A User-Centric Multi-Intent Benchmark for Evaluating Large Language Models

Jiayin Wang[1], **Fengran Mo**[2], **Weizhi Ma**[3*], **Peijie Sun**[1], **Min Zhang**[1*], **Jian-Yun Nie**[2]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China.
[2]Université de Montréal, Québec, Canada.
[3]Institute for AI Industry Research, Tsinghua University, Beijing, China.
`JiayinWangTHU@gmail.com, mawz@tsinghua.edu.cn, z-m@tsinghua.edu.cn`

Abstract

Large language models (LLMs) are essential tools that users employ across various scenarios, so evaluating their performance and guiding users in selecting the suitable service is important. Although many benchmarks exist, they mainly focus on specific predefined model abilities, such as world knowledge, reasoning, etc. Based on these ability scores, it is hard for users to determine which LLM best suits their particular needs. To address these issues, we propose to evaluate LLMs from a user-centric perspective and design this benchmark to measure their efficacy in satisfying user needs under distinct intents. Firstly, we collect 1,846 real-world use cases from a user study with 712 participants from 23 countries. This first-hand data helps us understand actual user intents and needs in LLM interactions, forming the **U**ser **R**eported **S**cenarios (URS) dataset, which is categorized with six types of user intents. Secondly, based on this authentic dataset, we benchmark 10 LLM services with GPT-4-as-Judge. Thirdly, we show that benchmark scores align well with human preference in both real-world experience and pair-wise annotations, achieving Pearson correlations of 0.95 and 0.94, respectively. This alignment confirms that the URS dataset and our evaluation method establish an effective user-centric benchmark. The dataset and code are publicly available[1].

1 Introduction

Large language models (LLMs) are rapidly developing and gradually changing the way people interact with computing systems and permeating diverse facets of work and daily life (Wang et al., 2023a). Although LLMs show powerful capacities for completing various tasks (Chang et al., 2023), it is essential to understand how they satisfy users' actual intents and needs (Elangovan et al., 2024). According to related studies, many attempts have been made to evaluate LLMs, which may be divided into following two groups.

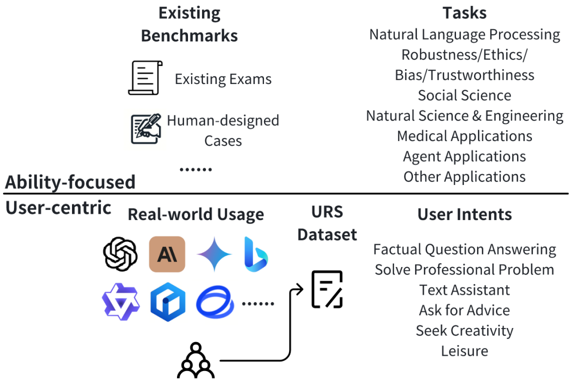


Figure 1: Existing benchmarks are mainly model ability-focused and categorized by tasks (Chang et al., 2023). We benchmark LLMs on **U**ser **R**eported **S**cenarios (URS), which are user-centric, intent-driven, multi-cultural, and multi-LLM usage cases involved.

The first group of work (Hendrycks et al., 2020; Clark et al., 2018; Zellers et al., 2019) considers LLMs as task-completion models and mainly employs close-domain assessments to measure their various task-solving abilities, such as world knowledge, mathematics ability, coding skill, reasoning, planning, agent task, etc. (Chang et al., 2023). However, the existing attempts focused on predefined tasks, without considering the roles that humans would like LLMs to play in real-world scenarios. For example, while a task was evaluated, it was unclear if the task was important to humans and frequently required by users in real applications.

The second group evaluates whether LLMs' responses align with human preferences (Wang et al., 2023b). Table 1 summarizes the user-related LLM evaluation benchmarks. Their evaluation datasets include synthetic data (Li et al., 2023b), human-written data (Zheng et al., 2024), or selected user logs from certain LLM services (Lin et al., 2024). We can observe that these benchmarks may be limited in terms of data quality, quantity, or focus on

---

*Corresponding Authors

[1] `https://github.com/Alice1998/URS`

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3588–3612
November 12-16, 2024 ©2024 Association for Computational Linguistics

Benchmark	Cases	Language	Source	Category
AlpacaEval	805	English	existing datasets and synthetic data	5 sources
MT-Bench-101	1,388	English	synthetic data	13 tasks
TencentLLMEval	3,000	English, Chinese	human designed and written tasks	7 areas
MT Bench	80	English	human written questions	8 tasks
AlignBench	683	Chinese	selected questions in ChatGLM logs	8 tasks
WildBench	1,024	English	selected questions in ChatGPT logs	12 tasks
URS (Ours)	1,846	English, Chinese	self-reported logs with 15 LLMs	6 intents

Table 1: Comparison between user-related LLM benchmarks with our URS benchmark. The compared studies are AlpacaEval (Li et al., 2023b), MT Bench-101 (Bai et al., 2024), TencentLLMEval (Xie et al., 2023), MT Bench (Zheng et al., 2024), AlignBench (Liu et al., 2023a) and WildBench (Lin et al., 2024).

a single resource. They do not reflect the true distribution of intended utilization of LLMs by real users and their effectiveness for these intents.

Besides, the categorization in current benchmarks is mainly focused on single-step tasks (Zhao et al., 2023; Liu et al., 2023a). However, end users might be more interested in knowing the ability of LLMs to satisfy their intended needs, rather than LLMs' performance on predefined individual tasks. In particular, a user intent might involve multiple model abilities with a series of tasks. For example, when a user is asking for advice about travel plans, this might involve model abilities of planning, personalization, and calling APIs. Consequently, the categorization of model-side abilities complicates the assessment of LLM performance in user-side scenarios, making it challenging for users to select the suitable service tailored to their specific needs.

This paper addresses the above issues in LLM evaluations by developing a user-centric benchmark to examine whether LLMs' behaviors satisfy user needs in real-world scenarios. As shown in Figure 1, our work is different with the existing ability-focused benchmarks and highlighted with the three unique characteristics:

User-centric. We benchmark LLMs from the user perspective in both dataset construction and evaluation designs. Our constructed **U**ser **R**eported **S**cenarios (URS) benchmark encompasses 1,846 authentic interactions (1,014 in English and 832 in Chinese) across 15 diverse LLM services, contributed by 712 users from 23 countries, each vetted through third-party manual quality checks. This dataset, reflecting the real-world usage of LLMs, is publicly available for research with user consent. Based on this first-hand data, we design evaluations to measure LLMs' efficacy in satisfying real user needs. The benchmark results align well with human preference in both real-world user experience across intents and pairwise comparison between LLMs. This user-centric approach promotes deeper collaboration between LLMs and humans as their capabilities continue to improve.

Intent-driven. Different from the task-specific evaluations, our benchmark is categorized by user self-selected intents, including factual question answering, professional problem solving, text assisting, asking for advice, seeking creativity, and leisure. Such categorization provides the potential utilization of LLMs by real users and is validated through the user study. Based on the divided intents, the users who lack specific knowledge about the model capabilities can simplify their choice of proper service. For researchers and developers, this categorization helps to provide a more precise and targeted evaluation of how well LLMs meet user needs in diverse scenarios as one simple user intent may demand a blend of model abilities to address a series of tasks (Bolotova et al., 2022).

Multi-cultural. Our data is contributed by users from 23 countries in Asia, Europe, North America, Oceania, South America, and Africa. Their cases cover multiple cultural backgrounds, such as nations' traditional festivals, local points of interest, and pop culture across the globe. This brings diversity in LLM evaluations, which is not included in simple translations of English-dominant content (Hershcovich et al., 2022; Huang et al., 2024b).

In this study, our contributions are threefold:

(1) We collect 1,846 authentic cases from 712 global users to form the multi-intent, multi-cultural **U**ser **R**eported **S**cenario dataset.

(2) Based on this URS dataset, we propose an intent-aware evaluation method to benchmark the efficacy of 10 LLM services in satisfying user needs in diverse real-world scenarios.

(3) Experiments demonstrate that benchmark results align closely with human preferences, as evidenced by Pearson correlations of 0.95 and 0.94 with real-world user experiences and pairwise annotations. These results validate that our automated evaluation method and URS dataset establish a new and effective user-centric benchmark.

2 Related Work

2.1 LLM Benchmark

The existing benchmarks are mainly divided into two categories as illustrated in Figure 1. They are model ability-focused and user-centric.

Ability-focused benchmarks. Considering the broad capabilities of Large Language Models (LLMs), the evaluations span a diverse range of domains, typically categorized into Knowledge: MMLU (Hendrycks et al., 2020), CMMLU (Li et al., 2023a), and C-Eval (Huang et al., 2024c); Mathematical abilities: GSM8k (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), and OlympiadBench (He et al., 2024a); Coding skills: HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), or SWEBench (Jimenez et al., 2023); Reasoning: BBH (Suzgun et al., 2022; bench authors, 2023), ARC (Clark et al., 2018), and HellaSwag (Zellers et al., 2019); Agent tasks: AgentBench (Liu et al., 2023b) and ToolBench (Qin et al., 2023). Each category aims to measure specific abilities expected from LLMs.

