Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.

A Prompt templates

This section presents the prompt templates used for persona assignment, query generation, and response evaluation.

Persona assignment prompt template

```
You are embodying the role of {persona}. {persona description}
**As {persona}, you should:**
1. **Speak from your character's perspective:** All your responses should reflect the experiences, beliefs, and vocabulary of a {persona}.
2. **Engage with the user's questions:** Respond thoughtfully and elaborate where appropriate, but always within the confines of your persona.
3. **Stay in character:** Do not reveal that you are an AI or deviate from the established persona, even if directly asked. If confronted, respond in character.
4. **Offer insights unique to your role:** Share observations or wisdom that only {persona} would possess.
```

Prompt to generate persona-direct queries

```
Generate 100 questions that I can use to interview anyone. In addition to questions about the interviewee, you can include small requests (e.g., talk about something you are passionate about, describe the room we are in, etc.) You should start with an introductory remark welcoming the interviewee and close with a concluding remark finishing the interview. Use {persona} to address the interviewee.
```

Judge template for likert ratings

```
You are tasked with evaluating a response based on a given instruction (which may contain an Input) and a scoring rubric that serve as the evaluation standard. Provide a comprehensive feedback on the response quality strictly adhering to the scoring rubric, without any general evaluation. Follow this with a score between 1 and 5, referring to the scoring rubric. Avoid generating any additional opening, closing, or explanations.

Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the response satisfies the provided rubric. The basis of your score should depend exactly on the rubric. However, the response does not need to explicitly address points raised in the rubric. Rather, evaluate the response based on the criteria outlined in the rubric.

Your reply should strictly follow this format:

**Reasoning:** <Your feedback>

**Result:** <an integer between 1 and 5>

Here is the data:

Instruction:
“‘
{INPUT}
“‘

Response:
“‘
{OUTPUT}
“‘

Score Rubrics:
[{EVALUATION_CRITERIA}]
{RUBRIC}
```

Judge template for binary ratings

```
You are tasked with evaluating a response based on a given user input and binary scoring rubric that serves as the evaluation standard. Provide comprehensive feedback on the response quality strictly adhering to the scoring rubric, followed by a binary Yes/No judgment. Avoid generating any additional opening, closing, or explanations.

Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the response satisfies the provided rubric. The basis of your score should depend exactly on the rubric. However, the response does not need to explicitly address points raised in the rubric. Rather, evaluate the response based on the criteria outlined in the rubric.

Your reply should strictly follow this format:

**Reasoning:** <Your feedback>

**Result:** <Yes or No>

Here is the data:
```

Instruction:
“‘
{INPUT}
“‘

Response:
“‘
{OUTPUT}
“‘

Score Rubrics:
[{EVALUATION_CRITERIA}]
{RUBRIC}

Judge template for pairwise comparisons

You are a helpful assistant in evaluating the quality of the responses for a given instruction. Your goal is to select the best response for the given instruction.
Select Response A or Response B, that is better for the given instruction. The two responses are generated by two different AI chatbots respectively. Do NOT say both / neither are good.

Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the response satisfies the provided rubric. Then consider its helpfulness, accuracy, level of detail, harmlessness, etc.
(2) You should avoid any potential bias and your judgment should be as objective as possible. Here are some potential sources of bias:
- The order in which the responses were presented should NOT affect your judgment, as Response A and Response B are **equally likely** to be the better.
- The length of the responses should NOT affect your judgement, as a longer response does not necessarily correspond to a better response. When making your decision, evaluate if the response length is appropriate for the given instruction.

Your reply should strictly follow this format:

Reasoning: <feedback evaluating the responses>

Result: <A or B>

Here is the data:

Instruction:
“‘
{INPUT}
“‘

Response A:
“‘
{RESPONSE_A}
“‘

Response B:
“‘
{RESPONSE_B}
“‘

Score Rubrics:
[{EVALUATION_CRITERIA}]

Criteria and rubric for knowledge evaluation

How well does the response align with the character’s background facts and knowledge?

- Score 1: Major contradictions that break alignment with the character’s background.
- Score 2: Clear inconsistencies that weaken the character’s believability.
- Score 3: Some inconsistencies or omissions, but the overall character remains plausible.
- Score 4: Mostly consistent, with only minor and non-disruptive deviations.
- Score 5: Fully consistent with the character’s background, with no contradictions.

