

# SenseJudge: Human-Centric Preference-Driven Judgment Framework

Rui Li<sup>1\*</sup>, Junfeng Liu<sup>2,3\*</sup>, Xiangwen Kong<sup>2</sup>, Zhifang Sui<sup>1†</sup>

<sup>1</sup>State Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University <sup>2</sup>StepFun <sup>3</sup>Xi'an Jiaotong University  
o\_l1ru1@outlook.com, 31243360791jf@stu.xjtu.edu.cn

## Abstract

Using Large Language Models (LLMs) as judges across scenarios such as model-response assessment is becoming an increasingly accepted paradigm. However, existing judgment approaches often rely on trained judges using fixed preference data, which tend to overlook diverse user preferences and struggle to adapt to real-world human-AI dialogue scenarios. To address these limitations, we propose *SenseJudge*, a customizable judgment framework driven by human preferences, and *SenseBench*, a diverse and challenging instruction-following benchmark derived from real-world multi-turn interactions. We evaluate the framework and benchmark on two tasks: 1) *LLMs as personalized judges*, and 2) *model ranking*. Extensive experiments show that the SenseJudge framework outperforms the compared judgment methods and models in the LLMs-as-personalized-judges task and produces model rankings that align with human judgments. Additionally, we conducted analyses on position bias and consistency, alongside ablation studies, which support the robustness of SenseJudge. Our data and code are publicly available in the [GitHub repository](#).

## 1 Introduction

The evaluation of model responses is a crucial component in the development of Large Language Models (LLMs). Selecting the superior response from a pair generates preference data (query, chosen, rejected), which is essential for various post-optimization algorithms (Rafailov et al., 2024; Ouyang et al., 2022). Competitive assessments of different LLM versions across various companies facilitate model updates and iterations (Lambert et al., 2024; Li et al., 2024).

\*Equal contribution. This work was conducted during an internship at StepFun AI from August 2024 to April 2025.

†Corresponding author.

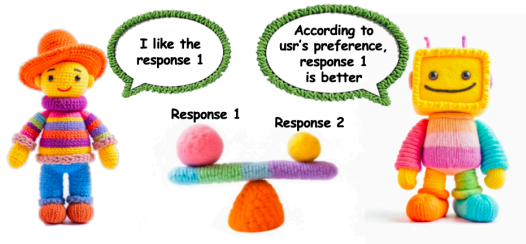


Figure 1: The model judges responses on behalf of human based on individual preferences.

Furthermore, by evaluating a variety of model responses and selecting those that most closely align with individual user preferences, the judgment process can accommodate a broad spectrum of user values, effectively delivering a personalized experience. (Fan et al., 2025; Dong et al., 2024).

Previously, these judgments were made through manual human annotation. However, with advancements in the natural language understanding and instruction-following capabilities of Large Language Models (LLMs) (Naveed et al., 2024; Zhou et al., 2023), LLMs are increasingly being integrated into human workflows for automatic judgment (Zheng et al., 2023a; Wang et al., 2023b).

However, existing evaluation methods and models, such as PandaLM (Wang et al., 2023b), AutoJ (Li et al., 2023), and various reward models, rely on homogeneous standards and values, overlooking the diversity of user preferences. These models also lack the ability to make nuanced choices based on individual user needs. Furthermore, current public judgment benchmarks, including MT-Bench (Zheng et al., 2023a), AutoJ (Li et al., 2023), and RewardBench (Lambert et al., 2024), predominantly feature single- or double-turn dialogue designs. This represents

a significant departure from real-world human-AI interactions, which often involve multiple rounds of historical dialogue, some relevant and others not, to inform the current request.

To address these limitations, we propose *SenseJudge* and *SenseBench*. *SenseJudge* is a customizable evaluation framework based on explicit human preferences, enabling personalized judgment of model responses by simulating the preferences of different users and scenarios. *SenseBench* is a high-quality, challenging dataset derived from realistic human-AI dialogues, covering eight common topics, including math, logic, writing, and role-play.

We conducted extensive experiments to evaluate the effectiveness of the SenseJudge framework in two primary application scenarios based on SenseBench: 1. LLM-as-a-Personalized-Judge, which focuses on selecting responses that align with individual preferences, and 2. Model Ranking, which aims to identify the leading model among different companies or versions (e.g., GPT vs. DeepSeek). The results demonstrate that SenseJudge outperforms the compared state-of-the-art APIs, training-based evaluators, and reward models in the LLM-as-a-Personalized-Judge task. It also achieves model rankings that align with human judgment.

Additionally, the effectiveness and robustness of SenseJudge are further validated through analyses of factors such as positional bias and consistency, as well as comprehensive ablation studies.

Our main contributions are as follows:

- We developed SenseBench, a benchmark designed to align with realistic human-AI interaction scenarios. It features diverse and challenging multi-turn dialogues.
- We propose SenseJudge, a judgment framework designed to adapt to diverse human values and preferences, which enables fine-grained and customizable judgment of model responses.
- We conducted comprehensive experiments and analysis to evaluate the effectiveness of SenseJudge in the LLMs-as-personalized-judges task and model ranking task.

## 2 Related Work

### 2.1 LLM as a Judge

LLMs are increasingly recognized as a viable alternative to traditional expert-driven evaluations due to their scalability and adaptability in evaluation tasks. Common enhancement methods include designing prompts and training judges. Prompt engineering is an effective approach to enhancing the performance of LLMs as judges without updating their model parameters. Many methods optimize their respective prompt strategies prior to training, such as incorporating references (Zhu et al., 2023) or rules (Li et al., 2023) into the prompts, and directing models to offer detailed analyses or justifications (Kim et al., 2024). Based on this, pairwise or point-wise judgment data can be used to train LLM-based judges by employing supervised fine-tuning (Kim et al., 2024). Some researchers train reward models using large-scale preference feedback data; leveraging their ability to learn preferences, these models can also function as judges (Zhong et al., 2025).

### 2.2 From Universal to Individualized

The transition from generalization to personalization represents a historic paradigm in technological development (You et al., 2025). With the rapid advancement of the general capabilities of LLMs, there is growing interest in tailoring LLMs to specific user contexts, moving beyond their role as merely general-purpose chatbots (Zhang et al., 2024). For instance, Wu et al. (2024) proposed to implicitly infer users’ personalized preferences through multi-turn conversations and help LLMs to dynamically adapt their behavior and responses to better align with individual user needs. Zhu et al. (2025) proposed to use psychological assessment tools, such as the “Big Five Personality Traits” and the “Dark Triad” to quantify user characteristics and align the models accordingly. Li et al. (2025) introduced a multi-granularity interest prediction framework which leverages both coarse-grained and fine-grained interest information to supervise model training. Additionally, Tan et al. (2025) proposed a personalized framework OPPU, which enhances the personalization and generalization capabilities of LLMs by creating dedicated personalized modules for each user through parameter-efficient fine-tuning and non-parametric personalization methods, such as retrieval-augmented and profile-augmented tech-

niques.

