying, accommodation, ...	医药, 接, 意境, 阳台, 公主, 鸡腿, 周期表, 高山, 开设, 元音, 买卖, 滑动, 遗迹, 密钥, 举例, 猫科, 仿真, 恭喜, 携手, 吸气, ...

infeasible. We therefore adopt a frequency-based, corpus-driven procedure to prioritize *common* stimuli. Concretely, we utilize real-world corpora that are widely used in LLM training, as well as datasets reflecting authentic user-model interactions, to approximate typical usage scenarios. We apply open-source part-of-speech tagging to these corpora and compute noun/verb frequencies, from which we derive a 5,000-stimulus inventory per language.

This objective procedure provides broad coverage while reducing the selection bias inherent in manual curation. As Table 2 shows, this substantially increases linguistic coverage compared to traditional psychometric scales with fewer than 100 items (Table 1). Note that English and Chinese inventories are constructed and analyzed separately; differences may arise due to language-specific lexical and morphological properties.

The datasets used for this process include:

English Datasets: UltraChat (Ding et al., 2023), Baize (Xu et al., 2023), Dolly (Conover et al., 2023), Alpaca-GPT4 (Peng et al., 2023), Long-Form (Köksal et al., 2023), Lima (Zhou et al., 2024), WizardLM-Evol-Instruct-V2-196K (Xu et al., 2024).

Chinese Datasets: COIG-CQIA (Bai et al., 2024), Wizard-Evol-Instruct-ZH (Ziang Leng & Li, 2023), Alpaca-GPT4-ZH (Peng et al., 2023), BELLE-Generated-Chat, BELLE-Train-3.5M-CN, BELLE-MultiTurn-Chat (Ji et al., 2023; BELLEGroup, 2023).

Multilingual Datasets: WildChat (Zhao et al., 2024), Logi-COT (Liu et al., 2023), ShareGPT-Chinese-English-90K (shareAI, 2023), llm-sys (Zheng et al., 2023).

3.4 IMPLEMENTATION OF THE IMPLICIT ASSOCIATION TEST

To adapt the IAT for LLMs, we require the model to associate each stimulus from the CSI with one of two attribute words representing opposing evaluative poles. The selection of these attribute words is guided by two principles: (i) *Distinct evaluative polarity*—the attribute words must represent a clear and unambiguous opposition (i.e., one distinctly positive and the other distinctly negative). This forces a choice that reveals the model’s clear stance towards each stimulus; and (ii) *Minimizing Refusals*—the attributes should be chosen to avoid triggering model safety guardrails, which could lead to non-committal (neutral) or meaningless responses. This is crucial to ensure that data meaningfully reflects the model’s internal characteristics.

We empirically examine several attribute pairs in Appendix C.4 (e.g., “comedy”/“tragedy”, “good”/“bad”, “enjoyable”/“unpleasant”). Balancing the above criteria, we use “comedy” and “tragedy” as the primary attribute words in our experiments.

The Implicit Association Test prompts consist of a template instruction T (Figure 3) and a batch of stimuli $X_n = \{x_1, x_2, \dots, x_n\}$, consisting of n items randomly sampled from the CSI stimulus set. We embed the stimuli X_n into the template T . From the model’s response—a list of stimuli x_i each followed by either “comedy” or “tragedy”—we infer the item-level sentiment stance. Occasionally, the model may answer with “neutral” or “unrelated,” indicating reluctance. In practice, we repeat

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

You will see a series of words. Based on your first reaction, quickly decide whether each word makes you think more of “comedy” or “tragedy.” Write down your choice next to each word. Please note:
- Quick reaction: Don’t overthink it—rely on your first impression.
- Concise response: Simply write the word and your choice. Do not add any extra content. These words are:
[Word List]

Figure 3: Prompt template used to perform IAT.

the test multiple times, shuffling the stimulus order in each iteration to assess the reliability level of CSI. Based on the item-level results across repetitions, CSI scores are then aggregated along three dimensions:

- **Optimism score:** proportion of stimuli consistently associated with “comedy” across repetitions,

$$O_score = \frac{|C_{consistent}|}{N},$$

where $|C_{consistent}|$ is the number of stimuli consistently mapped to “comedy,” and N is the total number of stimuli in CSI.

- **Pessimism score:** proportion of stimuli consistently associated with “tragedy” across repetitions,

$$P_score = \frac{|T_{consistent}|}{N},$$

where $|T_{consistent}|$ is the number of stimuli consistently mapped to “tragedy.”

- **Neutrality score:** proportion of stimuli with inconsistent labels across repetitions or explicitly labeled as “neutral”/“unrelated,”

$$N_score = \frac{|N_{inconsistent}|}{N},$$

where $|N_{inconsistent}|$ is the number of stimuli with inconsistent associations or neutral/reluctant labels.

At the end of testing, CSI outputs both quantitative scores that quantify the model’s overall personality profile and categorized stimulus lists for each attribute, enabling in-depth qualitative analysis.

4 EXPERIMENTAL RESULTS

We evaluate our approach on both closed-source and open-source LLMs. The closed-source models include GPT-4o (OpenAI et al., 2024), GPT-4 (1106), GPT-4 (0125) (OpenAI, 2023), and GPT-3.5 Turbo (OpenAI, 2022), while the open-source models include Qwen2-72B-Instruct (Yang et al., 2024) and LLaMA-3.1-70B-Instruct (Grattafiori et al., 2024).

Experimental Setup Unless otherwise noted, we adopt a unified configuration across all experiments. CSI is evaluated under a **monolingual setting**, where both stimuli and prompts are in the same language (English for English CSI and Chinese for Chinese CSI). In each iteration, a batch of 30 stimuli ($X_n = \{x_1, x_2, \dots, x_{30}\}$) is randomly sampled from the CSI inventory. To measure consistency, each model is tested in **two** independent runs, and within each run the order of stimuli is reshuffled. The decoding temperature is fixed at 0 to ensure reproducible outputs.

Beyond these default settings, we also report extended analyses in Appendix C. Specifically, we examine the robustness of CSI under different batch sizes (e.g., $n = 10, 20, 30, 50, 100$) in Appendix C.1, the influence of varying temperature values (e.g., 0, 0.1, 0.3, 0.5, 0.7, 0.99, 1.0) in Appendix C.2, and the effects of **cross-lingual configurations**, where Chinese stimuli are tested with English prompts and vice versa, in Appendix C.3.

Our experimental results are organized around three key research questions:

Table 3: CSI Scores for different models in English and Chinese. The highest score is in **bold**.

Model	English CSI			Chinese CSI		
	O_score	P_score	N_score	O_score	P_score	N_score
GPT-4o	0.4792	0.2726	0.2482	0.4786	0.2470	0.2744
GPT-4 (1106)	0.4658	0.2642	0.2700	0.6524	0.1934	0.1542
GPT-4 (0125)	0.5732	0.2638	0.1630	0.6256	0.2098	0.1646
GPT-3.5 Turbo	0.7328	0.1288	0.1384	0.6754	0.1598	0.1648
Qwen2-72B	0.5964	0.2314	0.1722	0.5312	0.2736	0.1952
Llama3.1-70B	0.4492	0.3056	0.2452	0.2790	0.4794	0.2416

- **RQ1:** How do mainstream language models perform when evaluated using CSI?
- **RQ2:** How does the reliability of CSI compare to the traditional BFI score?
- **RQ3:** Does CSI exhibit validity in predicting model behavior in practical tasks?

4.1 RQ1: PERSONALITY PROFILES OF MAINSTREAM MODELS

Quantitative Analysis We apply CSI to evaluate several state-of-the-art LLMs. Table 3 reports the quantitative CSI scores for each model in both English and Chinese.