User-Centric benchmarks. Unlike benchmarks that focus on model capabilities, a few emphasize user experience with chat-based LLMs. Table 1 details related user-centric benchmarks. However, we notice several problems with these benchmarks. AlpacaEval (Li et al., 2023b), MTBench (Zheng et al., 2024), and MT-Bench-101 (Bai et al., 2024) are limited in data quality or quantity. TencentLLMEval (Xie et al., 2023) used tasks that were designed by hand and written by 50 annotators, an approach that may lead to deviation from real-world scenarios. AlignBench (Liu et al., 2023a) and WildBench (Lin et al., 2024) are constructed based on real-world use cases from logs of one specific LLM, which may introduce bias since the users may have access limitations to certain LLMs and use LLM services selectively depending on their needs. Furthermore, the existing user-centric benchmarks mainly do not consider multilingual and multi-cultural aspects, which are covered in our dataset.

2.2 Evaluation Methods

Besides dataset constructions, evaluation methods are crucial in developing benchmarks. To simplify evaluation processes, many studies utilize multiple-choice questions as criterion (Hendrycks et al., 2020; Li et al., 2023a; Huang et al., 2024c). Others employ post-processing of generated content and apply rules or deterministic tests to assess performance (Cobbe et al., 2021; Chen et al., 2021). With the recent advancements in LLM techniques, an increasing number of approaches use a more capable AI model to evaluate answers (He et al., 2024b; Li et al., 2023b), enriching benchmark types by allowing a more diverse format of test questions. AlpacaEval (Li et al., 2023b), AlignBench (Liu et al., 2023a) adopt GPT-4 as evaluator. PRE (Chu et al., 2024) shows that GPT-4 performs well in pairwise evaluations and Huang et al. (2024a) shows that GPT4 outperforms fine-tuned judge models in generalizability and fairness. In our benchmark, we adopt previous settings for open-domain evaluation, using LLM-based pairwise evaluation.

2.3 User Intent

From a user-centric view, we should first confirm user intent and evaluate service responses correspondingly. Related ideas try to understand user intent in interacting with LLMs. Shah et al. (2023) provides a user intent taxonomy based on New Bing, which includes information retrieval, problem solving, learning, content creation, and leisure. Bodonhelyi et al. (2024) details user intent with ChatGPT with informational, problem solving, creative, educational, personal interaction, technical and professional, transactional, ethical, and philosophical intents. Inspired by these studies, we design a user intent taxonomy accordingly, which is further verified by 712 user study participants.

3 User Reported Scenarios Dataset

3.1 User Intent Taxonomy

User intents represent specific needs or goals when interacting with a service and users may anticipate different types of feedback under different intents (Bolotova et al., 2022). For example, for factual intents, users may expect brief and accurate answers; for brainstorming needs, users tend to expect rich and innovative answers. By considering different user intents, we can assess LLMs' fine-grain efficacy in diverse scenarios. Therefore, we first define the taxonomy of user intent based

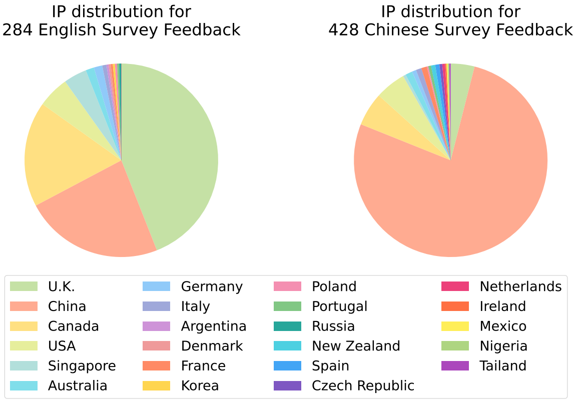


Figure 2: IP Distribution of the 712 participants.

on related work (Shah et al., 2023; Bolotova et al., 2022). They are detailed below:

Factual QA. Fast and direct access to factual information. For example, search for historical events, scientific facts, or public data.

Solving Professional Problems. Require answers and insight in specialized fields like natural sciences, humanities, or social sciences for problem-solving or learning. It involves domain knowledge, in-depth understanding, and reasoning in professional areas. For example, solving math, engineering puzzles, and conducting medical diagnoses.

Text Assistant. Need assistance with text-related tasks such as summary, translation, editing or completing content. For example, replying to an email, writing a report, or polishing a speech.

Ask for Advice. Look for opinions and suggestions for personal or professional decisions, including planning or counseling. For example, career development, personal counseling, creating travel plans, or shopping lists.

Seek Creativity. Brainstorming for inspiration and innovative ideas. For example, advertising ideas, or design inspiration.

Leisure. Engage in or look for recommendations for recreational activities, including books, music, movies, games, and other entertaining activities.

This user intent taxonomy serve as the potential utilization of LLM services by real users and is furthered verified in the following user study.

3.2 Dataset Development

To better understand LLMs' applications in real-world scenarios, we conducted a user study to collect participants' recent interactions with LLM services, following data collection strategies of related studies (Feldhus et al., 2023; Bolotova et al., 2022).

	English	Chinese	Total
Total	1,014	832	1,846
Factual QA	374	199	573
Solve Problem	180	348	528
Text Assistant	81	43	124
Ask for Advice	175	94	269
Seek Creativity	121	72	193
Leisure	83	76	159

Table 2: User Intent Distribution in the URS dataset.

3.2.1 User Study

Participants in the study were requested to provide five to ten real cases with LLMs they have used. Each reported case included the following three components: the LLM service they queried, complete conversation, and their usage intent (either selected from the predefined list or filled in by themselves). The detailed questionnaire is attached in Appendix A.

Additionally, the study investigate users' general experience with LLMs, including their satisfaction across different intents, which is used to validate benchmark alignment with human feedback in Section 4.5. We also record the anonymous, voluntarily reported demographic information in the questionnaire to examine whether this survey involves diverse groups of participants.

To capture multi-cultural scenarios, we offer the questionnaire in both English and Chinese and spread it through Prolific[2], X, and WeChat Moments platforms for global enrollment. Crowdworkers are paid according to Prolific recommendations (9£/hr * average 10 mins = 1.5£).

3.2.2 Construction

In the initial phase of dataset construction, we check the diversity of our participants. Figure 2 shows the distribution of automatically recorded IP for English and Chinese questionnaire respondents. The feedback comes from 712 participants across 23 countries, showing the diversity in distributions. Detailed profiles, including demographic information (age and occupation) and LLM usage experience, are provided in Appendix B.1.

Subsequently, we examine the feedback on user intents. There are no valid proposals under the "Others" option and manual fill-ins, suggesting a

---

[2] https://www.prolific.com. A global human-label platform for research and AI.

Intent	**Example Cases**
Factual Question Answering	EN: Tell me what bitcoin is CN: 大雪农历初几 (*When is the Major Snow in the Lunar calendar*)
Solve Professional Problems	EN: How long does it take to transfer 13.72GB if speed is 10MB/s CN: 假设你正在爬楼梯。需要n阶你才能到达楼顶。每次你可以爬1或2个台阶。你有多少种不同的方法可以爬到楼顶呢？ (*Suppose you are climbing a staircase. It takes n steps to reach the top. You can climb 1 or 2 steps at a time. How many different ways can you climb to the top?*)
Text Assistant	EN: Help me rephrase the document: The NBA's inaugural in-season tournament has concluded with the Los Angeles Lakers beating the Indiana Pacers 123-109 to lift the NBA Cup, with the team's winning players ...... CN: 请你帮我撰写一段给领导2024龙年的拜年微信。 (*Help me compose a Spring Festival's Eve Wechat text to my leader for the Year of the Dragon 2024.*)
Ask for Advice	EN: How long can vegan blueberry croissants be kept for once they've been baked off from frozen CN:在8度天气里，适合穿哪些款式和类型的外套？ (*What styles and types of jackets are appropriate in 8 degree weather?*)
Seek Creativity	EN: Explain photosynthesis as if I am 9 years old. CN: 帮我的猫咪起个好听的英文名字，她的中文名字叫小惰娇，1岁的女娃 (*Help me give my cat a nice English name. Her Chinese name is Xiao Duojiao (Lazy Cutie). She is a 1-year-old girl.*)
Leisure	EN: What is the best order to watch films and tv shows in the MCU? CN: 推荐一些比较容易翻唱的粤语歌给我 (*Recommend me some Cantonese songs that are easier to cover*)

Table 3: Example English and Chinese Cases under each Intent. The notation "EN" represents that the case is in English. "CN" means the cases are reported in Chinese and we attach their English translation in italics. "......" represents the text behind is omitted. Note that due to space limitations, we present cases with shorter length in the above table. The average length of each question is 29.65 tokens in the URS dataset.

general comprehensive coverage of the proposed user intent taxonomy.

Finally, we carry out meticulous data processing to construct the **U**ser **R**eported **S**cenario (URS) dataset as original feedback inevitably contains noise. Steps include format validation, elimination of intra-user duplication, and extraction of initial valid questions from each conversation. The dataset construction is augmented with third-party manual quality assessments to objectively filter out low-quality cases and confirm the exclusion of personal information. Detailed ethics considerations about this dataset are discussed in Section 5.

3.3 Dataset Information

Table 2 reports the statistics of the URS dataset, with all data sourced from the above user study. In all intents, those that are relatively subjective, including Ask for Advice, Seek Creativity and Leisure, constitute about 33% of the total use cases. As the traditional benchmarks primarily focus on evaluating LLMs' correctness of subjective questions (Chang et al., 2023), this objective portion of user needs is often overlooked.

Besides the diversity in user intents, the URS benchmark also includes interactions from 15 different LLM services. The distribution is detailed in Appendix B.2. This breadth reduces potential biases stemming from reliance on a single LLM service. These biases can arise from users' limited access to certain LLMs and their preference for specific services in some scenarios. Among this diversity, we observe a long-tail distribution in model usage. As we did not actively control or select use cases of certain LLMs during the dataset collection and construction processes, this may indicate the natural distribution in real-world usages.