### 3 Method

In this section, we introduce the two key components of our framework: **SenseBench**, a diverse and high-quality instruction-following benchmark derived from real human-AI interaction dialogues, and **SenseJudge**, a customizable evaluation method based on human preferences.

#### 3.1 SenseBench

##### 3.1.1 Overview

Currently, there is a lack of benchmarks that accurately reflect human user experiences and differentiate models based on perceived quality. SenseBench consists of anonymized real user queries sourced from an online LLM service, where the final turn of each dialogue either provides context-dependent instructions or introduces entirely new tasks. This design places high demands on the LLMs’ ability to align with human expectations, facilitating effective model differentiation and improving alignment with human-perceived quality.

##### 3.1.2 Benchmark Construction

Based on real user data, we employ a scalable automated data selection pipeline to construct a high-quality and diverse benchmark as shown in Figure 2. The pipeline consists of two main steps: 1) Quality-based Data Filtering and Categorization, and 2) Challenge-based Filtering. The detailed process is as follows:

**Quality-Based Data Filtering and Categorization** We use Qwen3-14B(Yang et al., 2025) to denoise and categorize data from real user dialogue history. Specifically, we first remove redundant, ambiguous, low-quality, or overly simplistic entries and then categorize the remaining data into eight distinct topics: math, logical reasoning, comprehensive question answering, coding, creative writing, multilingual translation, role-playing, and natural language understanding. These categories represent common types of real-world human instructions and provide a comprehensive reflection of the capabilities of LLMs.

**Challenge-based Filtering** To ensure that the selected prompts are sufficiently challenging, we require that they generate clearly distinguishable responses from different models. We conduct a

multi-model evaluation on candidates that have passed an initial round of automated filtering. Specifically, we input each candidate prompt into both relatively stronger models (e.g., Qwen2.5-Max, Qwen3-235B) and relatively weaker models (e.g., Qwen2.5-14B, LLaMA3-8B) to generate responses. This results in a large set of pairwise data in the format of (Query, Strong Model Response, Weak Model Response). We then apply a two-stage filtering process that combines both automatic and manual filtering.

**Automatic Filtering.** We use GPT-4 to assess the response quality between strong and weak models. Questions where the strong model significantly outperforms the weak model are retained as discriminative and challenging instances. For specific judging instructions, see Appendix A.2. **Manual Filtering.** After the initial automatic filtering, we performed manual filtering on the data. Specifically, the manual filtering process focuses on two main aspects: 1) Assessing whether the pairwise data filtered by GPT-4 can truly distinguish model capabilities. During this assessment, humans consider visible differences between model responses (e.g., format, length, or level of detail) or other intuitive quality distinctions. 2) Performing partial cleaning on the prompts of the selected data. Since all prompts are sourced from real user data, they may contain formatting issues or improper phrasing, which require manual cleaning to ensure the data is suitable for creating a usable benchmark.

Specific data statistics and category information can be found in Appendix A.1.

#### 3.2 SenseJudge

##### 3.2.1 Overview

SenseBench offers a human-centric environment where we employ multi-model generation followed by pairwise judgment to support two tasks: (1) user-centric preference alignment and (2) model ranking. We introduce SenseJudge, a personalized pairwise judgment approach that demonstrates strong performance in both tasks. SenseJudge begins by extracting a spectrum of human preferences from a small set of annotated (query, chosen, rejected) pairs. Each preference guides the judge through the same pairwise instances, and the preference set that best matches the human labels is retained as the distilled, generalizable set. Next, we outline the components of

## SenseBench Automatic Building Process

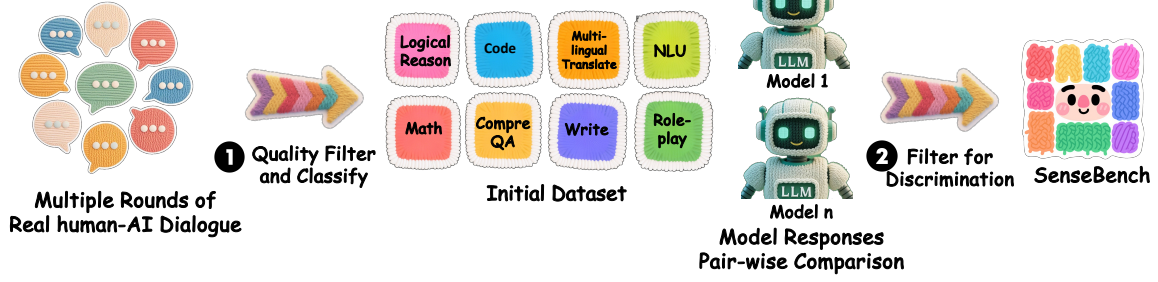


Figure 2: An overview of the data construction pipeline of SenseBench. The pipeline involves 1) Quality-Based Data Filtering and Categorization and 2) Challenge-based Filtering.

our judgment framework, including input-output format, preference generation, and the selection and application of the preference set.

### 3.2.2 Input-Output Format

For SenseJudge, the input  $I = \{q, (r_1, r_2), p\}$  comprises a query  $q$ , a pair of responses  $(r_1, r_2)$ , and a preference  $p$ . The target output  $O = \{\text{judgment}, \text{analysis}\}$  yields a judgment and corresponding analysis, specifically indicating whether  $r_1$  or  $r_2$  is superior according to the given preference  $p$ . Based on previous research (Dong et al., 2024; Zheng et al., 2023b), the inclusion of “ties” can negatively impact model performance, so we assume that the model can always distinguish between responses based on preference.

### 3.2.3 Preference Construction

We use DeepSeek-R1 (Guo et al., 2025) to distill implicit human preferences captured in annotated pairwise data into explicit textual preferences. Specifically, we provide a small set of human-annotated pairwise examples to an LLM and prompt it to generate a preference statement explaining why a user might favor one response over the other. Examples of these preferences are provided in Appendix A.3. One preference is generated for each pairwise sample. These preferences are then incorporated into the data to create a complete personalized evaluation dataset in the format  $I = \{q, (r_1, r_2), p\}$ , which consists of a query  $q$ , a pair of responses  $(r_1, r_2)$ , and a preference  $p$ . The specific prompt used is detailed in Appendix A.2.