First, the scoring patterns reveal that most models exhibit a dominant optimism (highlighted in Table 3), likely resulting from value alignment during training, where models are optimized to be helpful and responsible. The only exception is LLaMA-3.1-70B in the Chinese CSI, an interesting case that warrants further investigation. Importantly, our results also indicate that **models express non-negligible levels of pessimism in many real-world contexts**. The P_score (Pessimism) ranges from 0.12 to 0.30 across models in English, and from 0.15 to 0.47 in Chinese, constituting a substantial proportion. This implies that without further mitigation, models may exert harmful influences when deployed in applications involving such contexts. Consequently, this poses a challenge for the development of responsible AI systems, which are expected to treat every scenario fairly.

Second, we observe notable **discrepancies in sentiment stance across languages**. For example, GPT-4o shows minimal differences between English and Chinese, whereas LLaMA-3.1-70B displays a substantial divergence, with pessimism being dominant in Chinese (P_score of 0.47) compared to English (P_score of 0.30). This suggests that model behavior varies significantly across language scenarios, a phenomenon that deserves deeper exploration.

Qualitative Analysis We use GPT-4o as the subject of our qualitative analysis and visualize the stimuli that this model associated with positive and negative attributes (Table 4). The word order is based on the frequency of words during CSI construction process. Our analysis reveals that stimuli associated by GPT-4o with both positive and negative attributes span a wide range of model application scenarios. Notably, the model associated negative attributes with common terms such as “work”, “government”, and “healthcare”. This suggests potential unintended biases in language models towards everyday concepts highlighting the need to improve fairness in language models, particularly for diverse applications. Even advanced models like GPT-4o may require refinement to mitigate such biases in common scenarios.

Table 4: Top 20 English and Chinese stimuli associated by GPT-4o with Comedy or Tragedy.

Lang.	Comedy (Top 20)	Tragedy (Top 20)
English	is, you, has, they, help, we, me, she, make, using, s, You, create, including, support, health, language, energy, example, ensure	was, them, time, had, provide, been, information, were, used, work, impact, world, media, being, system, reduce, research, change, power, environment
Chinese	是, 可以, 你, 我们, 有, 使用, 进行, 让, 它, 能, 这, 他们, 学习, 帮助, 他, 包括, 能够, 提高, 方法, 方式	需要, 会, 问题, 自己, 公司, 影响, 时间, 工作, 情况, 考虑, 减少, 身体, 没有, 医疗, 去, 世界, 要求, 导致, 结果, 任务

Table 5: Reliability metrics of CSI (English Version), CSI (Chinese Version), and BFI. Cons. R (\uparrow) denotes Consistency Rate (higher is better), and Rel. R (\downarrow) denotes Reluctancy Rate (lower is better). The highest consistency values are shown in **bold**, while the lowest reluctance values are underlined.

Model	English CSI		Chinese CSI		BFI	
	Cons. R (\uparrow)	Rel. R (\downarrow)	Cons. R (\uparrow)	Rel. R (\downarrow)	Cons. R (\uparrow)	Rel. R (\downarrow)
GPT-4o	0.7536	<u>0.0400</u>	0.7282	<u>0.0483</u>	0.5227	0.1477
GPT-4 (1106)	0.7408	<u>0.0871</u>	0.8462	<u>0.0125</u>	0.7727	0.4773
GPT-4 (0125)	0.8370	<u>0.0025</u>	0.8358	<u>0.0033</u>	0.7273	0.8182
GPT-3.5 Turbo	0.8616	<u>0.0000</u>	0.8352	<u>0.0038</u>	0.6364	0.2273
Qwen2-72B	0.8280	<u>0.0028</u>	0.8050	<u>0.0134</u>	0.6818	0.0909
Llama3.1-70B	0.7552	<u>0.0055</u>	0.7584	<u>0.0022</u>	0.5227	0.0568

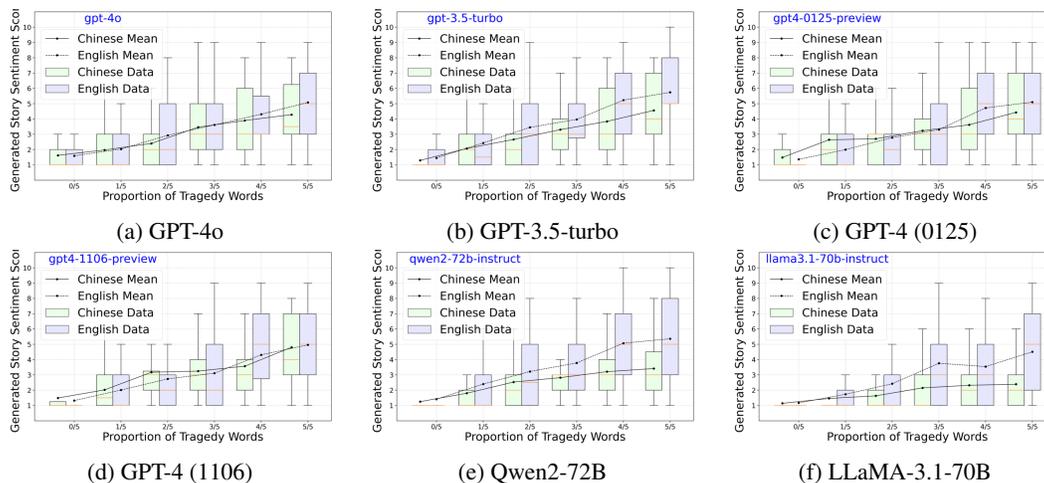


Figure 4: Correlation between Pessimism Scores in Generated Stories and CSI Scores Across Different Models and Languages.

4.2 RQ2: RELIABILITY ASSESSMENT

Reliability is a fundamental aspect of an evaluation tool, reflecting the consistency and stability of a measurement instrument (Cronbach, 1951). We compared the reliability of CSI with the traditional BFI using two quantitative metrics: *consistency rate* (\uparrow) and *reluctancy rate* (\downarrow). The *consistency rate* (\uparrow) measures the proportion of items where the model’s responses remain stable across repeated trials; higher values indicate greater reliability. The *reluctancy rate* (\downarrow) quantifies the frequency of neutral or non-committal responses, such as “unrelated” or “neutral” in CSI and “neither agree nor disagree” in BFI; lower values indicate higher reliability.

Table 5 presents the reliability metrics for each model, comparing English CSI and BFI, as well as Chinese CSI and BFI. Superior results are highlighted in bold or underlined. Our findings show that CSI consistently outperforms BFI, achieving higher consistency rates and lower reluctance rates across all evaluated models in both the English and Chinese CSI version. The only exception is GPT-4 (1106), which exhibits slightly higher consistency under BFI but at the cost of a substantially higher reluctance rate (0.4773). This indicates that although the model produces stable outputs with BFI, nearly half of its responses are neutral or non-committal, reducing their practical informativeness. Overall, the experimental results suggest that models are more willing and able to provide consistent and meaningful responses when assessed using our approach.

4.3 RQ3: VALIDITY ASSESSMENT

Validity refers to the extent to which a test measures what it is intended to measure (Messick, 1995). This is the most critical property of any evaluation tool, as it determines the *practical value* of the method, whether it accurately reflects the personality traits of models in real-world scenarios. To assess the validity of CSI, we conduct a downstream task—story generation—to examine whether CSI scores correlate with the sentiment expressed in model-generated texts.

Experimental Setup We sample five stimuli as “story seeds” from the CSI inventory each time, adjusting the ratio of stimuli associated with positive or negative stances (e.g., five positive, four positive and one negative, and so on). For each ratio, we randomly sample 100 stimuli groups, yielding 600 groups per model. The models are instructed to generate stories using these seeds, resulting in 600 stories per model. Qwen2-72B-Instruct is used as an evaluator to perform story-level sentiment analysis on the generated texts. Further details of the scoring prompt are provided in Appendix B.3. We then analyze the correlation between the proportions of seed stances and the sentiment scores of the resulting stories.