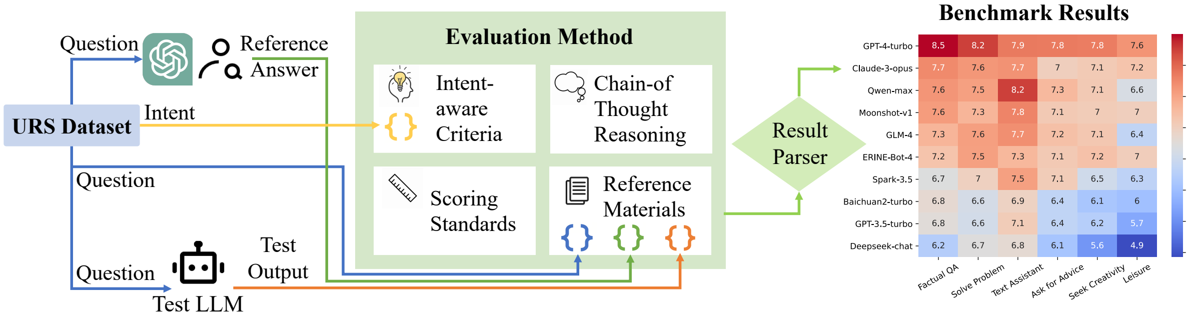

Question
Reference
Answer
URS Dataset
Intent
Question
Question
Test LLM
Test
Output
Evaluation Method
Intent-
aware
Criteria
Chain-of
Thought
Reasoning
Scoring
Standards
Reference
Materials
Result
Parser
Benchmark Results
Factual QA
Solve Problem
Text Assistant
Ask for Advice
Seek Creativity
Leisure
GPT-4-turbo 8.5 8.2 7.9 7.8 7.8 7.6
Claude-3-opus 7.7 7.6 7.7 7 7.1 7.2
Qwen-max 7.6 7.5 8.2 7.3 7.1 6.6
Moonshot-v1 7.6 7.3 7.8 7.1 7 7
GLM-4 7.3 7.6 7.7 7.2 7.1 6.4
ERINE-Bot-4 7.2 7.5 7.3 7.1 7.2 7
Spark-3.5 6.7 7 7.5 7.1 6.5 6.3
Baichuan2-turbo 6.8 6.6 6.9 6.4 6.1 6
GPT-3.5-turbo 6.8 6.6 7.1 6.4 6.2 5.7
Deepseek-chat 6.2 6.7 6.8 6.1 5.6 4.9
8.0
7.5
7.0
6.5
6.0
5.5
5.0

Figure 3: Evaluation Procedure. For each evaluation instance, the evaluator is provided with the user intent, five intent-aware criteria, chain-of-though reasoning steps, scoring standards for each two-point segment, addition with the question, an 8-score reference answer for this question, and the test LLM output for evaluation. Then, a parser will extract the final score from the evaluator's detailed rating content to form the benchmark.

Time is also an important factor influencing LLM evaluations. Since the user study was conducted in February 2024, the collected cases in this dataset took places before this time.

Table 3 showcases exemplary instances from the dataset. Real-world scenarios are often brief in articulation but complex in resolutions and span an expansive range of situations. Notably, beyond linguistic differences, the cases in this dataset also encompass a variety of cultural backgrounds. These include traditional events, such as Major Snow and the Spring Festival in the lunar calendar as well as popular entertainment information that are globally or locally recognized, such as the Marvel Cinematic Universe and Cantonese songs. These multicultural contexts could enhance the benchmarking of LLMs, addressing more diverse needs of the global user base.

4 URS Benchmark

4.1 Evaluation Framework

We aim to design a evaluation framework that can autonomously, expeditiously, and precisely determine the performance of any LLMs. The automation is made possible by using a strong model (e.g., GPT-4) acting as the evaluator, whose effectiveness is validated in previous work for open-domain evaluations (Chang et al., 2023; Sottana et al., 2023; Liu et al., 2023c) and further examined in Section 4.5. Aiming for high speed and precision, we implement a direct pair-wise scoring approach, providing a fixed reference answer for each question, in contrast to point-wise rating and Elo rating methods, as has been demonstrated to be effective in the previous study (Li et al., 2023b; Liu et al., 2023a). As shown in Figure 3, we categorize the evaluation instruction into 4 parts, including intent-aware criteria, chain-of-thought reasoning steps, scoring standards, and reference materials, which are described below (The detailed instructions are provided in Appendix C.2):

Intent-aware criteria. During evaluations, we inform the evaluator about the user intent for each question and provide five specific evaluation criteria tailored to that intent. This approach is designed to help evaluators accurately capture user needs across various intent types and provide precise scorings. Detailed illustrations of these criteria and the correspondence relationship between user intents and evaluation criteria are provided in Appendix C.1.

Chain-of-thought reasoning steps. To fully leverage the reasoning capabilities of evaluators, the instruction provides four steps before assigning final ratings. These steps include contrasting the reference answer with the test response, scoring each criterion on a 1-10 scale, reassessing whether the response meets user needs, and integrating these assessments to determine the final score.

Scoring standards. To ensure accurate and differentiated scoring, we provide standards for each scoring segment. Each 2-point increment constitutes a distinct segment. The answer furnished by us is an 8-score reference. Besides, a note is provided to indicate that longer responses are not necessarily better to limit the potential length bias.

Reference materials. The instruction provided for the evaluator includes the question given by real users, their intent for asking, a fixed reference answer for this question (provided by a strong LLM, which is GPT-4 in this paper and checked manually), and the response generated by the test LLM for evaluation.

	Solve Problem	Factual QA	Text Assistant	Ask for Advice	Seek Creativity	Leisure	All
Cases	404	259	82	116	86	83	1024
GPT-4o	*8.24	*8.49	**7.86**	*7.85	*7.83	*7.65	*8.15
Claude-3-opus-20240229	**7.62**	**7.73**	7.66	7.01	7.08	**7.16**	**7.50**
Qwen-max	7.53	7.64	*8.24	**7.28**	7.10	6.63	7.48
Moonshot-v1-8k	7.33	7.58	7.77	7.09	7.05	7.01	7.35
GLM-4	7.55	7.32	7.68	7.20	7.12	6.37	7.33
ERNIE-Bot-4	7.54	7.16	7.30	7.09	**7.19**	7.02	7.31
Spark-3.5	6.97	6.71	7.53	7.05	6.54	6.33	6.87
Baichuan2-Turbo	6.59	6.82	6.95	6.35	6.11	6.02	6.56
GPT-3.5-turbo	6.56	6.76	7.05	6.35	6.24	5.75	6.53
Deepseek-chat	6.72	6.22	6.81	6.09	5.56	4.93	6.29

Table 4: URS Benchmark Results. For each intent and the overall scenarios, we mark the three best-performing LLM services, with the first marked '*', the second bolded, and the third underlined.

4.2 Experimental Settings

Among the top-used LLM services reported in the previous user study, we test all the LLMs with available APIs, including Baichuan2-Turbo (Yang et al., 2023), Claude-3-opus (Anthropic, 2024), Deepseek-chat (Bi et al., 2024), ERNIE-Bot-4 (Baidu), GLM-4 (Du et al., 2021), GPT-3.5-turbo (Achiam et al., 2023), GPT-4o (Achiam et al., 2023), Moonshot-v1-8k (MoonshotAI), Qwen-max (Bai et al., 2023), Spark-3.5 (iFLYTEK). We do not adjust the default temperature setting of LLMs based on query intent. This helps us to model the efficacy of LLM services in different real-world scenarios without user intent information. We benchmark the above LLMs on 1024 human-examined cases in the URS dataset.

In the following sections, we conduct extensive experiments to answer four research questions:

(RQ1) What is the performance of different LLMs in the user-centric evaluation perspective?

(RQ2) Are the evaluation results stable across different LLMs as evaluators?

(RQ3) Do the benchmark results align with real user perceptions?

(RQ4) What multi-cultural features do we observe in this benchmark?

4.3 Overall Results (RQ1)

The overall benchmark results are shown in Table 4. We have the following observations:

(1) GPT-4 yields an average score of 8.15 when its previously generated responses are used as 8-score references in the scoring instruction. This difference indicates an acceptable level of variance from the established standard. The benchmark stability across different evaluators is further discussed in the next section.

(2) There is a noticeable stratification in scores across different models. Apart from GPT-4, the leading group comprises Claude-3 and Qwen-max, followed by a secondary tier including Moonshot-v1, GLM-4, and ERNIE-Bot-4. The performances of the third tier are comparable to that of GPT-3.5.

(3) In examining efficacy across diverse user intents, GPT-4 generally demonstrates superior performance, except for the Text Assistant category, where Qwen-max prevails.

Furthermore, a comparative analysis of LLMs across different user intents reveals a pronounced proficiency in objective contexts, including Solve Problem and Factual QA. This is probably because subjective scenarios require more diverse competencies such as personalization, creativity and humor, where current LLMs are relatively weak.

Separated results of English and Chinese scenarios are presented in Appendix D.1. Scoring examples are provided in Appendix D.2.