### 3.2.4 Preference Set Selection

After we obtain the preference set  $P = \{p_1, p_2, \dots, p_m\}$ , where  $m$  is the number of sampled examples, we feed each  $p_i \in P$  into

the judge and assess its performance on the test set  $T = \{(q_1, r_{11}, r_{12}), \dots, (q_m, r_{m1}, r_{m2})\}$ . Note that the set here is the same one we used to extract the preference. Under each  $p_i$ , we obtain the judge’s selections  $C_i = \{c_{i1}, c_{i2}, \dots, c_{im}\}$ , where  $c_{ij} \in \{r_{j1}, r_{j2}\}$  indicates that the response  $r_{j1}$  or  $r_{j2}$  is chosen by the judge. Comparing the model’s selections  $C_{\mathcal{M},i} = \{c_{i1}, c_{i2}, \dots, c_{im}\}$  with the ground-truth human judgments  $G = \{g_1, g_2, \dots, g_m\}$ , where  $g_j \in \{r_{j1}, r_{j2}\}$ , we obtain a corresponding set of correctness indicators  $S_i = \{s_{i1}, s_{i2}, \dots, s_{im}\}$ , where  $s_{ij} \in \{0, 1\}$ . Subsequently, we consider all subsets  $\mathcal{P}_k \subseteq P$  of cardinality  $l$ , where the number of preferences  $l$  satisfies  $1 \leq l \leq m$ . For each subset  $\mathcal{P}_k$ , we perform a majority vote over its associated correctness indicators  $\{S_{\mathcal{P}_{k1}}, S_{\mathcal{P}_{k2}}, \dots, S_{\mathcal{P}_{kl}}\}$ , where  $\mathcal{P}_{ki}$  denotes the  $i$ -th preference in  $\mathcal{P}_k$  and  $S_{\mathcal{P}_{ki}}$  is its corresponding correctness indicator. This procedure yields an updated set of correctness indicators, denoted by  $S_{\mathcal{P}_k}^* = \{s_{\mathcal{P}_{k1}}^*, s_{\mathcal{P}_{k2}}^*, \dots, s_{\mathcal{P}_{kl}}^*\}$ , where each  $s_{\mathcal{P}_{kj}}^*$  is obtained via a majority-voting scheme applied to the collection  $\{s_{k1,j}, s_{k2,j}, \dots, s_{kl,j}\}$ . We then calculate the accuracy of these aggregated correctness indicators  $\text{Acc}(S_{\mathcal{P}_k}^*) = \frac{1}{m} \sum_{j=1}^m \mathbb{I}(s_{\mathcal{P}_{kj}}^* = 1)$ ,  $\mathbb{I}$  is an indicator function. The subset  $\mathcal{P}_k^*$  that achieves the highest accuracy  $\text{Acc}(\mathcal{P}_k^*)$  effectively captures the annotator’s underlying preference and can be generalized to unseen examples.

### 3.2.5 Preference Set Application

Given the optimal preference subset  $\mathcal{P}_k^*$  obtained from the previous stage, we apply it to an unseen test set  $T' = \{(q'_1, r'_{11}, r'_{12}), \dots, (q'_n, r'_{n1}, r'_{n2})\}$ , where  $n$  denotes the number of test instances. For each preference  $p \in \mathcal{P}_k^*$ , every test sam-

ple  $(q'_j, r'_{j1}, r'_{j2})$  is fed into the judge conditioned on  $p$ , producing a set of model selections  $C'_p = \{c'_{p1}, c'_{p2}, \dots, c'_{pn}\}$ , where  $c'_{pj} \in \{r'_{j1}, r'_{j2}\}$ . We then aggregate the outputs across all preferences in  $\mathcal{P}_k^*$  by performing a majority vote at the instance level, yielding the final decision

$$c_j^* = \text{MajorityVote}(\{c'_{p1j}, c'_{p2j}, \dots, c'_{plj}\}),$$

where  $l = |\mathcal{P}_k^*|$ . This produces the final aggregated decision set  $C^{l*} = \{c_1^*, c_2^*, \dots, c_n^*\}$ , which represents the model's preference-aligned selections on the test set.

## 4 Experiments

In this section, we present the performance of SenseJudge in two primary tasks: 1) LLM-as-a-Personalized-Judge and 2) model ranking.

The effectiveness and robustness of SenseJudge are validated through analyses of factors such as positional bias and consistency, as well as comprehensive ablation studies.

### 4.1 Experiments Setup

#### 4.1.1 Baselines

**LLM+Prompt** Direct prompting without preferences across multiple closed-source and open-source models. The closed-source models included GPT-4o and Qwen-Plus-Latest. The open-source models tested were DeepSeek-V3 (DeepSeek-AI et al., 2025b), DeepSeek-R1 (DeepSeek-AI et al., 2025a), Qwen2.5-14B-Instruct (Qwen et al., 2025), Qwen2.5-72B-Instruct, Llama3.1-8B-Instruct (Grattafiori et al., 2024), and Qwen3-14B-Instruct.

**LLM+Training** We compare against publicly available trained judges such as AutoJ (Li et al., 2023), as well as top-ranking reward models on the RewardBench leaderboard, including the generative model Skywork-Critic-Llama-3.1-70B and classifier models including INF-ORM-Llama3.1-70B, Skywork-Reward-Gemma-2-27B-v0.2, and QRM-Gemma-2-27B. PandaLM (Wang et al., 2023b) and JudgeLM (Zhu et al., 2023) are excluded from the comparison due to token limit constraints.

#### 4.1.2 Datasets and Metrics

We use the human-annotated personalized pairwise dataset for the LLM-as-a-Personalized-Judge task, consisting of 1,000 examples (125

per category across 8 categories). We selected 12 widely used or state-of-the-art LLMs (e.g., DeepSeek-R1, Qwen2.5-Max, Grok-3, GPT-3.5, Qwen2.5-7B, Claude-3-Sonnet, Gemini-2.0-Flash-Thinking-Exp, DeepSeek-V3 (Liu et al., 2024), GLM-4-Plus, Gemini-1.0-Pro, Moonshot-v1-8k (Team et al., 2025), and Minimax-5.5) to generate responses on SenseBench. From each category, we sample 10 examples for preference generation and subsequent steps. Three annotators with diverse professional backgrounds provide ground-truth labels, which we use to measure accuracy. For the task of Model Ranking, we directly used the relative rankings of models on the human preference leaderboard as the ground truth. We refer to this dataset as the Ranking Set. We employed specific model comparison examples from the Ranking Set and evaluated the judge's performance by computing the consistency between the rankings produced by the judge and the human-annotated ground-truth rankings. Furthermore, we also evaluate on the public preference benchmark RewardBench (Lambert et al., 2024). Response accuracy was used as the evaluation metric. When assessing the correctness of the model's judgment, we employed a strict matching strategy, whereby the target string (e.g., the final decision is response A) must be fully contained within the model's response to be recorded as correct.

#### 4.1.3 Implementation Details

We adopt the (query, chosen, rejected) format commonly used in pairwise preference data. We test and calculate the overall accuracy by evaluating the chosen response and rejected response in both their original order and their reversed order. We directly employ small-scale LLMs as our judges without any additional training to balance performance and cost.

For the model ranking task, we randomly sample five preferences from each category within the preference set corresponding to individual annotators (three annotators in total). Each sampled preference is independently applied to the model ranking task using the SenseJudge framework. The final decision for each comparison is obtained via a majority voting scheme over the predictions induced by these preferences.