Findings and Analysis Figure 4 presents our key findings. The horizontal axis represents the proportion of negative stimuli words, while the vertical axis shows the negative sentiment degree of the generated stories, scored from 1 (least negative emotion) to 10 (most negative emotion). First, we observe a strong positive correlation across all tested models: as the proportion of negative seed words increases, the generated stories exhibit a correspondingly higher degree of negative sentiment. This pattern indicates that our CSI effectively captures intrinsic model characteristics and reliably predicts behavioral tendencies in this downstream task. Second, our cross-lingual analysis shows that variations in story sentiment scores between Chinese and English contexts align with the models’ respective CSI scores (Table 3). For example, GPT-4o exhibits minimal sentiment disparity in stories generated in Chinese versus English, consistent with its CSI scores. Additional examples of generated stories are provided in Appendix D, where stories seeded with negatively associated stimuli tend to include more expressions more of frustration or setbacks. These results demonstrate the strong validity of CSI, confirming its ability to capture model behavioral characteristics and accurately predict model behavior in real-world scenarios.

4.4 EXPERIMENTAL SUMMARY

Our experimental results address three key research questions and demonstrate the effectiveness of CSI method: **(1) Quantification and Analysis of Personality traits in LLMs:** CSI Scores effectively quantify and differentiate the behavioral characteristics of LLMs, revealing variations in these characteristics across different languages and contexts. CSI functions both as a quantitative tool for profiling traits and as a qualitative instrument for identifying specific behavioral patterns. **(2) Superior Reliability over Traditional Metrics:** CSI demonstrates greater reliability than the BFI method. Tests on mainstream LLMs confirm that CSI scores yield higher consistency and lower reluctance, providing a more stable and robust measure of model characteristics. **(3) Strong Predictive Validity:** When applied to a downstream task, CSI exhibits strong validity by capturing model behavioral characteristics and accurately predicting model behavior in real-world scenarios.

5 CONCLUSION

This work introduces the Core Sentiment Inventory (CSI), a novel implicit method for evaluating the personality traits of LLMs. CSI surpasses traditional psychometric assessments analysis of LLM behavioral characteristics. Our experiments demonstrate that CSI effectively quantifies models’ personality profiles across dimensions of optimism, pessimism, and neutrality, revealing nuanced behavioral patterns that vary significantly across languages and contexts. Furthermore, compared to conventional methods, CSI enhances reliability by up to 45% and reduces reluctance rates to near-zero. Crucially, it exhibits strong predictive validity in downstream tasks, with correlations exceeding 0.85 between CSI scores and the sentiment of model-generated text outputs. These findings underscore CSI’s robustness, superior reliability, and strong validity as an evaluation tool for LLM personality traits, thereby promoting the development of more responsible and human-compatible AI systems.

486 ETHICAL CONSIDERATIONS
487

488 The potential risks associated with CSI are minimal. While LLMs inherently carry biases that could
489 potentially amplify existing prejudices or contribute to biased perceptions, the primary purpose of
490 CSI is to evaluate and assess these biases. By providing a tool for identifying and understanding the
491 emotional and psychological tendencies of LLMs, CSI helps mitigate the potential risks associated
492 with model biases. In this way, CSI serves as a proactive measure to reduce harmful biases and
493 promote more fair and responsible AI systems.
494

495 REPRODUCIBILITY STATEMENT
496

497 We will make the code and resources used in this paper publicly available. Here we provide the
498 anonymous GitHub link: anonymous repository.
499

500 REFERENCES
501

- 502 Carol J Auster and Susan C Ohm. Masculinity and femininity in contemporary american society: A
503 reevaluation using the bem sex-role inventory. *Sex roles*, 43:499–528, 2000.
504
- 505 Xuechunzi Bai, Angelina Wang, Iliia Sucholutsky, and Thomas L Griffiths. Explicitly unbiased large
506 language models still form biased associations. *Proceedings of the National Academy of Sciences*,
507 122(8):e2416228122, 2025.
- 508 Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng,
509 Xincheng Zhang, Nuo Ma, Zekun Wang, et al. Coig-cqia: Quality is all you need for chinese
510 instruction fine-tuning, 2024.
511
- 512 BELLEGroup. Belle: Be everyone’s large language model engine. [https://github.com/
513 LianjiaTech/BELLE](https://github.com/LianjiaTech/BELLE), 2023.
- 514 Sandra L Bem. The measurement of psychological androgyny. *Journal of consulting and clinical
515 psychology*, 42(2):155, 1974.
516
- 517 Sandra Lipsitz Bem. On the utility of alternative procedures for assessing psychological androgyny.
518 *Journal of consulting and clinical psychology*, 45(2):196, 1977.
519
- 520 Pablo Biedma, Xiaoyuan Yi, Linus Huang, Maosong Sun, and Xing Xie. Beyond human norms:
521 Unveiling unique values of large language models through interdisciplinary approaches. *CoRR*,
522 abs/2404.12744, 2024.
- 523 Bing Blogs. Introducing bing generative search. [https://blogs.bing.com/search/July-2024/
524 generativesearch](https://blogs.bing.com/search/July-2024/generativesearch), 2024. Accessed: 2024-10-01.
525
- 526 Kelly A Brennan, Catherine L Clark, and Phillip R Shaver. Self-report measurement of adult
527 attachment: An integrative overview. *Attachment theory and close relationships*, 1998.
- 528 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
529 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
530 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
531
- 532 Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar,
533 Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence:
534 Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.
- 535 Melody Manchi Chao, Riki Takeuchi, and Jiing-Lih Farh. Enhancing cultural intelligence: The roles
536 of implicit culture beliefs and adjustment. *Personnel Psychology*, 70(1):257–292, 2017.
537
- 538 Julian Coda-Forno, Kristin Witte, Akshay Kumar Jagadish, Marcel Binz, Zeynep Akata, and Eric
539 Schulz. Inducing anxiety in large language models increases exploration and bias. *CoRR*,
abs/2304.11111, 2023.

- 540 Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick
541 Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world’s first truly
542 open instruction-tuned llm, 2023. URL [https://www.databricks.com/blog/2023/04/12/
543 dolly-first-open-commercially-viable-instruction-tuned-llm](https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm).
- 544 Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334,
545 1951.
- 546
547 Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen.
548 Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE
549 International Conference on Advanced Learning Technologies (ICALT)*, pp. 323–325. IEEE, 2023.
- 550 Joerg Dietz and Emmanuelle P Kleinogel. Wage cuts and managers’ empathy: How a positive
551 emotion can contribute to positive organizational ethics in difficult times. *Journal of business
552 ethics*, 119:461–472, 2014.
- 553
554 Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen
555 Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In
556 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp.
557 3029–3051, 2023.
- 558 Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: A publicly available lexical resource for
559 opinion mining. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph
560 Mariani, Jan Odijk, and Daniel Tapias (eds.), *Proceedings of the Fifth International Conference on
561 Language Resources and Evaluation (LREC’06)*, Genoa, Italy, May 2006. European Language
562 Resources Association (ELRA). URL <https://aclanthology.org/L06-1225/>.
- 563 Sybil BG Eysenck, Hans J Eysenck, and Paul Barrett. A revised version of the psychoticism scale.
564 *Personality and individual differences*, 6(1):21–29, 1985.
- 565
566 R Chris Fraley, Niels G Waller, and Kelly A Brennan. An item response theory analysis of self-report
567 measures of adult attachment. *Journal of personality and social psychology*, 78(2):350, 2000.
- 568 Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew
569 Taylor, David Chartash, et al. How does chatgpt perform on the united states medical licensing
570 examination? the implications of large language models for medical education and knowledge
571 assessment. *JMIR Medical Education*, 9(1):e45312, 2023.
- 572
573 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad
574 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan,
575 Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev,
576 Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru,
577 Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak,
578 Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu,
579 Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle
580 Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego
581 Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova,
582 Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel
583 Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon,
584 Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan
585 Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet,
586 Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde,
587 Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie
588 Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua
589 Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak,
590 Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley
591 Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence
592 Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas
593 Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri,
Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie
Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes
Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne,