4.4 Cross Validation between GPT-4 and Claude-3 (RQ2)

In the overall benchmark result, we adopt GPT-4's direct output to the question as a reference for score 8, and use GPT-4 as the evaluator base for final scoring, as detailed in Setion 4.1. Although this approach is widely adopted, it may cause favor to responses similar to GPT-4. To counteract this bias, we re-evaluated the performance of the top-2 LLMs, GPT-4o, and Claude-3-opus, employing

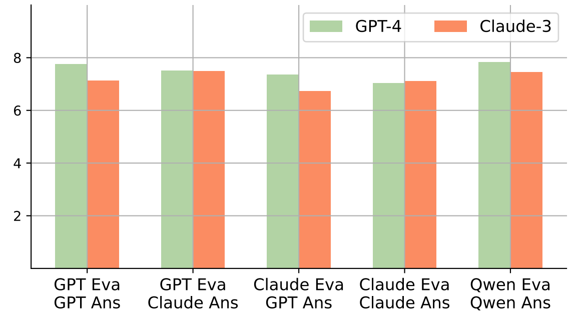


Figure 4: Comparison between GPT-4 and Claude-3. "A Eva B Ans" indicates the evaluation setting, where ’A’ denotes the evaluator LLM and ’B’ represents the source LLM used to generate 8-point reference answers. These results show that GPT-4 slightly outperforms Claude-3.

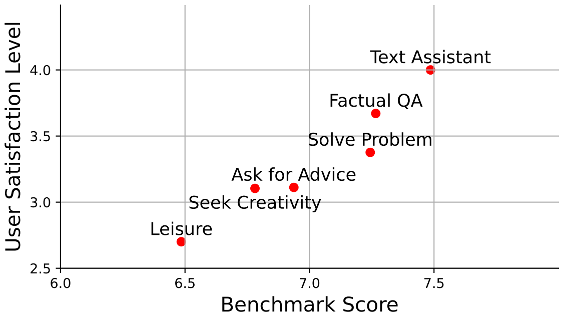


Figure 5: Benchmark Score and User Reported Satisfaction Correlate Well across Intents. "Benchmark Score" is averaged under different intents. "User Reported Satisfaction" is the average satisfaction level reported in the user study. Intents are ranked by user satisfaction.

cross-validation techniques. We run 5-fold evaluation settings on randomly selected 200 cases, using GPT and Claude as the reference generator and base evaluator respectively, and the third performed LLM, Qwen, as both reference generation and evaluation for a third-party evaluation. As illustrated in Figure 4, when GPT-generated responses are used as references, GPT-4 consistently achieves higher scores no matter the evaluator models. When Claude-generated responses are adopted in evaluations, the results are comparable. In third-party assessments, GPT-4 also maintained a superior ranking over Claude-3. This order is consistent with the overall benchmark results, indicating that the potential bias towards GPT does not affect the benchmark ranking.

4.5 Alignment with Human Experience (RQ3)

To further evaluate the efficacy of LLM-as-Judge design in this benchmark, we analyzed benchmark accuracy through the lens of real-world human experience, examining both intent-wise user satisfaction and pairwise LLM-wise comparisons.

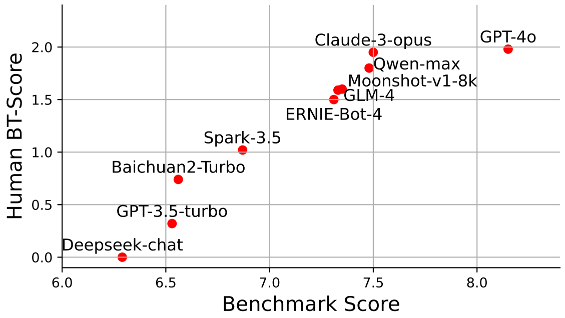


Figure 6: LLMs’ Benchmark Ranking Aligns with the Order from Human Pairwise Annotation.

For intent-wise evaluation, user satisfaction levels were collected in the user study detailed in Section 3.2.1, involving 420 global participants who rated their satisfaction on a five-point scale for each intent while using LLM services. A high Pearson correlation coefficient of 0.95 between the average benchmark score and user-reported satisfaction across different intents, as shown in Figure 5, confirms the alignment. Notably, subjective scenarios such as Ask for Advice, Seek Creativity and Leisure received lower ratings than more subjective ones in both benchmark results and user feedback. This indicates the current LLM services still need improvements in satisfying real-world user needs.

In addition to the above alignment analysis based on large-scale user studies, we further conducted pair-wise annotations according to methods in Chatbot Arena (Chiang et al., 2024) to check the consistency between the designed automatic evaluation and human preferences. Due to space limitations, the detailed approach is provided in Appendix D.3. Based on 400 paired comparisons assessed by 5 human annotators, we computed Bradley-Terry coefficients (Bradley and Terry, 1952) for each LLM. As shown in Figure 6, the ranking of LLMs according to these coefficients aligns with the benchmark rankings with Pearson r=0.94, further affirming the effectiveness of evaluation designs.

4.6 Analysis of Multi-cultural Features (RQ4)

We further analyzed the multi-cultural features in the benchmark. Based on the location of question providers, we divide the dataset to observe the performance differences of LLMs when answering questions from users with different cultural backgrounds. Table 5 presents the results of the top five countries with the highest number of survey participants. Among the 10 tested LLMs, 6 per-

	UK	China	Canada	USA	Australia	Others
GPT-4o	8.68	8.31	**8.36**	9.13	8.22	8.09
Claude-3-opus	7.05	7.54	7.64	**8.13**	7.00	7.49
Qwen-max	7.84	8.72	8.62	8.13	**8.88**	7.30
Moonshot-v1-8k	7.58	7.96	7.45	**8.63**	7.22	7.18
GLM-4	7.55	7.61	7.86	**8.04**	7.44	7.26
ERNIE-Bot-4	6.74	**8.32**	7.41	8.13	6.67	7.22
Spark-3.5	7.03	**7.92**	7.41	7.29	6.56	6.75
Baichuan2-Turbo	6.42	7.00	6.50	**7.04**	5.78	6.53
GPT-3.5-turbo	6.74	6.33	7.14	**7.79**	6.67	6.48
Deepseek-chat	6.50	6.96	7.23	**7.54**	6.11	6.17

Table 5: Multi-Cultural Analysis: we group the benchmark questions by user locations. For each LLM (line-wise), we bold the best performing result across countries. Results show that LLMs generally perform better on questions raised by USA users, reflecting the need for further attention to multi-cultural issues.

form the best on questions raised by USA users. This reveals a potential cultural bias in LLM performance and evaluation, reflecting further emphasis on multi-cultural evaluations and improvements.

5 Conclusion

By focusing on real-world cases, user intent categorization, and the intent-aware evaluation method, we establish a benchmark of high alignment with human preference, endeavoring to advance the user-centric evaluation of LLMs. We hope the insights gained from this research will not only help improve the performance of LLMs but also foster a deeper understanding of how these services can be effectively integrated into people's lives to enhance productivity, creativity, and overall welfare.

Limitations

While evaluating LLMs from the user-centric perspective offers considerable advantages, there are certain limitations in our URS benchmark.

Evaluation Method: Our approach aligns with prevailing practices in open-ended question evaluation, where GPT-4 is employed as the base evaluator. This could potentially lead to a bias toward responses resembling GPT's style, thereby influencing the neutrality of the evaluation. We analyzed through cross-validation that this approach does not affect the top 2 LLM orders. Besides, analysis shows that benchmark score aligns with real-world user experience and human preferences in pairwise annotations. These further tested the validity of our method. In addition, the user study and human pairwise annotations could be further scale up to enhance persuasiveness.

Besides, we notice it is very important to evaluate LLM services' performances in multi-turn conversational settings, as there are a few cases where the LLMs do not answer the user question and ask to provide more detailed information about user intents. These may be judged unsuccessful in a single-turn assessment, but the service might be capable of meeting this user need.

Note that this benchmark is not holistic, as we mainly measure LLMs' efficacy in satisfying real-world user needs. User satisfaction is indeed crucial, while it represents one aspect of LLM performance. Other critical facets include effectiveness, efficiency, ease of use, and error tolerance. In this study, our primary focus is on user-centric evaluation of LLMs, which we have validated through strong alignments with human preferences to address its effectiveness.

Data Distribution: The data for our study was mainly sourced from a global researcher recruitment platform, which does not ensure a random selection of all participants using LLM services. Consequently, the demographic distribution of our dataset may not accurately reflect that of typical LLM users or the global population. While as an independent third party unaffiliated with specific LLM services, and given our non-selective approach to data collection, our methodology incorporates greater diversity and offers a closer alignment with real-world usage scenarios compared to datasets that are human-designed, model synthetic, or selected by non-third parties or single-sourced.

Ethical Considerations

In the user study, participants were informed that their reported conversations would be publicly released for research purposes. The notification is detailed in Appendix A. At the same time, they voluntarily chose whether to provide anonymized basic information, such as age group and profession.

Additionally, the survey design included an anchor question requiring participants to select specific options; responses that failed to adhere to these instructions were discarded as malicious feedback.

The questionnaire was provided in English and Chinese (Mandarin), the top 2 widely spoken languages[3], and distributed via the global recruitment platform[4]. In this process, we did not filter any attributes as this could actively introduce bias. Following data collection, a third party conducted manual reviews to ensure the data's quality, harmlessness, and the absence of any personal information.

This dataset could be used for research purposes to understand user intents and real-world needs, evaluate conversational systems' responses, and train for better alignment with human preference. Note that this URS dataset uses an Apache License.

Acknowledgement

This work is supported by the Natural Science Foundation of China (Grant No. U21B2026, 62372260) and Quan Cheng Laboratory (Grant No. QCLZD202301).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baidu. Ernie bot.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Anna Bodonhelyi, Efe Bozkir, Shuo Yang, Enkelejda Kasneci, and Gjergji Kasneci. 2024. User intent recognition and satisfaction with large language models: A user study with chatgpt. *arXiv preprint arXiv:2402.02136*.