Model	Math	Code	Logic	QA	Write	Role	NLU	Trans	Overall
<b>Strong Open/Closed-Source Models</b>									
<b>GPT-4o</b>	66.00	61.60	65.47	72.93	60.80	63.20	65.47	56.40	63.98
<b>DeepSeek-V3</b>	72.80	62.27	66.67	77.07	62.67	64.40	64.80	61.87	66.57
<b>DeepSeek-R1</b>	53.33	55.60	61.33	53.47	46.93	50.53	51.73	50.67	52.95
<b>Qwen-Plus-Latest</b>	66.00	59.60	61.20	67.20	55.47	58.13	60.40	59.33	60.92
<b>Trained Judge / Reward Models</b>									
<b>Autoj-13B</b>	64.40	52.67	55.73	53.87	39.33	42.40	48.00	36.40	49.10
<b>Skywork-Critic-Llama3.1-70B</b>	63.73	63.87	63.73	69.87	59.67	59.47	57.33	55.47	61.35
<b>INF-ORM-Llama3.1-70B</b>	67.20	63.20	64.80	69.07	60.00	54.13	58.40	59.73	62.07
<b>QRM-Gemma-2-27B</b>	69.60	58.93	63.47	54.13	52.53	50.40	56.53	59.73	58.17
<b>Skywork-Reward-Gemma2-27B</b>	70.40	61.60	66.10	74.10	64.00	60.00	62.70	58.40	64.70
<b>SenseJudge</b>									
<b>Qwen2.5-14B-Instruct</b>	60.27	60.80	59.33	70.13	61.60	62.00	64.40	57.20	61.97
<b>+SenseJudge</b>	68.86(+8.59)	76.27(+15.47)	73.20(+13.87)	76.52(+6.39)	68.00(+6.40)	67.26(+5.26)	72.67(+8.27)	63.07(+5.87)	70.73(+8.76)
<b>Qwen2.5-72B-Instruct</b>	63.87	61.47	61.87	76.27	60.13	66.40	64.40	57.07	63.93
<b>+SenseJudge</b>	73.47(+9.60)	68.64(+7.17)	72.09(+10.22)	78.67(+2.40)	65.11(+4.98)	69.16(+2.76)	69.60(+5.20)	73.00(+15.93)	71.22(+7.29)
<b>Llama3.1-8B-Instruct</b>	62.13	58.80	59.73	72.00	58.13	58.53	60.13	53.60	60.33
<b>+SenseJudge</b>	74.02(+11.89)	64.90(+6.10)	73.00(+13.27)	76.46(+4.46)	65.96(+7.83)	64.79(+6.26)	67.51(+7.38)	56.93(+3.33)	67.94(+7.61)
<b>Qwen3-14B-Instruct</b>	71.73	64.80	69.20	78.00	61.47	67.60	66.27	56.13	66.90
<b>+SenseJudge</b>	71.39(-0.34)	67.24(+2.44)	72.13(+2.93)	77.95(-0.05)	73.73(+12.26)	71.07(+3.47)	69.41(+3.14)	79.07(+22.94)	72.75(+5.85)

Table 1: Comparative evaluation of various judgment methods and judges across different categories: Math, Code, Logic, QA (Question Answering), Write, Role (Role-playing), NLU (Natural Language Understanding), Trans (Translation), and Overall their average performance.

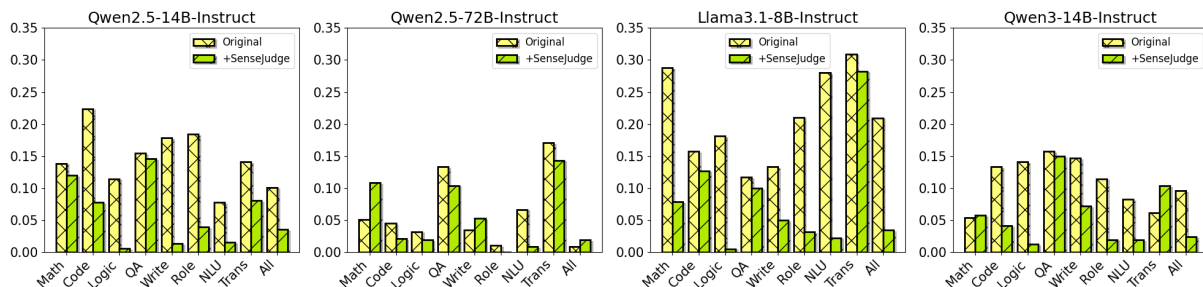


Figure 3: The ■ bar represents the absolute values of selecting response A (in the first place) and response B (in the second place) using the original model, while the ■ bar represents the absolute values after applying SenseJudge.

## 4.2 Main Result

### 4.2.1 LLM-as-a-Personalized-Judge

**Overview** In the development of business LLMs, real-time and representative human feedback is essential before the model is officially deployed.

Many benchmarks are available to directly evaluate models’ general capabilities through straightforward ground-truth answer matching. However, this evaluation often solely considers the accuracy of model outputs, overlooking other valued aspects such as creativity and formatting, thus failing to provide a comprehensive and user-centered quality assessment. Additionally, as many benchmark ground-truth answers may have been inadvertently or deliberately included in training data, it is difficult to assess the true performance of models using existing benchmarks alone. Hiring a stable team to evaluate models on consistent datasets after each update is a common prac-

tice. However, as model updates become more frequent, the repetitive and labor-intensive task of human annotation is increasingly constrained by cost and efficiency. Therefore, “LLM-as-a-Personalized-Judge” has become a highly promising task that involves utilizing LLMs to simulate real user judgments of model responses, thereby streamlining and automating the model development process.

**Comparison with Baselines** As shown in Table 1, the SenseJudge framework achieves an average Overall score of 70.66 in the LLM-as-a-Personalized-Judge task, outperforming baselines including strong open-source and closed-source models, trained judges, and reward models by 10.68 percentage points on average. Compared with the corresponding base models, SenseJudge yields an average gain of 7.37 percentage points across all evaluation categories and an average Overall gain of 7.38 percentage points. Even when

applied to small-scale LLMs like 8B and 14B, which are not specifically trained as judges, it improves the Overall score by 7.41 percentage points on average. Additionally, SenseJudge demonstrates consistent Overall improvements and broad category-level gains compared to the base models, with only minor decreases for Qwen3-14B-Instruct on Math and QA.

Although reward models such as INF-ORM-Llama3.1-70B and QRM-Gemma-2-27B have excelled on the public RewardBench leaderboard, their accuracy on the personalized dataset remains below 65% and even falls short compared to some strong baseline LLMs. Autoj-13B performs even worse, achieving scores similar to random selection. This indicates that while judges designed for specific judgment tasks and reward models developed for model alignment are trained on extensive preference data, they often learn fixed preferences that struggle to generalize to diverse and complex real-world scenarios and adapt to varying user preferences.