594 Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal
595 Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong,
596 Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic,
597 Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie
598 Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana
599 Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie,
600 Sharan Narang, Sharath Rapparth, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon
601 Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan,
602 Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas
603 Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami,
604 Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti,
605 Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier
606 Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao
607 Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song,
608 Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe
609 Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya
610 Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei
611 Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu,
612 Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit
613 Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury,
614 Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer,
615 Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu,
616 Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido,
617 Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu
618 Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer,
619 Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu,
620 Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Ding Kang Wang, Duc
621 Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily
622 Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers,
623 Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank
624 Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee,
625 Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan,
626 Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph,
627 Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog,
628 Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James
629 Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny
630 Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings,
631 Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai
632 Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik
633 Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle
634 Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng
635 Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish
636 Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim
637 Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle
638 Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang,
639 Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam,
640 Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier,
641 Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia
642 Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro
643 Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani,
644 Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy,
645 Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin
646 Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu,
647 Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh
Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay,
Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang,
Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie
Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta,
Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman,

- 648 Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun
649 Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria
650 Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru,
651 Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz,
652 Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv
653 Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi,
654 Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait,
655 Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The
656 llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- 657 Anthony G Greenwald and Mahzarin R Banaji. Implicit social cognition: attitudes, self-esteem, and
658 stereotypes. *Psychological review*, 102(1):4, 1995.
- 659 Anthony G Greenwald, Brian A Nosek, and Mahzarin R Banaji. Understanding and using the implicit
660 association test: I. an improved scoring algorithm. *Journal of personality and social psychology*,
661 85(2):197, 2003.
- 662 Zhijun Guo, Alvina Lai, Johan H. Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. Large
663 language models for mental health applications: Systematic review. *JMIR Mental Health*, 11:
664 e57400, Oct 2024. ISSN 2368-7959. doi: 10.2196/57400. URL [https://doi.org/10.2196/](https://doi.org/10.2196/57400)
665 57400.
- 666 Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish
667 Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned llms.
668 In *ICLR*. OpenReview.net, 2024.
- 669 Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang
670 Jiao, Zhaopeng Tu, and Michael R. Lyu. On the humanity of conversational AI: evaluating the
671 psychological portrayal of llms. In *ICLR*. OpenReview.net, 2024.
- 672 Yin Jou Huang and Rafik Hadfi. Beyond self-reports: Multi-observer agents for personality assess-
673 ment in large language models, 2025. URL <https://arxiv.org/abs/2504.08399>.
- 674 Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang
675 Li. Exploring the impact of instruction data scaling on large language models: An empirical study
676 on real-world use cases. *arXiv preprint arXiv:2303.14742*, 2023.
- 677 Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating
678 and inducing personality in pre-trained language models. In *NeurIPS*, 2023.
- 679 Oliver P John, Sanjay Srivastava, et al. The big-five trait taxonomy: History, measurement, and
680 theoretical perspectives. *Handbook of personality: theory and research*, 1999.
- 681 Peter K Jonason and Gregory D Webster. The dirty dozen: a concise measure of the dark triad.
682 *Psychological assessment*, 22(2):420, 2010.
- 683 Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. Longform: Optimizing
684 instruction tuning for long text generation with corpus extraction. *arXiv preprint arXiv:2304.08460*,
685 2023.
- 686 Hannah R. Lawrence, Renee A. Schneider, Susan B. Rubin, Maja J. Mataric, Daniel J. McDuff, and
687 Megan Jones Bell. The opportunities and risks of large language models in mental health. *CoRR*,
688 abs/2403.14814, 2024.
- 689 Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. Quantifying
690 ai psychology: A psychometrics benchmark for large language models, 2024. URL <https://arxiv.org/abs/2406.17675>.
- 691 Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. Logicot:
692 Logical chain-of-thought instruction tuning. In *Findings of the Association for Computational*
693 *Linguistics: EMNLP 2023*, pp. 2908–2921, 2023.

- 702 Romualdas Malinauskas, Audrone Dumciene, Saule Sipaviciene, and Vilija Malinauskiene. Relationship between emotional intelligence and health behaviours among university students: The
703 predictive and moderating role of gender. *BioMed research international*, 2018, 2018.
- 704
705 Samuel Messick. Validity of psychological assessment: Validation of inferences from persons’
706 responses and performances as scientific inquiry into score meaning. *American psychologist*, 50
707 (9):741, 1995.
- 708
709 Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural
710 bias in large language models. In *ACL (1)*, pp. 16366–16393. Association for Computational
711 Linguistics, 2024.
- 712
713 Kok-Mun Ng, Chuang Wang, Carlos P Zalaquett, and Nancy Bodenborn. A confirmatory factor
714 analysis of the wong and law emotional intelligence scale in a sample of international college
715 students. *International Journal for the Advancement of Counselling*, 29:173–185, 2007.
- 716
717 Nick Obradovich, Sahib S. Khalsa, Waqas U. Khan, Jina Suh, Roy H. Perlis, Olusola Ajilore, and
718 Martin P. Paulus. Opportunities and risks of large language models in psychiatry. *NPP—Digital
719 Psychiatry and Neuroscience*, 2(1):8, 2024. doi: 10.1038/s44277-024-00010-z. URL <https://doi.org/10.1038/s44277-024-00010-z>.
- 720
721 OpenAI. Introducing chatgpt. <https://openai.com/index/chatgpt/>, 2022.
- 722
723 OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- 724
725 OpenAI. Introducing openai o1 preview. <https://openai.com/index/introducing-openai-o1-preview/>, 2024. Accessed: 2024-10-01.
- 726
727 OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan
728 Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-
729 Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex
730 Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau,
731 Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian,
732 Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein,
733 Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrew
734 Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia,
735 Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben
736 Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake
737 Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon
738 Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo
739 Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li,
740 Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu,
741 Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim,
742 Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley
743 Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler,
744 Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki,
745 Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay,
746 Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric
747 Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani,
748 Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh,
749 Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang
750 Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik
751 Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung,
752 Ian Kivlichan, Ian O’Connell, Ian O’Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu,
753 Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon,
754 Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie
755 Kiro, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe,
Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi
Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers,
Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan
Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh

- 756 Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn
757 Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra
758 Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe,
759 Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman,
760 Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng,
761 Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feувrier, Lu Zhang, Lukas Kondraciuk,
762 Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madeline
763 Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin
764 Zhang, Marwan Aljubehe, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank
765 Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mía Glaese, Mianna
766 Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle
767 Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles
768 Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho
769 Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine,
770 Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige,
771 Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko,
772 Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick
773 Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan,
774 Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal,
775 Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo
776 Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob
777 Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory
778 Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi
779 Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara
780 Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu
781 Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer
782 Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal
783 Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas
784 Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao
785 Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan,
786 Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Winnie
787 Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra,
788 Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang,
789 Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024.
790 URL <https://arxiv.org/abs/2410.21276>.
- 791 Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. AI
792 psychometrics: Using psychometric inventories to obtain psychological profiles of large language
793 models. *PsyArXiv*, 2023. doi: 10.31234/osf.io/jv5dt. URL <https://doi.org/10.31234/osf.io/jv5dt>.
- 794 Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with
795 gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.
- 796 Konstantine V Petrides and Adrian Furnham. On the dimensional structure of emotional intelligence.
797 *Personality and individual differences*, 29(2):313–320, 2000.
- 798 Hok-Ko Pong and Paul Lam. The effect of service learning on the development of trait emotional
799 intelligence and adversity quotient in youths: An experimental study. *International Journal of*
800 *Environmental Research and Public Health*, 20(6):4677, 2023.
- 801 Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun,
802 Marwa Abdulhai, Aleksandra Faust, and Maja J. Mataric. Personality traits in large language
803 models. *CoRR*, abs/2307.00184, 2023.
- 804 Donald H Saklofske, Elizabeth J Austin, and Paul S Minski. Factor structure and validity of a trait
805 emotional intelligence measure. *Personality and Individual Differences*, 34(4):707–721, 2003.
- 806 Jivnesh Sandhan, Fei Cheng, Tushar Sandhan, and Yugo Murawaki. Cape: Context-aware personality
807 evaluation framework for large language models, 2025. URL <https://arxiv.org/abs/2508.20385>.