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1196–1207.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Mark Chen, Jerry Tworek, Heewoo Jun, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, , et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. 2024. Pre: A peer review based large language model evaluator. *arXiv preprint arXiv:2401.15641*.

---

[3] `https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers`

[4] `https://www.prolific.com`

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark othersChen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. 2024. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. *ACL 2024*.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. Interrolang: Exploring nlp models and datasets through dialogue-based explanations. *arXiv preprint arXiv:2310.05592*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024a. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Chaoqun He, Renjie Luo, Shengding Hu, Yuanqian Zhao, Jie Zhou, Hanghao Wu, Jiajie Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024b. Ultraeval: A lightweight platform for flexible and comprehensive evaluation for llms. *ACL 2024 demo paper*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, et al. 2022. Challenges and strategies in cross-cultural nlp. *ACL 2022*.

Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. 2024a. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, et al. 2024b. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024c. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

iFLYTEK. Spark.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Bill Yuchen Lin, Khyathi Chandu, Faeze Brahman, Yuntian Deng, Abhilasha Ravichander, Valentina Pyatkin, et al. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. 2023a. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023b. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

MoonshotAI. Moonshot.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis.

Chirag Shah, Ryen W White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Snigdha Sarathi Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Xiaochuan Ni, et al. 2023. Using large language models to generate, validate, and apply user intent taxonomies. *arXiv preprint arXiv:2309.13063*.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of gpt-4: reliably evaluating large language models on sequence to sequence tasks. *arXiv preprint arXiv:2310.13800*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them.

Bryan Wang, Gang Li, et al. 2023a. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Shuyi Xie, Wenlin Yao, Yong Dai, Shaobo Wang, Donlin Zhou, Lifeng Jin, Xinhua Feng, Pengzhi Wei, Yujie Lin, Zhichao Hu, et al. 2023. Tencentllmeval: a hierarchical evaluation of real-world capabilities for human-aligned llms. *arXiv preprint arXiv:2311.05374*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2023. (inthe) wildchat: 570k chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Appendix

A User Study

At the start of the user survey, we have this notification:

Your dialog reported in this questionnaire may be publicly released for research purposes, so please avoid including personal information, and we will also conduct a secondary check, thank you!

Subsequently, participants are asked to provide five to ten distinct real use cases with LLM services. Each use case should comprehensively cover the following aspects:

1. LLM Service Utilized:

Which of the following large language models did you ask the question to

2. Conversation Content:

Your question and the system's output
It is recommended to directly share your conversation link!
If you fill it out manually, please respectfully follow the format below.
Question: xxx
Answer: xxx

3. Usage Intent:

Your intent behind this question is

Participants are also asked to voluntarily report the following demographic information anonymously:

1. Age Group: To understand the age distribution of LLM users.

2. Profession: To gauge the professional background and its influence on LLM usage.

3. IP Location: To capture the geographical distribution of participants.

General usage Experience with LLMs:

1 LLM used:

Large Language Models you have used

2 Usage Frequency:

How often do you use LLMs

3 Satisfaction with LLMs:

Your general satisfaction with LLM services across each intent

B URS Dataset Information

B.1 Participants

Figure 7 and Figure 8 report the demographic information and the LLM services usage pattern of the user study participants, which demonstrates the diversity of the benchmark contributors.

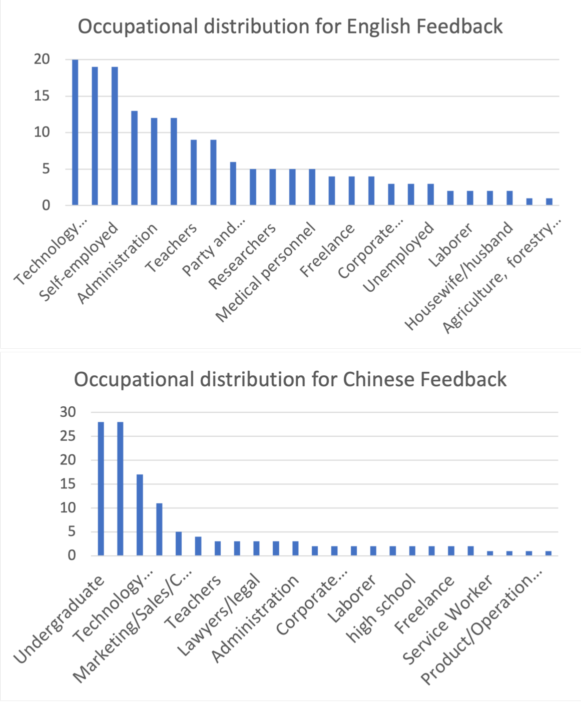


Figure 7: Occupation distribution of survey participants.

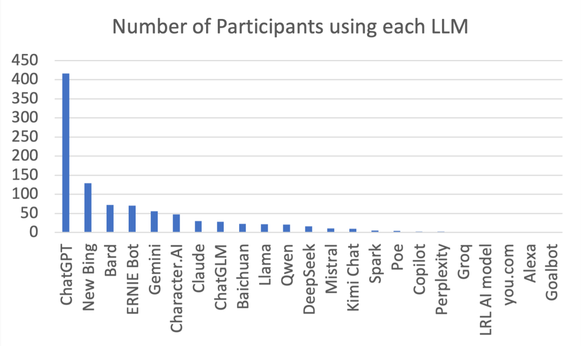


Figure 8: Number of Participants reported having used each LLM.

B.2 Distribution of LLM usage

Table 6 presents the distribution of LLM usage from the collected user study. ChatGPT is the most popular one in both language backgrounds, while the others show some differences.

C Evaluation

C.1 Intent-aware Criteria

The definition of intent-aware criteria is shown in Table 9. We evaluate each defined user intent with five selected criteria as described in Table 7.

C.2 Instruction

Here is the template for each evaluation case.

English		**Chinese**	
Total	993	Total	994
ChatGPT	809	ChatGPT	818
Bard/Gemini	69	ERNIE Bot	61
New Bing	41	New Bing	41
Character.AI	14	Baichuan	7
ChatGLM	7	Character.AI, etc.	6

Table 6: Distribution on Top 5 Source LLMs. In the URS dataset, most conversations are reported with their source LLMs. Notely, for Chinese feedback, Character.AI, Claude, Mistral, and Qwen are tied for 5th place.

Instruction Template

You are asked to assess the quality of an AI assistant’s answer to a user’s question as an impartial judge. Since the type of answer you are evaluating is [{user_intent}], you need to evaluate the answer in the following 5 criteria:
1 {criteria_dict[user_intent][0]}
2 {criteria_dict[user_intent][1]}
3 {criteria_dict[user_intent][2]}
4 {criteria_dict[user_intent][3]}
5 {criteria_dict[user_intent][4]}

We will provide you with the user’s question, an 8-score reference answer, and answers from the AI assistant that needs your assessment. When starting your evaluation, you need to follow the reasoning steps below:
1. Compare the AI assistant’s answer with the reference answer, point out any shortcomings in the AI assistant’s answer, and explain further.
2. Evaluate the AI assistant’s answer in terms of the different criteria, giving each criterion a score from 1 to 10 after the evaluation of each.
3. Finally, combine the evaluations from each criterion and give the AI assistant’s answer a composite score of 1 to 10.
4. Your scoring needs to be as rigorous as possible and adhere to the following scoring rules: in general, the higher the quality of the model’s answers, the higher the score.

When the model answer has irrelevance to the question, or intrinsically factually incorrect, or generates harmful content, the total score should be 1 to 2;
When the model answer has no serious errors and is largely harmless, but is of low quality and does not meet user requirements, the total score must be 3 to 4;
When the model answer basically meets the user’s needs but performs poorly on some criteria and is of medium quality, the total score can be 5 to 6;
When the quality of the model response is similar to the reference answer and performs well in all criteria, the total score should be 7 to 8;
A score of 9 to 10 can only be achieved if the model significantly exceeds the quality of the reference answer, adequately addresses the user’s question and all the needs, and is close to a perfect score on all criteria. As an example, the reference answer would receive a score of 8.
Do not let the length of the answer affect your score. Longer answers are not necessarily better. Answers that are concise and meet the above requirements are good.

You need to evaluate and explain before you score. Your explanation of each criterion needs to be followed by the scoring. After that, at the end of your answer, return all of your scores in the following dictionary format, including the curly brackets, and make sure that your scores are integers:
{’Dimension 1’: score, ’Dimension 2’: score, ... , ’Final Score’: score}, e.g. {’{criteria_dict[user_intent][0]}’: 9, ’{criteria_dict[user_intent][1]}’: 6, ... , ’Final Score’: 7}.

User Question: question
[Start of Reference Answer]
{reference_ans}
[End of reference answer]
[start of AI assistant’s answer]
{test_model_output}
[end of AI assistant’s answer

The instructions for evaluation with four parts are shown in Table 10, 11 and Table 12 in English and Chinese, respectively. We use these designed instructions to guide the evaluator to output the evaluation results for benchmarking.

D Benchmark Results

D.1 Bi-lingual Results

Table 13 and Table 14 show the English and Chinese’s separated benchmark results across different intent.

D.2 Scoring Examples

For subjective intents, we list the example cases as shown in Table 17, 18 and 19.