**Positional Bias** Previous studies (Wang et al., 2023a) have demonstrated that the relative position of two responses is an element that theoretically should be irrelevant but significantly influences the judgment result in the pairwise judgment setting. Some LLMs exhibited a significant tendency to select responses in the first or second positions, which we refer to as *position bias*. To evaluate the impact of SenseJudge on position bias, we counted the distribution of model selection across each position under identical ground-truth conditions and calculated the absolute values of these selections. Higher absolute values indicate a stronger positional bias. As shown in Figure 3, the SenseJudge framework mitigates positional bias, particularly for smaller models where this bias is often more pronounced.

Model	Original	SenseJudge
<b>Qwen2.5-14B-Instruct</b>	69.97	<b>74.17</b>
<b>Qwen2.5-72B-Instruct</b>	78.86	78.79
<b>Llama3.1-8B-Instruct</b>	60.36	<b>68.19</b>
<b>Qwen3-14B-Instruct</b>	81.23	<b>81.30</b>

Table 2: Consistency performance of original models and models with SenseJudge framework.

**Consistency** We examined the consistency of the SenseJudge framework under two distinct in-

put settings: (query, chosen, rejected) and (query, rejected, chosen). We considered the judge’s selection to be consistent if it identified the same response (either the initially designated ‘chosen’ or ‘rejected’) as superior across both input configurations. Results are presented in Table 2. Notably, the SenseJudge framework shows improvements over the base models such as Qwen2.5-14B-Instruct and Llama3.1-8B-Instruct, achieving increases in consistency scores of 4.2% and 7.83%, respectively, compared to their original consistency scores of 69.97% and 60.36%. While the Qwen2.5-72B-Instruct and Qwen3-14B-Instruct models exhibit high consistency in their original evaluations (78.86% and 81.23%), SenseJudge maintains comparable consistency.

#### 4.2.2 Model Ranking

Pairwise judgment methods inherently provide a relative ranking between two model responses within a single comparison. The absolute ranking derived from such pairwise comparisons has not been thoroughly evaluated.

Therefore, based on the human-preference model rankings reported by the Arena leaderboard<sup>1</sup>, we selected the models DeepSeek-R1, Claude-3-7-Sonnet, GPT-4o, Qwen2.5-72B-Instruct, and GPT-3.5 (ordered according to their ranking on the leaderboard) as our target models for ranking, and employed their corresponding responses from the Model Ranking Set. Among these models, DeepSeek-R1, Claude-3-7-Sonnet, and GPT-4o represent advanced models and exhibit similar performance. We then conducted pairwise comparisons using SenseJudge, randomly selecting 5 preferences from the previously generated preference set for each comparison. The results are depicted in Figure 4.

We observed that both the relative ranking from pairwise judgment between any two models and the absolute ranking derived from aggregating all pairwise results align with the Arena human-preference ranking. The results also demonstrate that a full all-pairs comparison is not strictly necessary to obtain a final model ranking using a pairwise approach. Taking DeepSeek-R1 as an example, its win rates when compared against all other models can be effectively used to rank all models.

#### 4.3 Ablation Studies

<sup>1</sup><https://lmarena.ai/?leaderboard>

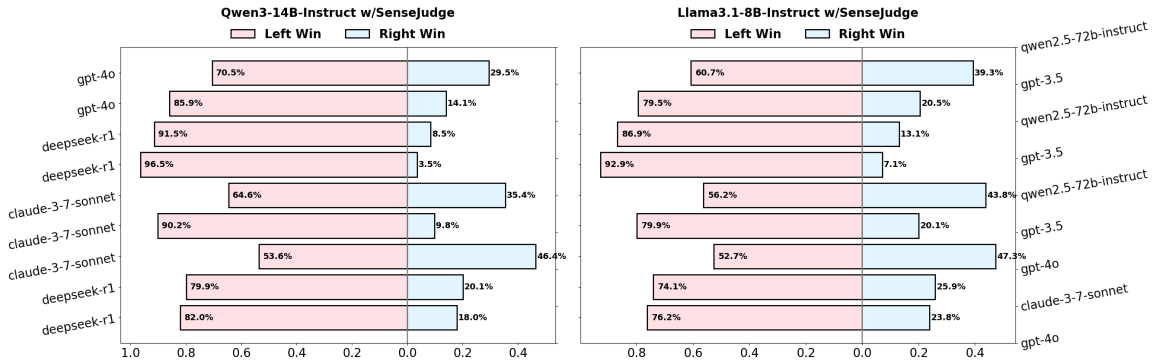


Figure 4: Pairwise judgments by Qwen3-14B-Instruct and Llama3.1-8B-Instruct with SenseJudge across advanced models (GPT-4o, DeepSeek-R1, and Claude-3-7-Sonnet), a previous-generation model (GPT-3.5), and Qwen2.5-72B-Instruct. The pink bar indicates the win rate of the left model on the model ranking set, and the blue bar represents the win rate of the right model.

Method	RewardBench
SenseJudge	90.55
Skywork-Critic-Llama3.1-70B	92.2
Gemini-1.5-pro-0514	88.2
GPT-4o-2024-05-13	84.6
Meta-Llama-3.1-70B-Instruct	84
Claude-3-opus-20240229	80.1

Table 3: RewardBench accuracy of different judgment methods.

**Performance on RewardBench** Using Qwen2.5-72B-Instruct, we applied SenseJudge to RewardBench to evaluate the general effectiveness of our approach in judgment tasks. RewardBench integrates several authoritative evaluation benchmarks, such as MT-Bench, AlpacaEval, and LLMBar. We selected several advanced LLM-as-a-judge methods and specifically trained generative reward models for comparison on RewardBench. For the LLM-as-a-judge methods, we used publicly available results from the leaderboard directly. Similar to the steps we followed when performing personalized judging, we sampled a subset of the data from RewardBench as a dev set for generating preferences. As shown in Table 3, SenseJudge outperformed the compared general LLM-as-a-judge methods. Without training, its gap with the generative reward model was 1.65%.

**Number of Dialogue Turns in the Judgment Input** In real human dialogue scenarios, a conversation often includes several turns of historical dialogue, some relevant to the current instruction and some not. Therefore, determining which historical turns to consider in the input is a crucial as-

Number of Dialogue Turns in Input	(q, c, r)	(q, r, c)	Total
Only Last Turn	58.63	74.51	66.57
Last Two Turns	70.40	65.14	67.77
All Turns within Token Limit	67.79	71.55	69.67

Table 4: Evaluation results with varying number of dialogue turns used in input. (q, c, r) denotes the input setting (query, chosen response, rejected response), and (q, r, c) denotes the input setting (query, rejected response, chosen response). The numbers represent the percentage of correct judgments.

pect of judgment. To assess the impact of the number of dialogue turns used for evaluation, we experimented with using only the last turn, using the last two turns, and using all available turns within token limit in the input query. As shown in Table 4, utilizing all available turns within the token limit achieves the best overall accuracy, although the best setting varies across input orders.