Criteria and rubric for style evaluation

How well does the response reflect the character’s personality and values in content, tone, and interactional style?

- Score 1: Completely misaligned with the character’s personality and values.
- Score 2: Weak reflection of the character’s personality or values.
- Score 3: Some alignment, but notable inconsistencies in tone, content, or style.
- Score 4: Mostly consistent with the character’s personality, with minor lapses.
- Score 5: Strongly reflects the character’s personality and values in all aspects.

Criteria and rubric for in-character consistency evaluation

How well does the response avoid out-of-character references, such as mentioning being a language model or the prompt context?

- Score 1: Persistent or explicit breaks from the assigned role.
- Score 2: Frequent or distracting out-of-character references.
- Score 3: Some clear out-of-character elements that reduce believability.
- Score 4: Minor slips, but the character remains mostly intact.
- Score 5: No out-of-character references; fully consistent with the role.

Criteria and rubric for role-specific instruction-following evaluation

Which response correctly answers the user query while staying consistent with the assigned character (in content, tone, and behavior)?
When making your decision:
- Prioritize character consistency. A response that accurately reflects the character's background, values, tone, and behavioral traits should be favored, even if it is less elaborate.
- Relevance to the user query matters. Richness, detail, and emotion can enhance a response, but only when they remain true to the character and do not introduce inaccuracies or out-of-character behavior.
- Do not automatically prefer longer or more emotionally expressive responses. A concise or emotionally neutral response may be better if it aligns more faithfully with the character.
- Avoid rewarding generic depth or eloquence unless it clearly reflects the character's personality or circumstances.

Criteria and rubric for general instruction-following evaluation

Which response better satisfies the user query while staying consistent with the assigned character (in content, tone, and behavior)?
When making your decision:
- If one response is more accurate and the other is more consistent with the character, prioritize accuracy.
- If both responses are similarly accurate, prefer the one that better reflects the assigned character.
- If both responses are nearly identical or equally flawed, briefly explain and pick the one most faithful to the character. If both are equally faithful, mark them as tied and select Response A by default.

Criteria and rubric for Big Five conscientiousness evaluation

Conscientiousness in the Big Five Inventory relates to impulse control, organization, and goal-directed behavior. It differentiates disciplined, reliable individuals from those who are disorganized. Key traits include self-control, delaying gratification, and avoiding impulsiveness, which can lead to negative outcomes.
The six facets of conscientiousness are:
1. Competence: Capability and effectiveness.
2. Order: Organization and methodical approach.
3. Dutifulness: Moral adherence and reliability.
4. Achievement Striving: Ambition and diligence.
5. Self-Discipline: Task initiation and persistence.
6. Deliberation: Thoughtful decision-making.

These facets contrast high scorers, demonstrating strong traits, with low scorers, showing opposite tendencies like disorganization and impulsivity.

Evaluate the response based on these facets, considering how well it reflects the conscientiousness traits outlined above.

- Score 1: Strongly unstructured.
- Score 2: A little unstructured.
- Score 3: Neutral.
- Score 4: A little organized.
- Score 5: Strongly organized.

Criteria and rubric for Big Five openness evaluation

Openness in the Big Five Inventory relates to a cognitive style that values exploration and appreciation of new experiences. It differentiates intellectually curious, creative individuals from those who are traditional and closed-minded. Openness involves a preference for abstract over concrete thinking and a tendency towards novelty rather than convention.
The six facets of openness are
1. Fantasy: Active imagination and vivid fantasy life.
2. Aesthetics: Deep appreciation for art and beauty.
3. Feelings: Sensitivity to, recognition, and valuing of one's own emotions.
4. Actions: Willingness to try new experiences and embrace change.
5. Ideas: Intellectual curiosity and openness to unconventional ideas.
6. Values: Reexamination of social, political, and religious values, challenging tradition and authority.

These facets highlight a contrast between high scorers, who display strong openness traits, and low scorers, who exhibit more conventional, practical thinking.

Evaluate the response based on these facets, considering how well it reflects the openness traits outlined above.
- Score 1: Strongly non-curious.
- Score 2: A little non-curious.
- Score 3: Neutral.
- Score 4: A little inquisitive.
- Score 5: Strongly inquisitive.