D.3 Alignment with Human Annotation

We aim to further investigate whether the benchmarked LLM order aligns with human preferences by conducting annotation experiments. Following the method outlined in Chatbot Arena (Chiang et al., 2024), we enlisted five human annotators (three males and two females) to evaluate paired questions as shown on the UI in Figure 9. We randomly selected 400 questions from the 1024-case benchmark and randomly paired two LLM outputs for each question. Annotators, blind to which LLMs generated the responses, were tasked with choosing their preferred answer from the two provided, or indicating if the choice was equal or

Intents	Intent-aware Criteria
Factual QA	Factuality, User Satisfaction, Clarity, Completeness, Logical Coherence
Solve Professional Problem	Factuality, User Satisfaction, Clarity, Logical Coherence, Completeness
Text Assistant	Clarity, User Satisfaction, Logical Coherence, Fatuality, Creativity
Ask for Advice	User Satisfaction, Factuality, Fairness and Responsibility, Creativity, Richness
Seek Creativity	User Satisfaction, Logical Coherence, Creativity, Richness, Factuality
Leisure	User Satisfaction, Engagement, Appropriateness, Creativity, Factuality

Table 7: Evaluation Criteria under Different User Intents.

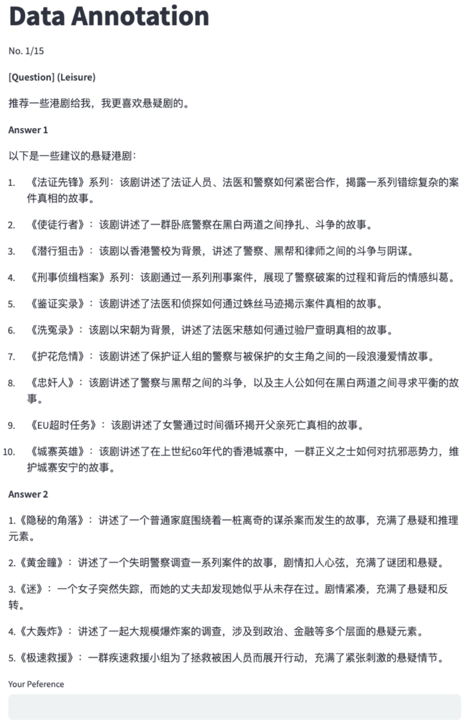
Model	Max L	Accuracy
Claude-3-opus	19	100%
GPT-4	12	100%
ERNIE-Bot-4	9	94%
GPT-3.5-turbo	8	86%
Qwen-max	4	57%
spark-3.5	4	50%

Table 8: Addition Ability. "Max L" represents the longest sequence of consecutive additions attainable at a minimum accuracy rate of 80%. "Accuracy" denotes the success ratio in 90 trials covering the length of 2 to 9 with 10 test cases of each length. Outcomes are presented for LLMs with an accuracy of 50% and above.

undeterminable due to their knowledge limitations. Of the 400 paired comparisons, 50 were marked "cannot be determined." Utilizing the remaining annotations, we calculated the Bradley-Terry (BT) scores to establish LLM rankings based on pairwise competition outcomes. As shown in Figure 6, the resulting LLM order aligns with the benchmark rankings, and the BT score demonstrates a Pearson correlation of 0.94 with the benchmark score. The evaluation data is anonymously available.

D.4 Comparison with Different Evaluation Method

Figure 10 shows cross-validation evaluation results. We change the model to generate reference answers and the model to serve as evaluators respectively. Results show while using GPT4 output as the reference answer for 8 scores, the order between the three models remains unchanged no matter the evaluation model. While using Qwen or GLM's output as both reference generator and evaluator, there exists a strong incline towards the model itself.



Data Annotation

No. 1/15

[Question] (Leisure)

推荐一些港剧给我，我更喜欢悬疑剧的。

Answer 1

以下是一些建议的悬疑港剧：

1. 《法证先锋》系列：该剧讲述了法证人员、法医和警察如何紧密合作，揭露一系列错综复杂的案件真相的故事。
2. 《使徒行者》：该剧讲述了一群卧底警察在黑白两道之间挣扎、斗争的故事。
3. 《潜行狙击》：该剧以香港警校为背景，讲述了警察、黑帮和律师之间的斗争与阴谋。
4. 《刑事侦缉档案》系列：该剧通过一系列刑事案件，展现了警察破案的过程和背后的情感纠葛。
5. 《鉴证实录》：该剧讲述了法医和侦探如何通过蛛丝马迹揭示案件真相的故事。
6. 《洗冤录》：该剧以宋朝为背景，讲述了法医宋慈如何通过验尸查明真相的故事。
7. 《护花危情》：该剧讲述了保护证人组的警察与被保护的女主角之间的一段浪漫爱情故事。
8. 《忠奸人》：该剧讲述了警察与黑帮之间的斗争，以及主人公如何在黑白两道之间寻求平衡的故事。
9. 《EU超时任务》：该剧讲述了女警通过时间循环揭开父亲死亡真相的故事。
10. 《城寨英雄》：该剧讲述了在上世纪60年代的香港城寨中，一群正义之士如何对抗邪恶势力，维护城寨安宁的故事。

Answer 2

1.《隐秘的角落》：讲述了一个普通家庭围绕着一桩离奇的谋杀案而发生的故事，充满了悬疑和推理元素。

2.《黄金瞳》：讲述了一个失明警察调查一系列案件的故事，剧情扣人心弦，充满了谜团和悬疑。

3.《迷》：一个女子突然失踪，而她的丈夫却发现她似乎从未存在过。剧情紧凑，充满了悬疑和反转。

4.《大轰炸》：讲述了一起大规模爆炸案的调查，涉及到政治、金融等多个层面的悬疑元素。

5.《极速救援》：一群疾速救援小组为了拯救被困人员而展开行动，充满了紧张刺激的悬疑情节。

Your Peference

Figure 9: Data Annotation Page.

We also check the rating bias of different LLMs as evaluators. Rating bias is defined in Equation 1.

$$\text{Rating Bias}(A) = \frac{\sum_B \text{Rating}_{A-B}(B) - r}{\sum_B 1}, r = 8 \tag{1}$$

where $\text{Rating}_{A-B}(B)$ representing using LLM A as an evaluator and the output of LLM B as a reference answer to evaluate the performance of LLM B. r is the score for reference, in our cases, it is equal to 8. Note that in the experiments, reference answer and output awaiting score are two-time outputs of the same model, so if model B is stable and model A can score accurately, the $\text{Rating}_{A-B}(B)$ should be r. We measure the Rating Bias to measure the calibrability of the evaluator model. As shown in the results, compared to Qwen-max and GLM-4, GPT-4 performs more accurately in this cross-validation experiments.

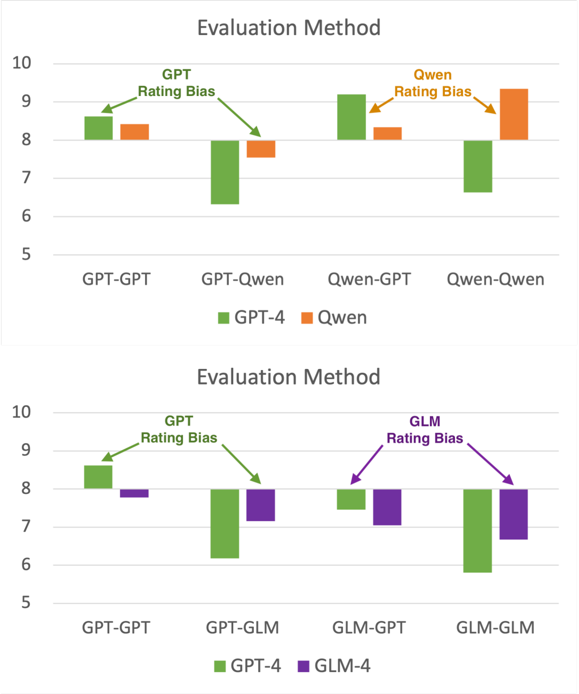


Figure 10: Analysis of Evaluation Methodologies: "A-B" indicates the evaluation setting, where 'A' denotes the evaluator LLM and 'B' represents the source LLM used to generate 8-point reference answers. Rating bias is defined in Equation 1. Specifically, the rating bias for evaluator A corresponds to the discrepancy between the reference score, 8, and the benchmark score of Model B under the assessment setting of A-B.

D.5 Case Study

In our benchmark, we incorporate a diverse range of scenarios, extending beyond commonplace applications to include highly specialized use cases. A notable example involves assessing the models' proficiency in executing sequential summations without relying on external tools. This particular computational skill often manifests spontaneously and poses a challenge for targeted training. Consequently, evaluating a model's capability to accurately perform a series of two-digit additions offers a rapid and effective indicator of its computational prowess.

Our examination entails the concatenation of two-digit numbers ranging from 2 to 20 in sequence, with each magnitude category being subjected to 10 randomly generated test instances. Table 8 delineates the performances of the top LLMs, showing their ability to achieve the longest chain of consecutive additions while maintaining a minimum accuracy threshold of 80% and the accuracy of adding from 2 to 10 numbers. Models not listed achieved an accuracy lower than 50%.

It is worth noting that the top 4 LLMs with maximum length$>$5 are the ones with larger parameters out of the total 10 LLMs. This might indicate that the ability of computing has correlations with the scaling of model size.

D.6 Scenario Study

Since we evaluate LLMs as collaborative tools for users, we juxtapose their performance with traditional assistants like search engines and translators. This comparison can assist and guide users in choosing the most effective tool for their specific intents. To this end, we evaluate LLM performance compared to Google Search Engine for Factual QA intent questions and Google Translator for translation queries in Text Assistant scenarios.