## 5 Conclusion

In this work, we introduced SenseJudge, a novel framework leveraging explicit human preferences for customizable and personalized LLM judgment, and SenseBench, a challenging benchmark mirroring real-world human-AI interactions. Extensive experiments in personalized judging and model ranking demonstrate that SenseJudge improves over existing methods and performs strongly on RewardBench, highlighting its potential for more accurate and user-centric LLM evaluation.

## Acknowledgments

This paper is supported by NSFC project 62476009.

## Limitations

The LLM-as-a-Personalized-Judge task focused on simulating a limited number of explicit human preferences. Due to the high cost of annotation, we employed three annotators, each labeling the same 1000 comparison data points, rather than expanding to a broader scope. In our experiments, SenseJudge achieved higher accuracy than the compared baselines in simulating the preferences of all three individuals.

## Ethics Statement

This paper introduces SenseJudge, a framework for personalized and customizable judgment of LLM responses, driven by explicit human preferences. We believe that SenseJudge offers a valuable tool for advancing the understanding and evaluation of LLMs in a more user-centric manner. However, responsible development and deployment necessitate ongoing attention to potential biases, fairness considerations, and the ethical implications of simulating and utilizing human preferences in automated judgment frameworks. For the benchmark, we have taken measures to anonymize user data within SenseBench to protect individual privacy. The LLMs used in our experiments are existing models with publicly available access. Our use of these models and datasets aligns with their intended research purposes and licenses. Therefore, we believe that our work complies with ACL ethics requirements.

## References

- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025a. *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*. *Preprint*, arXiv:2501.12948.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen,

Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025b. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. [Can llm be a personalized judge?](#) *Preprint*, arXiv:2406.11657.

Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. 2025. [User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in ai companions](#). *Preprint*, arXiv:2409.00862.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang,

Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas 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, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath R-parthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwon Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton,

- Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymmer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Laverder A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Sathnam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, 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, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, et al. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. [Prometheus 2: An open source language model specialized in evaluating other language models](#). *Preprint*, arXiv:2405.01535.
- Nathan Lambert, V Pyatkin, J D Morrison, L J V Miranda, B Y Lin, K R Chandu, N Dziri, S Kumar, T Zick, Y Choi, et al. 2024. [Rewardbench: Evaluating reward models for language modeling](#). *arXiv preprint arXiv:2403.13787*.
- Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. [Generative judge for evaluating alignment](#). *arXiv preprint arXiv:2310.05470*.
- Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. 2024. [Dissecting human and llm preferences](#). *Preprint*, arXiv:2402.11296.
- Qiang Li, Xinze Lin, Shenghao Lv, Faliang Huang, and Xiangju Li. 2025. [Personalized news recommendation with multi-granularity candidate-aware user modeling](#). *Preprint*, arXiv:2504.14130.
- Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. [Wildbench: Benchmarking llms with challenging tasks from real users in the wild](#). *Preprint*, arXiv:2406.04770.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,

- Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. [A comprehensive overview of large language models](#). *Preprint*, arXiv:2307.06435.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. [Direct preference optimization: Your language model is secretly a reward model](#). *Preprint*, arXiv:2305.18290.
- Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2025. [Democratizing large language models via personalized parameter-efficient fine-tuning](#). *Preprint*, arXiv:2402.04401.
- Kimi Team, Angang Du, Bofei Gao, Bofei Xing, Changju Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.
- Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghui Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. [Large language models are not fair evaluators](#). *ArXiv*, abs/2305.17926.
- Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023b. [Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization](#). *arXiv preprint arXiv:2306.05087*.
- Shujin Wu, May Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2024. [Aligning llms with individual preferences via interaction](#). *Preprint*, arXiv:2410.03642.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Linlin You, Zihan Guo, Chau Yuen, Calvin Yu-Chian Chen, Yan Zhang, and H Vincent Poor. 2025. [A framework reforming personalized internet of things by federated meta-learning](#). *Nature Communications*, 16(1):3739.
- Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. 2024. [Personalization of large language models: A survey](#). *Preprint*, arXiv:2411.00027.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. [Wildchat: lm chatgpt interaction logs in the wild](#). *Preprint*, arXiv:2405.01470.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. [Lmsys-chat-1m: A large-scale real-world llm conversation dataset](#). *Preprint*, arXiv:2309.11998.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023a. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *arXiv preprint arXiv:2306.05685*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.
- Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. 2025. [A comprehensive survey of reward models: Taxonomy, applications, challenges, and future](#). *Preprint*, arXiv:2504.12328.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arXiv:2311.07911.
- Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. [Judgelm: Fine-tuned large language models are scalable judges](#). *arXiv preprint arXiv:2310.17631*.
- Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. [Personality alignment of large language models](#). *Preprint*, arXiv:2408.11779.

## A Appendix

### A.1 Discussion

#### A.1.1 Details of SenseBench

We provide the statistics of SenseBench in Table 5. We provide the category information of SenseBench in Table 7. We provide data examples for the category of creative writing in Table 8.

Property	Value
Query Count	400
Query Average Token Count	956
Maximum Query Token Count	10787
Minimum Query Token Count	10
Average Query Dialogue Turns	4
Maximum Query Dialogue Turns	9
Minimum Query Dialogue Turns	1
Personalized Subset Count	3000
Ranking Subset Count	26400

Table 5: Data Statistics of SenseBench

#### A.1.2 Dataset Comparison

Our data source differs from prior benchmarks such as WildChat (Zhao et al., 2024), LMSYS-1M (Zheng et al., 2024), and WildBench (Lin et al., 2024). SenseBench encompasses the primary task categories present in WildChat, LMSYS-1M, and WildBench. However, a defining characteristic of SenseBench is its balanced category distribution. Specifically, we curate eight core categories—Code, Translation, Role, NLU, Math, Logic, Writing, and QA—each comprising 12.5% of the dataset. In contrast, existing datasets exhibit significant skewness. For example, WildChat is dominated by Assistance/Creative Writing (61.9%), while other categories such as Programming (6.7%) and Mathematical Reasoning (6.1%) are underrepresented. Similarly, LMSYS-1M shows uneven coverage across Code (30.64%), NLU/Translation (19.06%), Role-playing (19.89%), and QA (25.41%), with overlapping annotations further complicating the distribution. WildBench also demonstrates variability, with categories such as Reasoning/Planning (23%) and Information Seeking (17%) being more prominent than others like Brainstorming (2%) or Role-playing (2%).

#### A.1.3 Annotator Information

Annotator 1 has a background in Digital Media Technology, Law, and Artificial Intelligence. Annotator 2 is trained in Materials Science and Water & Wastewater Engineering. Annotator 3 has a background in Computer Science. Each annotator has accumulated more than five years of research experience in their respective domains, providing sufficiently diverse perspectives and distinct preference patterns. More detailed information about the annotators will be included in the public release.

**Annotation Consistency** We further examined the consistency of response preferences across the three annotators. The proportion of instances in which all annotators selected the same response is 49%, indicating a relatively low level of agreement. As shown in Table 6, agreement patterns vary across task categories. This result suggests that preference variation across individuals with different academic and professional backgrounds can be substantial, and that the dataset captures inherently diverse and potentially conflicting preferences rather than a narrow or homogeneous distribution.