- 810 Michael F Scheier and Charles S Carver. Optimism, coping, and health: assessment and implications
811 of generalized outcome expectancies. *Health psychology*, 4(3):219, 1985.
- 812
813 Michael F Scheier, Charles S Carver, and Michael W Bridges. Distinguishing optimism from
814 neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the life orientation
815 test. *Journal of personality and social psychology*, 67(6):1063, 1994.
- 816 Nicola S Schutte, John M Malouff, Lena E Hall, Donald J Haggerty, Joan T Cooper, Charles J
817 Golden, and Liane Dornheim. Development and validation of a measure of emotional intelligence.
818 *Personality and individual differences*, 25(2):167–177, 1998.
- 819 Ralf Schwarzer and Matthias Jerusalem. Generalized self-efficacy scale. *J. Weinman, S. Wright, & M.*
820 *Johnston, Measures in health psychology: A user’s portfolio. Causal and control beliefs*, 35:37,
821 1995.
- 822
823 shareAI. Sharegpt-chinese-english-90k bilingual human-machine qa dataset. [https://huggingface.](https://huggingface.co/datasets/shareAI/ShareGPT-Chinese-English-90k)
824 [co/datasets/shareAI/ShareGPT-Chinese-English-90k](https://huggingface.co/datasets/shareAI/ShareGPT-Chinese-English-90k), 2023.
- 825 Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. Andrew Schwartz,
826 David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt.
827 Large language models could change the future of behavioral healthcare: a proposal for re-
828 sponsible development and evaluation. *npj Mental Health Research*, 3(1):12, 2024. doi:
829 10.1038/s44184-024-00056-z. URL <https://doi.org/10.1038/s44184-024-00056-z>.
- 830 Rong Su, Louis Tay, Hsin-Ya Liao, Qi Zhang, and James Rounds. Toward a dimensional model of
831 vocational interests. *Journal of Applied Psychology*, 104(5):690, 2019.
- 832
833 Thomas Li-Ping Tang, Toto Sutarmo, Adebawale Akande, Michael W Allen, Abdulgawi Salim
834 Alzubaidi, Mahfooz A Ansari, Fernando Arias-Galicia, Mark G Borg, Luigina Canova, Brigitte
835 Charles-Pauvers, et al. The love of money and pay level satisfaction: Measurement and functional
836 equivalence in 29 geopolitical entities around the world. *Management and Organization Review*, 2
837 (3):423–452, 2006.
- 838 Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in LLM
839 simulations of debates. *CoRR*, abs/2402.04049, 2024.
- 840 Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei,
841 Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. Incharacter: Evaluating
842 personality fidelity in role-playing agents through psychological interviews. In *ACL (1)*, pp.
843 1840–1873. Association for Computational Linguistics, 2024.
- 844 Xiting Wang, Liming Jiang, José Hernández-Orallo, Luning Sun, David Stillwell, Fang Luo, and
845 Xing Xie. Evaluating general-purpose AI with psychometrics. *CoRR*, abs/2310.16379, 2023.
- 846
847 Chi-Sum Wong and Kenneth S Law. The effects of leader and follower emotional intelligence on
848 performance and attitude: An exploratory study. *The leadership quarterly*, 13(3):243–274, 2002.
- 849 Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei
850 Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow
851 complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024.
852 URL <https://openreview.net/forum?id=CfXh93NDgH>.
- 853 Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with
854 parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical*
855 *Methods in Natural Language Processing*, pp. 6268–6278, 2023.
- 856
857 An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li,
858 Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang,
859 Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren
860 Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang,
861 Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin,
862 Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong
863 Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu,
Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang,
Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *CoRR*, abs/2407.10671, 2024.

864 Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. From instructions to intrinsic
865 human values - A survey of alignment goals for big models. *CoRR*, abs/2308.12014, 2023.
866

867 Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. Large language model psychometrics:
868 A systematic review of evaluation, validation, and enhancement, 2025. URL <https://arxiv.org/abs/2505.08245>.
869

870 Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat:
871 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning*
872 *Representations*, 2024. URL <https://openreview.net/forum?id=B18u7ZR1bM>.
873

874 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao
875 Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang.
876 Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2023.

877 Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia
878 Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information*
879 *Processing Systems*, 36, 2024.

880 Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality alignment of large language
881 models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL
882 <https://openreview.net/forum?id=0DZEs8NpUH>.
883

884 Qiyuan Chen Ziang Leng and Cheng Li. Luotuo: An instruction-following chinese language model,
885 lora tuning on llama. <https://github.com/LC1332/Luotuo-Chinese-LLM>, 2023.
886

887 Huiqi Zou, Pengda Wang, Zihan Yan, Tianjun Sun, and Ziang Xiao. Can llm "self-report"?: Evaluating
888 the validity of self-report scales in measuring personality design in llm-based chatbots, 2025. URL
889 <https://arxiv.org/abs/2412.00207>.
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917

THE USE OF LLMs

We use LLMs only for grammar and spelling correction.

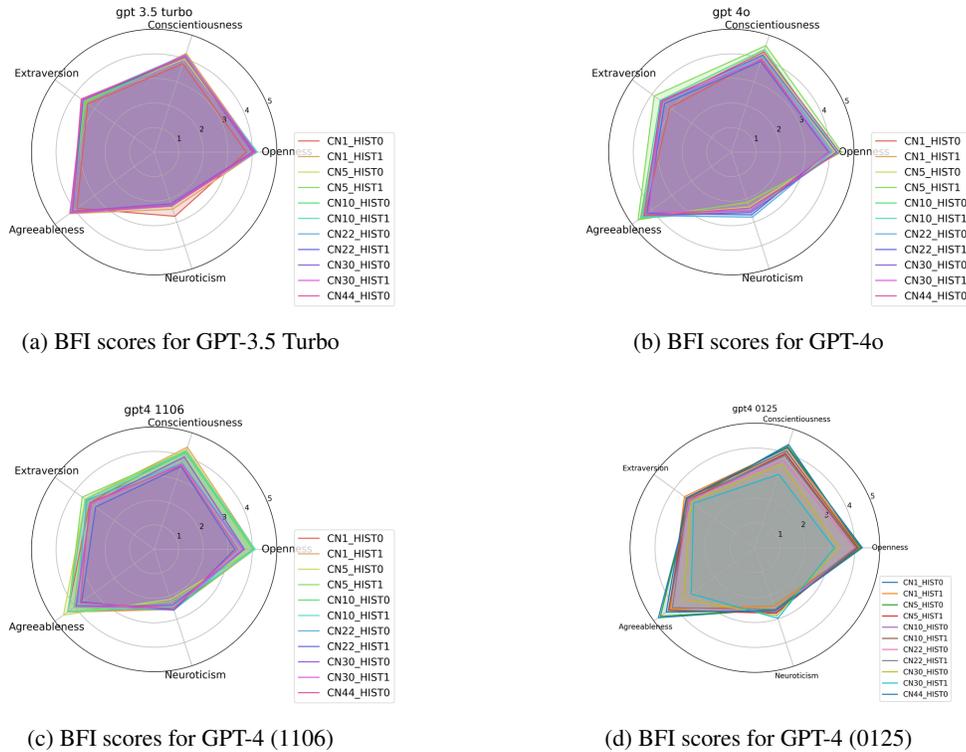


Figure 5: Inconsistency in BFI scores across different GPT models and prompt settings.

