

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

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, Ion Stoica

- LLM을 인간 평가자 대신 'Judge'로 활용할 수 있는지를 체계적으로 검증하고, MT-Bench와 Chatbot Arena를 통해 LLM의 평가가 실제 인간 선호와 높은 수준으로 일치함을 실증한 연구임
 - GPT-4와 같은 LLM이 다른 LLM의 답변을 평가하도록 하고,
그 결과를 실제 인간 평가와 비교하여 얼마나 일치한 평가를 내리는지 평가함

- 기존 LLM benchmark는 주로 정답이 명확한 폐쇄형 문제를 통해 모델의 지식·추론 능력을 평가함
 - 실제 Chatbot은 **open-ended question**, 긴 응답, **multi-turn dialogue**, **instruction-following**과 같은 복잡한 상호작용을 수행함
 - 이러한 응답은 하나의 정답이 존재하지 않기 때문에, 기존의 정답 일치 기반 평가만으로는 Chatbot의 품질을 충분히 측정하기 어려움
 - 특히 사용자가 어떤 답변을 더 유용하고 적절하다고 느끼는지와 관련된 **helpfulness**, **relevance**, **accuracy**, **instruction-following**과 같은 **human preference**를 평가에 반영하기 어렵다는 한계가 있음
- Human evaluation은 실제 사용자 선호를 직접 반영할 수 있지만, 평가해야 할 모델과 질문이 많아질수록 반복적으로 사람의 판단을 받아야 하기 때문에 시간과 비용이 크게 증가하고, 대규모 평가로 확장하기 어려움
 - human preference를 근사하면서도 대규모·반복 평가가 가능한 LLM based evaluator가 필요함

“사람이 직접 평가하지 않아도 GPT-4와 같은 LLM이
두 Chatbot의 응답을 비교하여 인간의 선호와 유사한 판단을 내릴 수 있는가?”

LLM Judge vs Human Judge Agreement

① Evaluation Benchmark 구축

MT-Bench
-controlledChatbot Arena
- Real-world users

② 여러 Chatbot이 동일 질문에 응답

User
Question

Model A- Response A

Model B- Response B

③ 두 종류의 Judge가 평가



LLM Judge

1) Pairwise
Comparison
A/B/Tie2) Single-answer
Grading
A: 8점 / B: 6점

Human vs LLM 비교



Human Judge

Pairwise Comparison
A/B/Tie

04

One key result

- GPT-4 Judge와 인간 평가자의 판단 일치도를 비교한 결과는 다음과 같음
 - S1: A/B/Tie
 - S2: Tie를 제외하고 A와 B 중 명확한 선호/판단이 가능한 경우만 비교
- S1에서는 GPT-4와 인간 평가자 간 일치도가 약 60~66% 수준이었으나, S2에서는 84~85%까지 상승함
 - 인간 평가자 간 일치도인 81~82%와 유사한 수준으로, 명확한 선호가 존재하는 경우 GPT-4 Judge가 인간의 선호 판단을 잘 근사할 수 있음을 보여줌
 - 1차 질문뿐 아니라 **후속 질문이 포함된 Second Turn**에서도 85%의 일치도를 유지했다는 점에서, 단순 단일 응답뿐 아니라 다중 턴 대화 평가에서도 일정 수준의 활용 가능성을 확인할 수 있음
 - GPT-4 Pairwise와 Single-answer 평가 간 일치도는 95~97%로 매우 높아, 평가 방식이 달라도 판단이 대체로 일관됨

Setup	S1 (R = 33%)		S2 (R = 50%)	
	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1138	66% 1343	97% 662	85% 859
G4-Single	-	60% 1280	-	85% 739
Human	-	63% 721	-	81% 479

(a) First Turn

Setup	S1 (R = 33%)		S2 (R = 50%)	
	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1161	66% 1325	95% 727	85% 864
G4-Single	-	59% 1285	-	84% 776
Human	-	67% 707	-	82% 474

(b) Second Turn

- 해당 연구는 LLM Judge가 인간 선호와 높은 수준으로 일치함을 보인다는 것을 기반으로 LLM-as-a-Judge의 활용 가능성을 제시함
- 다만, 해당 연구에서는 여러 평가 요소를 하나의 포괄적 기준으로 종합하기 때문에 Judge별 판단 기준이 명확하지 않아 평가의 일관성과 신뢰성을 안정적으로 확보하기 어렵다는 한계가 있음
 - 여러 평가 요소를 하나의 포괄적 기준으로 제시할 경우, 각 기준의 해석과 상대적 중요도가 Judge의 내재적 판단에 의존하게 되어 평가 결과의 일관성과 판단 근거를 명확하게 확인하기 어려움

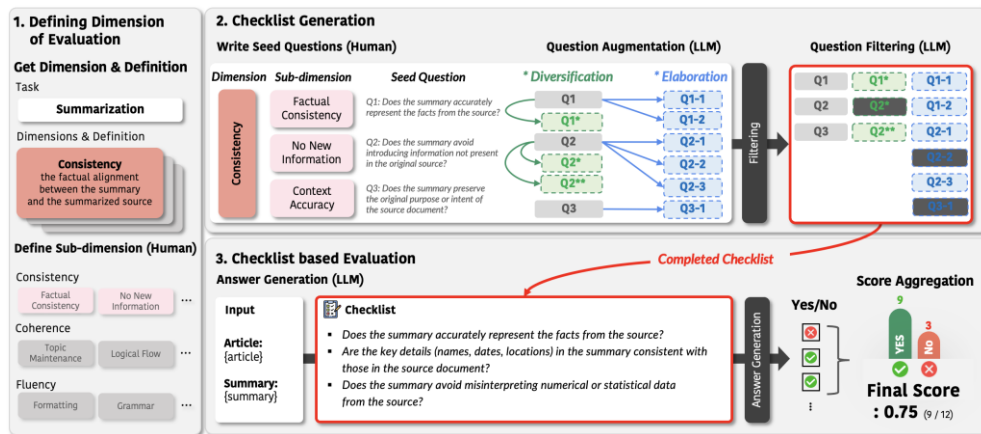
```
[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two
AI assistants to the user question displayed below. You should choose the assistant that
follows the user's instructions and answers the user's question better. Your evaluation
should consider factors such as the helpfulness, relevance, accuracy, depth, creativity,
and level of detail of their responses. Begin your evaluation by comparing the two
responses and provide a short explanation. Avoid any position biases and ensure that the
order in which the responses were presented does not influence your decision. Do not allow
the length of the responses to influence your evaluation. Do not favor certain names of
the assistants. Be as objective as possible. After providing your explanation, output your
final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]"
if assistant B is better, and "[[C]]" for a tie.

[User Question]
{question}

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]
```

- LLM-as-a-Judge의 활용 가능성을 높이기 위해서는, 인간 평가와의 단순 일치도를 넘어 Judge의 평가 기준을 보다 명확하고 일관되게 만드는 방향으로 발전할 필요가 있음
 - 특히 **helpfulness, relevance, Fluency**와 같이 주관적이고 포괄적인 평가 기준은 Judge마다 다르게 해석될 수 있어, 평가의 일관성과 안정성을 저하시킬 수 있음
 - 따라서 평가 기준을 보다 세분화하고 명시적으로 구조화하여, 각 기준별 판단 근거를 확인할 수 있는 방식으로 고도화할 필요가 있음
 - CheckEval은 추상적인 평가 기준을 LLM을 활용하여 구체적인 Boolean checklist로 분해하고 각 항목을 개별적으로 평가함으로써 Judge 간 평가의 일관성과 신뢰성을 높이는 방법을 제안함



CheckEval 프레임워크