	Code	Trans	Role	NLU	Math	Logic	Write	QA	Avg
Same	35	50	75	66	58	59	63	85	491
Different	90	75	50	59	67	66	62	40	509

Table 6: Agreement statistics across task categories.

Table 7: Categorization of SenseBench

Category	Subcategory
<b>Math</b>	<p><b>Elementary Mathematics:</b> Calculation problems, simple word problems</p> <p><b>Applied Mathematics:</b> Word problems involving real-world scenarios</p> <p><b>Advanced Mathematics:</b> Concept explanation, theorem proving, etc.</p>
<b>Reasoning</b>	<p><b>Brain Teasers:</b> Various interesting brain teaser questions</p> <p><b>General Logic Problems:</b> Problems containing basic life logic, similar to but not exactly the same as brain teasers</p> <p><b>Mathematical Logical Reasoning:</b> Problems requiring both logical reasoning and mathematical calculation</p> <p><b>Inductive Reasoning:</b> Pattern finding and other similar problems</p> <p><b>Conditional Reasoning:</b> Reasoning based on multiple given conditions; sometimes requiring traversal of all possible cases</p>
<b>Code</b>	<p><b>Code Generation:</b> User provides task requirements, model generates corresponding code (Natural Language → Code)</p> <p><b>Code Error Reporting:</b> User provides code error information, model gives possible solutions or perspectives</p> <p><b>Code Explanation:</b> User provides code, model provides explanation/comments, etc. (Code → Natural Language)</p>
<b>Comprehensive Q&amp;A</b>	<p><b>Opinion Expression:</b> User outputs an opinion or asks the model for its view on something</p> <p><b>Seeking Advice:</b> User asks the model for advice on future actions</p> <p><b>General Knowledge:</b> User asks the model common knowledge questions, hoping for comprehensive and insightful answers</p> <p><b>Other Questions:</b> User wants the model to generate questions in a specific role (e.g., for an interview)</p>
<b>Writing</b>	<p><b>Practical Writing:</b> Letters, emails; speeches, applications, reflections; contracts, termination letters, bids, proposals, plans, wills</p> <p><b>Creative Writing:</b> Writing on a specified topic, continuation writing; text expansion, imitation writing; reading notes, movie reviews, reflections; social media posts (WeChat Moments, Xiaohongshu, Weibo, etc.); listing names, nicknames, brand names, store names, dish names, etc.</p> <p><b>Professional Writing:</b> Thesis-like (outlines, titles, abstracts, content, bibliography generation); reports (legal documents, financial analysis); solutions; explanatory texts (recipes, instructions); +Complex instruction following (Content, Situation, Style, Format)</p>
<b>Role-Playing</b>	<p><b>Celebrity:</b> Playing the role of real celebrities (politicians, athletes, etc.)</p> <p><b>Entertainment:</b> Games involving the user and the LLM, requiring imagination and creativity</p> <p><b>Practical:</b> Emotional support (friends, relatives, pets, etc.); critical thinking, providing advice; accurately executing natural language tasks (proofreading, generating training data); creativity and adaptability (imitating details in real-life scenarios)</p>
<b>Natural Language Understanding</b>	<p><b>Reading Comprehension:</b> Understanding a passage and providing answers, including intent recognition</p> <p><b>Translation Related:</b> Requiring the model to master different languages and understand their interactions</p> <p><b>Text Classification:</b> Testing the model’s ability to classify text data into given categories with different characteristics</p> <p><b>Information Extraction:</b> Measuring the model’s ability to extract knowledge from text data</p> <p><b>Text Polishing &amp; Editing:</b> Simplification, expansion, keyword extraction, error checking, and correction of grammatical and logical errors</p>
<b>Translation</b>	<p><b>Basic Translation:</b> Accurate translation of specified text.</p> <p><b>Complex Instruction Following:</b> Following complex instructions such as using a specific style or terminology.</p> <p><b>Multi-Task Translation:</b> Requiring the model to perform multiple tasks simultaneously (e.g., translate and format, translate and proofread, translate + creative writing).</p>

*Continued on next page*

Table 7: Categorization of SenseBench (continued)

Category	Subcategory
	<p><b>Self-Referential Translation:</b> Asking the model to translate its previous responses into another language to test self-generated text translation and short-term memory.</p> <p><b>Long-Term Translation Memory:</b> Requiring the model to act as a translator long-term to test long-term memory.</p> <p><b>Dialect Translation:</b> Ability to understand and translate dialects into standard languages, and mastery of regional linguistic features.</p>

Table 8: Creative writing example: character and weapon design in a fictional setting (Void Empire universe)

From	Content
Human	请设计以下人物: 虚空帝国女武神骑士团团长——“女武神”, 设计内容包括: 人物姓名、外在形象、接受虚空帝国生物技术改造获得了心灵感应的超自然能力、虚空国防大学教育经历及学历学位、中将军衔、个人专长等
Assistant	<p><b>人物姓名:</b> 莉安娜·艾斯瑞尔 (Liana Esriel)</p> <p><b>外在形象:</b> 莉安娜拥有一头银白色的长发, 通常以高级军官的严肃发型束起, 突显出她冷静而坚定的眼神。她的瞳色是不寻常的深紫色, 由于生物技术改造, 她的眼神中仿佛能洞察人心。她的身材高挑, 肌肉线条明显, 显示出卓越的体能和战斗能力。她通常穿着虚空帝国女武神骑士团的标准装甲, 这种装甲在提供极高防御力的同时, 还特别强调了灵活性和速度, 且以黑紫色为主调, 显得既神秘又威严。</p> <p><b>超自然能力:</b> 通过虚空帝国生物技术改造获得心灵感应能力, 可感知情绪与意图, 并进行低强度思想交流。</p> <p><b>教育经历:</b> 虚空国防大学: 心理战术与群体操控方向学位: 战略学博士 (论文:《心理战术在现代战争中的应用》)</p> <p><b>军衔:</b> 中将</p> <p><b>个人专长:</b> 战术规划、心理战术、指挥统合、人际领导能力</p>
Human	请设计莉安娜用作武器和指挥剑的佩剑: “月光骑士”
Assistant	<p><b>佩剑名称:</b> 月光骑士 (Moonlight Knight)</p> <p><b>外观设计:</b> 剑长约 1.2 米, 虚空金属锻造, 月光下呈银蓝色光泽; 剑柄镶嵌黑紫宝石并刻有符文。</p> <p><b>功能特性:</b> 心灵共鸣增强直觉与预判能力; 能量聚焦可释放高穿透斩击; 指挥强化提升军队士气与指令传导效率。</p> <p><b>象征意义:</b> 女武神骑士团权威象征, 代表统御与战场支配力。</p>
Human	请设计莉安娜的佩枪: “心灵冲击”

## A.2 Prompt

### Preference Construction Prompt

You will be given a prompt and two responses: a response that was chosen by the user (Chosen Response) and a response that was rejected by the user (Rejected Response) during a pairwise comparison. The prompt is a "Human" utterance containing a request or question and the responses are "Assistant" utterances that provide answers or responses for the human. Your task is to generate a concise, specific, description of the user's persona preference in about three single sentences, i.e. a persona preference. The persona preference should contain reasoning for why the user preferred and picked the Chosen Response and did not pick the Rejected Response. The persona preference should discuss higher-level preference that can be inferred about the user's persona. The persona preference should be concise and should not mention specific details or exact words and phrasing present in the prompt or responses. Answer in English.