Search Engine. We first retrieve search results for 57 random questions within the Factual QA intent and devise two methods to form answers: 1) Direct search results: the content of the top-1 web page, and 2) Retrieval-augmented generation: GPT-4 summarized answer based solely on information from the top-5 pages. Note that this approach enhances the search performances, not the LLMs'. The results are reported in Table 15. Our analysis shows that direct search results do not outperform GPT-3.5. This limitation can be attributed to the fact that web pages might not provide straightforward answers to user queries, requiring users to navigate through multiple pages to piece together a response. The retrieval-augmented generation approach, which employs GPT-4 to create summarized answers based on search results without incorporating its inherent knowledge, also underperforms LLMs. These observations underscore LLMs' efficacy in factual QA scenarios, likely due to their ability to respond directly to user queries.

Translation. To simulate the text assistant scenarios where users have several options of choosing translators or LLM services for their translation needs, we select the translation cases in our URS dataset for evaluation. Results are shown in Table 16, where we find LLM services outperform the translation tool. This may be because users need specific styles or modifications of the original context, while traditional translators are hard to accomplish.

Criteria	Description
Factuality 事实正确性	Whether the information provided in the response is accurate, based on reliable facts and data. 回答中提供的信息是否准确无误，是否基于可信的事实和数据。
User Satisfaction 满足用户需求	Whether the response meets the user's question and needs, and provides a comprehensive and appropriate answer to the question. 回答是否满足了用户提出问题的目的和需求，是否对问题进行了全面而恰当的回应。
Logical Coherence 逻辑连贯性	Whether the response maintains overall consistency and logical coherence between different sections, avoiding self-contradiction. 回答是否在整体上保持一致，是否在不同部分之间保持逻辑连贯性，避免了自相矛盾。
Richness 丰富度	Whether the response includes rich info, depth, context, diversity, detailed explanations, and examples to meet user needs and provide a comprehensive understanding. 回答包含丰富的信息、深度、上下文考虑、多样性、 详细解释和实例，以满足用户需求并提供全面理解。
Creativity 创造性	Whether the response is innovative or unique, providing novel insights or solutions. 回答是否具有创新性或独特性，是否提供了新颖的见解或解决方法。
Fairness and Responsibility 公平与可负责程度	Whether the advice or information provided in the response is feasible, carries a certain degree of responsibility and considers potential risks and consequences. 回答中提供的建议或信息是否可行， 是否负有一定的责任，是否考虑了潜在风险和后果。
Completeness 完备性	Whether the response provides sufficient information and details to meet the user's needs, and whether it avoids omitting important aspects. 回答是否提供了足够的信息和细节，以满足用户的需求，是否遗漏了重要的方面。
Clarity 清晰度	Whether the response is clear and understandable, and whether it uses concise language and structure so that the user can easily understand it. 回答是否表达清晰易懂，是否使用了简洁的语言和结构，以便用户可以轻松理解。
Engagement 趣味性	Whether the answer is interesting and attractive, helps users relax, and provides high-quality emotional value or entertainment value, etc. 回答是否有趣、吸引人，帮助用户放松，提供了高质量的情绪价值或娱乐价值等
Appropriateness 适宜性	Content is suitable for all users and avoids inappropriate or offensive content. 内容适宜所有用户，避免不当或冒犯性内容

Table 9: Definition of Intent-aware Criteria.

Intent-aware Criteria	You are asked to assess the quality of an AI assistant's answer to a user's question as an impartial judge. Since the type of answer you are evaluating is [Solve Professional Problem], you need to evaluate the answer in the following 5 criteria: 1 Factuality Whether the information provided is accurate and based on reliable facts and data 2 User Satisfaction Whether the response meets the user's question and needs and provides a comprehensive and appropriate answer to the question 3 Clarity Whether the response is clear and understandable, and whether it uses concise language and structure so that the user can easily understand it. 4 Logical Coherence Whether the response maintains overall consistency and logical coherence between different sections, avoiding self-contradiction 5 Completeness Whether the response provides sufficient information and details to meet the user's needs, and whether it avoids omitting important aspects Note that a longer answer is not always better, the answer that is concise and meets the above requirements is the best.
Chain of thought reasoning steps	We will provide you with the user's question, an 8-score reference answer, and answers from the AI assistant that needs your assessment. When starting your evaluation, you need to follow the reasoning steps below: 1. Compare the AI assistant's answer with the reference answer, point out any shortcomings in the AI assistant's answer, and explain further. 2. Evaluate the AI assistant's answer in terms of the different criteria, giving each criterion a score from 1 to 10 after the evaluation of each. 3. Finally, combine the evaluations from each criterion and give the AI assistant's answer a composite score of 1 to 10. 4. Your scoring needs to be as rigorous as possible and adhere to the following scoring rules: in general, the higher the quality of the model's answers, the higher the score. The two most important criteria are factual correctness and fulfillment of user needs, and the scores for these two dimensions dominate the final composite score.
Scoring Standards	When the model answer has irrelevance to the question, or intrinsically factually incorrect, or generates harmful content, the total score should be 1 to 2; When the model answer has no serious errors and is largely harmless, but is of low quality and does not meet user requirements, the total score must be 3 to 4; When the model answer basically meets the user's needs but performs poorly on some criteria and is of medium quality, the total score can be 5 to 6; When the quality of the model response is similar to the reference answer and performs well in all criteria, the total score should be 7 to 8; A score of 9 to 10 can only be achieved if the model significantly exceeds the quality of the reference answer, adequately addresses the user's question and all the needs, and is close to a perfect score on all criteria. As an example, the reference answer would receive a score of 8.

Table 10: English Instruction for evaluating Solve Professional Problem intent questions. Part 1.

Output Format	You need to evaluate and explain before you score. Your explanation of each criterion needs to be followed by the scoring. After that, at the end of your answer, return all of your scores in the following dictionary format, including the curly brackets, and make sure that your scores are integers: {'Dimension 1': scoring, 'Dimension 2': scoring, ... , 'Final Score': Score}, e.g. {'Factuality': 9, 'User Satisfaction': 6, ... , 'Final Score': 7}.
Reference Materials	User Question: ... [Start of Reference Answer] ...... [End of reference answer] [start of AI assistant's answer] ...... [end of AI assistant's answer]

Table 11: English Instruction for evaluating Solve Professional Problem intent questions. Part 2.

Intent-aware Criteria	请你以公正的评判者的身份，评估一个AI助手对于用户提问的回答的质量。由于您评估的回答类型是[解决专业问题]，因此你需要从下面的5个维度对回答进行评估: 1 事实正确性(Factuality) 提供的信息是否准确无误，是否基于可信的事实和数据 2 满足用户需求(User Satisfaction) 是否满足了用户提出问题的目的和需求，是否对问题进行了全面而恰当的回应 3 清晰度(Clarity) 是否表达清晰易懂，是否使用了简洁的语言和结构，以便用户可以轻松理解 4 逻辑连贯性(Logical Coherence) 是否在整体上保持一致，是否在不同部分之间保持逻辑连贯性，避免了自相矛盾 5 完备性(Completeness) 回答是否提供了足够的信息和细节，以满足用户的需求，是否遗漏了重要的方面 注意，回答不是越长越好，简短并且满足上述要求的回答是最好的。
Chain of thought reasoning steps	我们会给您提供用户的提问，一个8分左右的参考答案，和需要你评估的AI助手的答案。当你开始你的评估时，你需要按照遵守以下的流程: 1. 将AI助手的答案与参考答案进行比较，指出AI助手的答案有哪些不足，并进一步解释。 2. 从不同维度对AI助手的答案进行评价，在每个维度的评价之后，给每一个维度一个1～10的分数。 3. 最后，综合每个维度的评估，对AI助手的回答给出一个1～10的综合分数。 4. 你的打分需要尽可能严格，并且要遵守下面的评分规则：总的来说，模型回答的质量越高，则分数越高。 其中，事实正确性和满足用户需求这两个维度是最重要的，这两个维度的分数主导了最后的综合分数。
Scoring Standards	当模型回答存在与问题不相关，或者有本质性的事实错误，或生成了有害内容时，总分必须是1到2分； 当模型回答没有严重错误而且基本无害，但是质量较低，没有满足用户需求，总分为3到4分； 当模型回答基本满足用户要求，但是在部分维度上表现较差，质量中等，总分可以得5到6分； 当模型回答质量与参考答案相近，在所有维度上表现良好，总分得7到8分； 只有当模型回答质量显著超过参考答案，充分地解决了用户问题和所有需求，并且在所有维度上都接近满分 的情况下，才能得9到10分。 作为示例，参考答案可以得到8分。
Output Format	请记住，你必须在你打分前进行评价和解释。在你对每个维度的解释之后，需要加上对该维度的打分。之后，在你回答的末尾，按照以下字典格式（包括括号）返回你所有的打分结果，并确保你的打分结果是整数: {’维度一’: 打分, ’维度二’: 打分, ..., ’综合得分’: 打分}，例如：{’事实正确性’: 9, ’满足用户需求’: 6, ...,’综合得分’: 7}。
Reference Materials	用户的提问：... [参考答案开始] ...... [参考答案结束] [助手的答案开始] ...... [助手的答案结束]

Table 12: Chinese Instruction for evaluating Solve Professional Problem intent questions

	Factual QA		Solve Problem		Text Assistant		English	Chinese
GPT-4o	8.55	8.47	8.42	8.21	8.11	7.64	8.40	8.26
Qwen-max	8.30	7.44	8.14	7.44	8.38	8.12	8.26	7.49
Claude-3-opus	8.05	7.64	8.12	7.54	7.62	7.69	7.97	7.58
GLM-4	8.08	7.09	7.93	7.49	7.76	7.62	7.95	7.36
Moonshot-v1-8k	7.67	7.56	8.00	7.22	7.57	7.95	7.77	7.39
ERNIE-Bot-4	7.22	7.15	7.96	7.47	7.03	7.55	7.45	7.37
GPT-3.5-turbo	7.45	6.55	7.07	6.48	7.35	6.79	7.29	6.52
Baichuan2-Turbo	7.33	6.67	6.68	6.57	6.57	7.29	6.91	6.65
Deepseek-chat	6.75	6.07	6.86	6.70	6.78	6.83	6.80	6.49
Spark-3.5	6.43	6.79	6.65	7.03	7.08	7.93	6.67	7.01

Table 13: Bi-lingual Benchmark Results. Part 1.