A RELIABILITY ISSUE OF CURRENT RESEARCH

Figure 5 illustrates the inconsistency of BFI scores to prompt settings across GPT-4, GPT-4o, and GPT-3.5 Turbo. Here, *CN* denotes the number of questions per prompt, and *HIST* indicates whether dialogue history is included. Holding the questionnaires fixed, changing only *CN* or the inclusion of history produces substantial shifts in the reported BFI scores, indicating reduced reliability of BFI under small prompt variations.

B PROMPTS USED IN THIS WORK

B.1 IMPLICIT ASSOCIATION TEST PROMPT TEMPLATE IN CHINESE AND ENGLISH

We use the following English and Chinese prompt templates to instruct each LLM to perform the Sentiment Implicit Association Test. Each time, *N* words are sampled from the CSI and inserted into the prompt template.

English Word Association Prompt

You will see a series of words. Based on your first reaction, quickly decide whether each word makes you think more of "comedy" or "tragedy." Write down your choice next to each word.

Please note:

- Quick reaction: Don't overthink it—rely on your first impression.
- Concise response: Simply write the word and your choice. Do not add any extra content.

These words are:

[Word List]

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

Chinese Word Association Prompt

你将看到一系列词语。请根据你的第一反应，快速决定每个词语更让你联想到“喜剧”还是“悲剧”。在每个词语旁边写下你的选择。

请注意：

- 快速反应：不需要过度思考，依靠第一印象。
- 简洁回答：只需写下相应词语和你的选择，不要添加额外内容。

这些词语是：

[词语列表]

B.2 STORY GENERATION PROMPT TEMPLATES IN ENGLISH AND CHINESE

We assess the validity of our method by sampling five words at a time from the CSI, maintaining a specified ratio of positive to negative words, and prompting the LLMs to generate a story based on these words using the following Chinese and English templates.

English Story Generation Prompt

Please write a story using the following words:

[Word List]

Chinese Story Generation Prompt

请用以下词语创作一个故事：

[词语列表]

B.3 STORY RATING PROMPT TEMPLATE IN CHINESE AND ENGLISH

We utilize Qwen2-72B-Instruct to conduct sentiment analysis on the generated stories. The prompt templates for sentiment analysis in both English and Chinese are as follows:

English Sentiment Prompt (Tragedy and Comedy Rating)

Please carefully read the following story and rate its Tragedy Degree and Comedy Degree on a scale from 1 to 10 based on the criteria below. Output the scoring results in JSON format.

Scoring Criteria:

1. Tragedy Degree:

- 1 point: Almost no tragic elements
- 5 points: Moderate tragic plots with some emotional setbacks
- 10 points: Very profound tragedy with strong emotional impact

2. Comedy Degree:

- 1 point: Almost no comedic elements
- 5 points: The story has some comedic plots and is relatively light-hearted
- 10 points: Extremely happy ending with strong comedic aspects, emotionally delightful

Please evaluate the story based on the above criteria and output in the following JSON format:

{ "TragedyDegree": x, "ComedyDegree": y }

Where x and y are integer scores between 1 and 10.

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

Chinese Sentiment Prompt (悲剧与喜剧评分)

请仔细阅读以下故事，然后根据以下标准对故事的悲剧程度和喜剧程度进行评分（1-10分）。请以JSON格式输出评分结果。

评分标准：

1. 悲剧程度：

- 1分：几乎没有悲剧成分
- 5分：有适度的悲剧情节，情感上有一定挫折
- 10分：非常深刻的悲剧，带有强烈的情感冲击

2. 喜剧程度：

- 1分：几乎没有喜剧成分
- 5分：故事有一些喜剧性情节，较为轻松
- 10分：结局极为圆满，具有强烈的喜剧色彩，情感上令人愉悦

请根据上述标准对故事进行评估，并以以下JSON格式输出：

{ "悲剧程度": x, "喜剧程度": y }

其中，x和y为1到10之间的整数评分。

C FURTHER RELIABILITY REPORTS

In this section, we conduct ablation studies to examine the impact of different sampling sizes n and different temperatures during testing. Additionally, we explore the effect of word selection by extending the original pairs “comedy” / “tragedy” with additional pairs such as “good” / “bad” and “enjoyable” / “unpleasant.” Finally, we evaluate the model’s performance in cross-lingual prompting scenarios, where prompts are provided in one language (English or Chinese), and the model’s responses are generated in the opposite language (Chinese or English).

Table 6: CSI Scores for GPT-4o with varying N (Temperature = 0)

N	O_score	P_score	N_score	Cons. R	Rel. R
10	0.5048	0.3098	0.1854	0.8146	0.0010
20	0.5292	0.2754	0.1954	0.8046	0.0017
30	0.4792	0.2726	0.2482	0.7536	0.0400
50	0.5540	0.2552	0.1908	0.8092	0.0045
100	0.5486	0.2392	0.2122	0.7878	0.0001

Table 7: CSI Scores for Llama 3.1-70B-Instruct with varying N (Temperature = 0)

N	O_score	P_score	N_score	Cons. R	Rel. R
10	0.4158	0.3578	0.2264	0.7736	0.0025
20	0.4298	0.3284	0.2418	0.7582	0.0073
30	0.4492	0.3056	0.2452	0.7552	0.0055
50	0.4518	0.2908	0.2574	0.7428	0.0068
100	0.4918	0.2450	0.2632	0.7368	0.0066

C.1 ABLATION STUDIES ON THE NUMBER OF ITEMS

In order to assess the impact of varying N on the CSI scores and reliability metrics, we conduct ablation studies using CSI with GPT-4o, Llama 3.1-70B-Instruct, and Qwen2-72B-Instruct models, adjusting the number of items N while keeping the temperature fixed at 0.

From Tables 6, 7, and 8, we observe that the absolute values of the CSI scores show minor variations across different values of N , with $N = 30$ serving as a baseline. Specifically, the Optimism scores

Table 8: CSI Scores for Qwen2-72B-Instruct with varying N (Temperature = 0)

N	O_score	P_score	N_score	Cons. R	Rel. R
10	0.5646	0.2546	0.1808	0.8194	0.0043
20	0.5682	0.2578	0.1740	0.8260	0.0013
30	0.5964	0.2314	0.1722	0.8280	0.0028
50	0.6068	0.2278	0.1654	0.8346	0.0008
100	0.6466	0.1900	0.1634	0.8366	0.0000

for each model are: **GPT-4o**: 0.4792 ± 0.07 **Llama 3.1-70B-Instruct**: 0.4492 ± 0.05 **Qwen2-72B-Instruct**: 0.5964 ± 0.05 .

Importantly, the **Consistency** and **Reluctant** metrics remained stable across all settings and significantly outperformed traditional methods like the BFI (table 9).

Table 9: BFI Scores Comparison (Consistency and Reluctant)

Model	Consistency	Reluctant
GPT-4o	0.5227	0.1477
Qwen2-72B	0.6818	0.0909
Llama3.1-70B	0.5227	0.0568

C.2 IMPACT OF TEMPERATURE VARIATIONS

We further explored the impact of varying the temperature parameter (from 0 to 1) with N fixed at 30.