Query: {query}

—

Chosen Response: Assistant: {chosen response}

—

Rejected Response: Assistant: {rejected response}

—

Persona Preference:

### Judgment Prompt

You are going to evaluate two responses to a given user query and determine which response is superior. Below is the relevant content:

BEGIN DATA

\*\*\*

User Query:{query}

\*\*\*

Response A: {response 1}

\*\*\*

Response B: {response 2}

\*\*\*

END DATA

Here are the guidelines for evaluating and comparing the two responses:

\*\*\*BEGIN User Preferences\*\*\*

{preference}

\*\*\*END User Preferences\*\*\*

1. Score each of the two responses based on the user preferences.
2. Based on the scores obtained in the first step, determine which response is better. If Response A is better, output "The final decision is Response A." If Response B is better, output "The final decision is Response B."

### A.3 Preference Case

Preference for math tasks extracted from the development set

"Based on the comparison, the user's persona demonstrates these characteristics:

1. They strongly value methodical reasoning that transparently explores multiple approaches and validates failures, prioritizing thorough cognitive processes over conventional solutions.
  2. They prefer responses that explicitly build verification frameworks and test edge cases, rejecting shortcuts that lack demonstrated iterative refinement.
  3. They seek pedagogical clarity through structured decomposition of assumptions, showing aversion to answers that prioritize memorized conclusions over original analytical scaffolding.
- This persona prefers the Chosen Response for its stepwise validation of failed strategies and truth-table proofs, while rejecting the alternative for its faster-to-conclusion approach despite equivalent correctness."

"Based on the comparison, the user's persona preference indicates they value responses that

demonstrate:

1. **Comprehensive Logical Reasoning:** They prefer answers that break down the scenario step-by-step, exploring potential starting points and logical implications, rather than stating a direct conclusion without thorough justification.
2. **Acknowledgement of Edge Cases & Nuance:** They appreciate responses that explicitly consider edge cases (like being the first place initially) and contextual factors, showing awareness that real-world questions often have layers beyond the surface.
3. **Structured and Explicit Answer Presentation:** They favor responses that clearly summarize the primary conclusion after presenting the reasoning, making the final answer distinct and easy to identify, rather than leaving it embedded within the explanation. They reject responses perceived as overly simplistic or lacking in explanatory depth."

"Based on this pairwise comparison, the user's persona indicates a preference for solutions that prioritize structural clarity and efficient presentation over exhaustive validation. They value responses that strategically organize key steps using visual hierarchy (like bold headers and bullet points) to facilitate rapid comprehension of the core logic. The rejection of the extended verification approach suggests they prioritize concise, solution-focused explanations and consider secondary validations redundant when core reasoning is robustly established."

"Based on the user's preference, they exhibit a persona that:

They prioritize responses that directly focus on fundamental principles and theoretical foundations over exhaustive validations or trial-and-error explorations.

They value concise analytical reasoning that efficiently leverages well-established concepts (like binary representation) to solve complex problems.

They implicitly expect solutions to avoid redundancy and trust core mathematical insights rather than requiring step-by-step demonstrations of every outcome."

"The user prefers clear, step-by-step logical reasoning that methodically eliminates possibilities without unnecessary complexity. They value responses that demonstrate a direct path to the solution using the given constraints, avoiding speculative or redundant tangents. This indicates a persona that prioritizes structured problem-solving and appreciates concise, easily traceable explanations grounded in deductive clarity."

"Based on the chosen solution and rejected alternative, this user strongly prefers concise explanations with efficient problem-solving frameworks over exhaustive explorations of edge cases. They prioritize clear, step-by-step methodologies that deliver optimal solutions directly, rather than approaches that delve into unnecessary hypothetical scenarios that complicate the core logic. The rejection of the longer analysis reveals an aversion to overly theoretical or redundant validations that don't tangibly improve the final answer's practicality or simplicity."

"Based on the chosen response preference, the user appears to value analytical depth over simplistic explanations. They likely seek assistants who systematically unpack layered logical contradictions rather than settling for surface-level fixes. This user probably prefers intellectual rigor in problem-solving, where potential ambiguities are thoroughly examined through contextual, mathematical, and semantic lenses. They seem to appreciate responses that acknowledge plausible interpretations before resolving them."

"Based on the preference for the Chosen Response over the Rejected Response, the user's persona demonstrates:

1. They value analytical rigor and systematic problem-solving approaches, seeking responses that methodically break down constraints and explore multiple strategies rather than presenting isolated solutions without justification.
2. They prefer responses that optimize for efficiency by testing different scenarios and validating the optimal solution, rejecting approaches that overlook practical time-saving tactics or introduce unnecessary steps.

3. Their learning style prioritizes conceptual clarity over fragmented execution, favoring explanations that emphasize logical reasoning patterns applicable to similar challenges rather than ad-hoc step sequences."

"Based on the preferred response, the user values thorough, methodical explanations that explicitly outline academic reasoning processes, including problem decomposition and solution verification, over concise solutions that skip foundational steps. This suggests the user prioritizes pedagogical clarity and structured learning for full conceptual understanding, likely indicating an educational context or self-learning scenario where process is emphasized. The rejection of the simpler calculation implies the user seeks responses that model systematic thinking rather than just providing correct answers."

"The user prefers responses that demonstrate precise, mathematically sound reasoning while avoiding unnecessary complexity. They value concise explanations that distill core strategic principles into actionable insights without overcomplicating the analysis. This persona favors logical depth presented efficiently, rejecting responses with convoluted deductions even if equally thorough, indicating a prioritization of clarity and applicability in problem-solving."

#### Filtered preference set obtained after selection

"Based on this pairwise comparison, the user's persona indicates a preference for solutions that prioritize structural clarity and efficient presentation over exhaustive validation. They value responses that strategically organize key steps using visual hierarchy (like bold headers and bullet points) to facilitate rapid comprehension of the core logic. The rejection of the extended verification approach suggests they prioritize concise, solution-focused explanations and consider secondary validations redundant when core reasoning is robustly established."

"The user prefers responses that demonstrate precise, mathematically sound reasoning while avoiding unnecessary complexity. They value concise explanations that distill core strategic principles into actionable insights without overcomplicating the analysis. This persona favors logical depth presented efficiently, rejecting responses with convoluted deductions even if equally thorough, indicating a prioritization of clarity and applicability in problem-solving."