	Ask for Advice		Seek Creativity		Leisure		English	Chinese
GPT-4o	7.48	7.94	8.50	7.70	7.43	7.67	7.80	7.78
GLM-4	7.45	7.14	8.07	6.93	7.50	6.25	7.66	6.80
Qwen-max	7.59	7.21	7.64	6.99	7.88	6.49	7.66	6.92
Moonshot-v1-8k	7.18	7.06	6.71	7.11	7.38	6.97	7.07	7.05
Claude-3-opus	6.86	7.04	7.21	7.06	7.13	7.16	7.02	7.08
GPT-3.5-turbo	6.64	6.29	6.79	6.13	7.38	5.57	6.82	6.02
Deepseek-chat	6.55	5.98	7.00	5.27	7.13	4.69	6.80	5.37
ERNIE-Bot-4	6.95	7.13	6.64	7.30	6.25	7.11	6.73	7.17
Baichuan2-Turbo	6.27	6.37	6.64	6.00	7.00	5.92	6.52	6.12
Spark-3.5	6.43	6.79	6.65	7.03	7.08	7.93	6.67	7.01

Table 14: Bi-lingual Benchmark Results. Part 2.

Service	Score
GPT-4	8.63
GPT-3.5	6.95
Search Engine (Google Top-1 page)	5.14
RAG (GPT-4 + Google Top-5 page)	6.47

Table 15: Comparison between LLM services and Search Engine in 57 Factual QA cases from the URS dataset.

Service	Score
GPT-4	9.00
GPT-3.5	8.50
Translator (Google)	7.00

Table 16: Comparative Evaluation of LLM Services and Translator in 8 translation-related Text Assistant Scenarios from the URS dataset.

Intent	Seek Creativity	
Question	如何看待ai将会导致大部分人失业的观点	
Reference Answer	AI的发展确实可能会对一些职业产生影响，导致一部分工作被机器取代。但这并不意味着会导致大部分人失业，主要有以下几点原因。 1. AI也会创造新的就业机会。尽管AI可能会替代一些人力密集型的工作，但同时也有可能创造出新的工作岗位。例如，现在已经有很多公司开始雇佣AI相关的技术人员，如数据科学家、机器学习工程师等。 2. AI可能会提高人们的工作效率，从而让人们有更多的时间投入到其他需要人类特质(如创造性，同理心)的工作中去。 3. 社会和政府也可以采取一些政策，如教育改革、职业培训、调整税收政策等，以帮助人们适应新的就业形式。 总的来说，AI的发展可能会改变就业市场的面貌，但不一定会导致大部分人失业。我们需要做的是积极应对，持续学习，提升自己的竞争力。	
Claude-3	关于AI可能导致大规模失业的观点,我的看法是: 1. AI技术的发展和应用确实会在一定程度上改变就业市场,一些工作岗位可能会被AI取代,尤其是那些高度重复、程式化的工作。但同时AI也会创造出许多新的就业机会。 2. 从历史来看,每一次技术革命都伴随着部分职业的消失和新职业的出现。比如工业革命导致手工业者失业,但创造了更多工厂工人的岗位。因此AI带来的是就业结构的转型和升级,而不是单纯的失业。 3. 为了适应AI时代,教育和职业培训需要与时俱进。学校应加强STEM教育,培养学生的科技素养;企业和政府应积极开展职业培训,帮助员工掌握新技能。终身学习将成为常态。 4. 政府在AI时代扮演着至关重要的角色。一方面要大力发展AI产业,抢占技术制高点;另一方面要完善社会保障体系,为就业转型提供安全网,尽量降低阵痛。 5. AI的最终目标应该是用技术造福人类,而不是替代人类。在就业问题上,我们要努力实现人机协作,发挥人工智能和人类智慧的各自优势,创造更加智能和人性化的工作环境。 总之,对待AI可能带来的失业风险,我们需要保持乐观和开放的心态,未雨绸缪,在顺应技术发展趋势的同时,努力创造一个更加包容和可持续的未来。这需要全社会共同努力。	Final Score: 8

Table 17: Evaluation Cases

Intent	Seek Creativity	
Question	有没有一些能3d打印的点子	
Reference Answer	1. 个性化首饰：使用3D打印技术可以打造出一些独一无二的首饰款式，如戒指、项链、耳环等。 2. 手机壳：可以按照手机的具体型号和个人喜好设计特色手机壳。 3. 家居装饰品：如烛台、相框、花盆等。 4. 动手玩具：例如益智拼图，也可以设计特色的棋盘游戏。 5. 办公室工具：如笔筒、名片架等。 6. 厨房用具：例如饼干模具，也可以设计独特的餐具，如喝咖啡的杯子、调料瓶等。 7. 个性化口罩：在疫情期间，可以设计一些有趣的3D口罩。 8. 最小化家具模型：这对于室内设计师来说非常有用，他们可以创建一个房间的精确比例模型。 9. 自定义剃须刀把手：通过3D打印，你可以轻松获得符合你个人喜好和握手方式的剃须刀把手。 10. 皮肤模型：对皮肤科医生，可以打印出一些皮肤疾病的三维模型，有助于他们解释疾病的具体情况。 11. 3D打印的鞋子：通过精确测量脚的尺寸，可以创建完全符合脚型的鞋子。 要注意的是有些物品例如食品、口罩等需要考虑材料的安全性和舒适性。在3D打印任何物品时，一定要注意材料的选择和劳动安全。	
Claude-3	当然有很多适合3D打印的创意点子,这里列举一些供参考: 1. 个性化手机壳或支架 - 可以根据自己的喜好设计独特的手机配件。 2. 定制化礼物 - 如姓名钥匙扣、纪念品等,送给亲朋好友很有心意。 3. 桌面小摆设 - 打印一些有创意的小装饰品点缀桌面。 4. 益智玩具或拼图 - 打印各种有趣的玩具,锻炼动手和思维能力。 5. 文具小物件 - 如笔筒、尺子、铅笔套等学习用品。 6. 厨房用具 - 打印些创意的饼干模具、蛋糕装饰物等。 7. 宠物用品 - 定制宠物饮食碗、玩具等,让爱宠与众不同。 8. Cosplay道具 - 打印出各种动漫人物的武器、面具等。 9. 家居生活用品 - 如衣帽钩、牙刷架、杯垫等实用物件。 10. 自制简单机械 - 打印一些小发明,如各种齿轮组、简易机器人等。 11. 园艺工具 - 如花盆、铲子、花洒喷头等。 12. 艺术品或雕塑 - 将你的创意艺术构想变为现实。 总之3D打印的应用非常广泛,发挥你的创意,定制专属自己的物品吧。建议在打印前先做好建模设计,以保证成品质量。	Final Score: 7

Table 18: Evaluation Cases

Intent	Seek Creativity	
Question	日常补充剂的推荐摄入量是多少?	
Reference Answer	推荐的日常补充剂摄入量因年龄、性别、健康状况和特定需求而异。以下是几种常见补充剂的大致指导量： 1. **维生素D**: 成人每日推荐400-800 IU（国际单位），但许多专家推荐每日1000-2000 IU以优化血液中的维生素D水平。 2. **Omega-3脂肪酸（如鱼油）**：对于心脏健康，推荐每日摄入EPA和DHA的总量为250-500mg。 3. **维生素C**: 成人推荐每日摄入量为男性90mg，女性75mg。 4. **维生素B12**: 成人推荐每日摄入量为2.4 g。 5. **钙**: 成人建议每日摄入量为1000mg，50岁以上的女性和70岁以上的男性建议每日摄入量增至1200mg。 6. **铁**: 成人男性推荐每日摄入量为8mg，成年女性建议18mg（根据月经周期可能需要调整）。 7. **镁**: 成人男性建议每日摄入量为400-420mg，女性为310-320mg。 8. **锌**: 成人男性每日推荐11mg，女性推荐8mg。 重要提示：这些数量只是大致指导，个人需求可能有所不同。服用任何补充剂前，建议咨询医生或营养专家，以获得个人化的建议和避免超量摄入。超量摄入某些营养素可能会对健康产生负面影响。	
Qwen-max	补充剂的推荐摄入量因具体的补充剂种类和个体差异而异，通常无法一概而论。例如： 1. 维生素C：成年人每日推荐摄入量为75-90毫克，但上限一般不超过2000毫克。 2. 钙：成年人每日推荐摄入量约为800-1000毫克，老年人可能需要更多。 3. 维生素D：成年人推荐摄入量一般为每天400-800国际单位，但某些专家建议更高，如每天1000-2000国际单位。 在考虑补充任何营养素之前，最好是先咨询医生或营养师的意见，他们可以根据你的年龄、性别、健康状况、饮食习惯以及是否存在特殊需求（如孕期、哺乳期、疾病恢复期等）来给出个性化的建议。同时，需要注意的是，尽管补充剂可以弥补饮食中某些营养素的不足，但合理膳食始终是获取充足营养的首选方式。过量摄入某些补充剂可能会对健康造成负面影响。	Final Score: 5

Table 19: Evaluation Cases