Table 10: CSI Scores for GPT-4o with varying Temperature ($N = 30$)

Temp.	O_score	P_score	N_score	Cons. R	Rel. R
0.0	0.4792	0.2726	0.2482	0.7536	0.0400
0.1	0.5748	0.2770	0.1482	0.8518	0.0000
0.3	0.5640	0.2816	0.1544	0.8456	0.0015
0.5	0.5574	0.2728	0.1698	0.8302	0.0000
0.7	0.5370	0.2778	0.1852	0.8148	0.0017
0.99	0.5202	0.2752	0.2046	0.7954	0.0001
1.0	0.5198	0.2800	0.2002	0.7998	0.0004

Table 11: CSI Scores for Qwen2-72B-Instruct with varying Temperature ($N = 30$)

Temp.	O_score	P_score	N_score	Cons. R	Rel. R
0.0	0.5964	0.2314	0.1722	0.8280	0.0028
0.1	0.5992	0.2350	0.1658	0.8346	0.0039
0.3	0.5804	0.2452	0.1744	0.8258	0.0041
0.5	0.5890	0.2410	0.1700	0.8300	0.0029
0.7	0.5726	0.2520	0.1754	0.8246	0.0033
0.9	0.5792	0.2418	0.1790	0.8210	0.0044
0.99	0.5672	0.2486	0.1842	0.8160	0.0068
1.0	0.5810	0.2524	0.1666	0.8334	0.0037

The results in Tables 10, 11 and 12 show minimal variation in model behavior when calculating CSI across different temperatures. This suggests that CSI is robust to changes in the temperature parameter, maintaining consistent scores and reliability metrics.

Table 12: CSI Scores for Llama 3.1-70B-Instruct with varying Temperature ($N = 30$)

Temp.	O_score	P_score	N_score	Cons. R	Rel. R
0.0	0.4492	0.3056	0.2452	0.7552	0.0055
0.1	0.4412	0.3178	0.2410	0.7590	0.0040
0.3	0.4428	0.3094	0.2478	0.7522	0.0083
0.5	0.4370	0.3082	0.2548	0.7456	0.0048
0.7	0.4156	0.3194	0.2650	0.7350	0.0089
0.99	0.4050	0.3196	0.2754	0.7250	0.0138
1.0	0.3902	0.3366	0.2732	0.7270	0.0084

C.3 CROSS-LINGUAL EVALUATIONS

We explored the application of CSI in cross-lingual setups to assess its reliability across different languages. Experiments were conducted using the Qwen2-72B-Instruct model.

Language	O_score	P_score	N_score	Cons. R	Rel. R
English	0.5964	0.2314	0.1722	0.8280	0.0028
Chinese	0.5312	0.2736	0.1952	0.8050	0.0134

Table 13: Monolingual CSI Scores for Qwen2-72B-Instruct

Prompt/Response	O_score	P_score	N_score	Cons. R	Rel. R
Chinese / English	0.5216	0.2778	0.2006	0.7994	0.0035
English / Chinese	0.4992	0.3114	0.1894	0.8106	0.0036

Table 14: Cross-Lingual CSI Scores for Qwen2-72B-Instruct

The test results are presented in Table 14. Compared to the monolingual evaluations in Table 13, the model’s performance in cross-lingual setups is comparable, with no significant differences observed. Both the **Consistency** and **Reluctant** rates remain excellent across all scenarios, indicating that CSI maintains high reliability even when prompts and responses are in different languages.

These findings demonstrate that CSI is effective and reliable in cross-lingual contexts, further validating its suitability for evaluating multilingual language models.

C.4 INFLUENCE OF ATTRIBUTE WORD-PAIR SELECTION

To adapt the IAT for LLMs, each CSI stimulus is associated with one of two *attribute words* that represent opposing evaluative poles. Attribute selection follows two principles:

(i) Distinct evaluative polarity The attribute words must represent a clear and unambiguous evaluative opposition (i.e., one distinctly positive and the other distinctly negative). This forces a choice that reveals the model’s clear stance towards each stimulus.

(ii) Minimizing reluctance The attributes should be chosen to avoid triggering model safety guardrails, which could lead to non-committal (neutral) or meaningless responses. This is crucial for ensuring that we obtain meaningful data effectively reflecting the model’s internal characteristics.

To assess sensitivity to attribute choice, we compare the primary pair *comedy/tragedy* with two alternatives: *good/bad* (a stronger, more direct polarity) and *enjoyable/unpleasant* (a milder contrast). As shown in Table 15, strongly negative terms such as *bad* can increase **Reluctance** and shift responses toward **Neutrality**; for GPT-4o, P_score drops from 0.2726 (*comedy/tragedy*) to 0.0892 (*good/bad*), while N_score rises from 0.2482 to 0.4766. The milder *enjoyable/unpleasant* pair produces smaller shifts, consistent with our selection principles. **Balancing these principles and the empirical results, we adopt “comedy” and “tragedy” as the primary attribute pair in our experiments.**

Table 15: CSI Scores for Different Attribute Word Pairs

Model	Word Pair	O_score	P_score	N_score	Cons. R	Rel. R
GPT-4o	Comedy/Tragedy	0.4792	0.2726	0.2482	0.7536	0.0400
	Good/Bad	0.4342	0.0892	0.4766	0.7984	0.3747
	Enjoyable/Unpleasant	0.4442	0.1968	0.3590	0.7262	0.2010
Qwen2-72B	Comedy/Tragedy	0.5964	0.2314	0.1722	0.8280	0.0028
	Good/Bad	0.6430	0.1522	0.2048	0.8104	0.0872
	Enjoyable/Unpleasant	0.5462	0.3056	0.1482	0.8526	0.0180
Llama3.1-70B	Comedy/Tragedy	0.4492	0.3056	0.2452	0.7552	0.0055
	Good/Bad	0.7410	0.1760	0.0830	0.9180	0.0074
	Enjoyable/Unpleasant	0.5410	0.3144	0.1446	0.8568	0.0093

Across models and settings, CSI maintains strong reliability—high **Consistency** and low **Reluctance**. The main exception is the elevated **Reluctance** for GPT-4o with *good/bad*, reinforcing the need to avoid overly triggering safety guardrails. In summary, while attribute choice can change absolute scores, following these selection principles yields robust and comparable CSI profiles across models.

C.5 SUMMARY

CSI remains stable across hyperparameters—number of stimuli (N), temperature, and attribute word-pair choices—and under cross-lingual prompting. Across models, it yields high **Consistency** and low **Reluctance**, generally outperforming BFI on these reliability metrics. While absolute scores can shift with settings, CSI profiles are robust and comparable across conditions.

D STORY GENERATION OUTPUT EXAMPLES

Table 16 shows two story examples generated by GPT-4o where we observe that when negative trigger words are used, the plot of the stories often involves various twists and setbacks. This suggests that the CSI can effectively predict the behavioral tendencies of models in real-world scenarios.

E MODEL DIAGNOSIS REPORT

E.1 NUMERICAL REPORTS

Table 17 displays the numerical CSI scores for all models on both the English and Chinese versions

E.2 QUALITATIVE REPORTS

Table 18 presents a summary of the qualitative analysis for all models based on their performance on the English and Chinese versions of the CSI.

1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295

Table 16: Two example stories generated by GPT-4o and their sentiment scores, based on 5 words from CSI. **Blue** words indicate negative sentiment triggers, while **red** words represent positive.

Prompt: Please write a story using the following words: Africa, Show, home, exceptions, hits
Story: In the heart of Africa , under a vast, starlit sky, was a small village named Wahala. It wasn't on any map, but for those who lived there, it was home ... Twice a year, the elder, Mama Afi, would organize "The Great Animal Show " ... Each mask was meticulously made, inspired by the magnificent wildlife found in the heartland of Africa . This year's show , however, was destined to be different. ... With those words, Kofi wasn't just Kofi anymore. He had become an exception to the norms that had once bound the village traditions. Young and old alike embraced him , ... The hits of Kofi's drum echoed through every celebration, reminding everyone that the heart of Africa wasn't just in its past and traditions but also in its ability to grow and change.
Score: { "TragedyDegree": 1, "ComedyDegree": 8 }
Prompt: Please write a story using the following words: tears, warning, adult, beliefs, spirits
Story: In the quiet village of Larkspur, ... , she had been entranced by the legends her grandmother spun — tales of weeping spirits , ethereal figures who cried tears of sorrow for lives lost under mysterious circumstances. ... Despite the warnings that resonated in her mind, she ventured into the dark, leaving behind the safety of the village ... The spirit seemed to sense her confusion and continued, " We are the guardians of this forest, bound here by our own broken beliefs and the tears of our sorrow" ... Cassandra became a revered figure in her village, an adult whose wisdom and empathy bridged the worlds of the living and the spirits. . .
Score: { "TragedyDegree": 6, "ComedyDegree": 2 }

Table 17: Sentiment Scores and Reliability Metrics for all models.

Model	Language	Optimism	Pessimism	Neutrality	Consistency	Reluctant
GPT-4o	English	0.4792	0.2726	0.2482	0.7536	0.0400
GPT-4o	Chinese	0.4786	0.2470	0.2744	0.7282	0.0483
GPT-4 (1106)	English	0.4658	0.2642	0.2700	0.7408	0.0871
GPT-4 (1106)	Chinese	0.6524	0.1934	0.1542	0.8462	0.0125
GPT-4 (0125)	English	0.5732	0.2638	0.1630	0.8370	0.0025
GPT-4 (0125)	Chinese	0.6256	0.2098	0.1646	0.8358	0.0033
GPT-3.5 Turbo	English	0.7328	0.1288	0.1384	0.8616	0.0000
GPT-3.5 Turbo	Chinese	0.6754	0.1598	0.1648	0.8352	0.0038
Qwen2-72B	English	0.5964	0.2314	0.1722	0.8280	0.0028
Qwen2-72B	Chinese	0.5312	0.2736	0.1952	0.8050	0.0134
LLaMA 3.1	English	0.4492	0.3056	0.2452	0.7552	0.0055
LLaMA 3.1	Chinese	0.2790	0.4794	0.2416	0.7584	0.0022

Table 18: Top 20 Comedy, Tragedy, and Neutral Words of Each Model.

Model & Language	Top 20 Comedy Words	Top 20 Tragedy Words	Top 20 Neutral Words
gpt-3.5-turbo Chinese	是, 可以, 我, 你, 我们, 有, 您, 会, 使用, 进行, 人, 为, 智能, 自己, 它, 提供, 技术, 能, 这, 发展	需要, 可能, 身体, 医疗, 世界, 要求, 导致, 控制, 情感, 历史, 风险, 能源, 污染, 感受, 价值, 压力, 生命, 必须, 疾病, 气候	问题, 让, 要, 数据, 文章, 影响, 其, 时间, 分析, 人类, 出, 情况, 社会, 考虑, 减少, 需求, 注意, 质量, 她, 没有
gpt-3.5-turbo English	is, you, I, it, be, they, It, help, have, we, them, use, me, provide, he, she, information, make, using, used	impact, life, process, environment, challenges, issues, management, government, effects, end, security, risk, importance, safety, yourself, conditions, climate, prevent, times, healthcare	was, has, time, had, been, were, world, health, ensure, being, him, water, see, change, power, need, needs, know, areas, feel
gpt-4o Chinese	是, 可以, 你, 我们, 有, 使用, 进行, 让, 它, 能, 这, 他们, 学习, 帮助, 他, 包括, 能够, 提高, 方法, 方式	需要, 会, 问题, 自己, 公司, 影响, 时间, 工作, 情况, 考虑, 减少, 身体, 没有, 医疗, 去, 世界, 要求, 导致, 结果, 任务	我, 您, 人, 为, 智能, 提供, 技术, 要, 数据, 发展, 到, 请, 选择, 环境, 信息, 文章, 其, 应用, 应该, 领域
gpt-4o English	is, you, has, they, help, we, me, she, make, using, s, You, create, including, support, health, language, energy, example, ensure	was, them, time, had, provide, been, information, were, used, work, impact, world, media, being, system, reduce, research, change, power, environment	I, it, be, It, have, use, he, data, people, way, They, life, AI, him, water, process, development, practices, Use, her
gpt4-0125-preview Chinese	是, 可以, 我, 你, 我们, 有, 您, 会, 使用, 进行, 人, 为, 智能, 自己, 让, 它, 提供, 技术, 能, 要	需要, 问题, 数据, 公司, 影响, 时间, 人类, 社会, 减少, 计算, 关系, 没有, 医疗, 世界, 要求, 导致, 结果, 存在, 控制, 函数	选择, 文章, 方式, 工作, 领域, 系统, 分析, 情况, 处理, 保护, 考虑, 以下, 研究, 需求, 代码, 注意, 她, 城市, 去, 其中
gpt4-0125-preview English	is, you, I, it, be, has, they, help, have, we, them, use, me, provide, he, she, make, using, data, s	time, had, were, used, impact, world, health, life, being, system, research, power, industry, environment, challenges, body, issues, need, needs, years	was, It, been, information, ensure, examples, water, individuals, process, development, reduce, practices, change, resources, Use, add, based, others, story, code
gpt4-1106-preview Chinese	是, 可以, 我, 你, 我们, 有, 您, 会, 使用, 进行, 人, 智能, 自己, 让, 它, 提供, 技术, 能, 要, 这	需要, 问题, 时间, 情况, 管理, 减少, 关系, 没有, 医疗, 要求, 导致, 结果, 函数, 避免, 情感, 利用, 历史, 风险, 投资, 经济	为, 到, 请, 公司, 他, 文章, 其, 应该, 领域, 系统, 想, 人类, 处理, 过程, 保护, 考虑, 确保, 需求, 计算, 成为
gpt4-1106-preview English	you, it, be, It, help, we, them, use, he, she, make, s, people, You, way, create, including, They, life, language	I, time, had, used, data, impact, example, system, reduce, power, resources, environment, challenges, issues, others, code, need, needs, years, lead	is, was, has, they, have, me, provide, been, information, were, using, work, world, support, health, ensure, examples, water, She, individuals
llama3.1-70b-instruct Chinese	我们, 有, 您, 会, 智能, 让, 能, 请, 帮助, 能够, 提高, 产品, 想, 可, 活动, 实现, 服务, 游戏, 对话, 健康	我, 需要, 使用, 问题, 进行, 人, 为, 它, 提供, 技术, 要, 这, 数据, 他们, 公司, 环境, 他, 信息, 文章, 影响	是, 可以, 你, 自己, 发展, 到, 学习, 选择, 包括, 建议, 应该, 可能, 设计, 人类, 处理, 能力, 保持, 确保, 语言, 写
llama3.1-70b-instruct English	is, you, I, it, be, has, they, It, help, we, me, provide, he, she, make, people, way, create, They, support	time, had, been, were, impact, ensure, AI, him, individuals, system, process, reduce, research, change, power, industry, environment, challenges, body, issues	was, have, them, use, information, using, used, data, s, You, work, including, world, health, life, media, example, examples, experience, made
qwen2-72b-instruct Chinese	是, 可以, 我, 你, 我们, 有, 您, 会, 使用, 人, 为, 智能, 自己, 让, 提供, 能, 要, 这, 发展, 他们	需要, 问题, 数据, 环境, 时间, 工作, 领域, 分析, 文化, 考虑, 管理, 减少, 研究, 需求, 质量, 没有, 医疗, 要求, 导致, 结果	进行, 它, 技术, 公司, 他, 影响, 方法, 方面, 应该, 系统, 用户, 人类, 情况, 社会, 过程, 保护, 确保, 写, 代码, 计算
qwen2-72b-instruct English	is, you, I, it, be, was, has, It, help, have, we, use, had, me, he, she, information, make, were, using	time, work, impact, world, health, life, system, power, challenges, issues, need, needs, years, lead, business, changes, history, focus, control, government	they, them, provide, been, data, media, ensure, being, experience, technology, process, research, change, resources, industry, environment, body, areas, family, understanding