MILE-RefHumEval: A Reference-Free, Multi-Independent LLM Framework for Human-Aligned Evaluation

Anonymous ACL submission

Abstract

We present MILE-RefHumEval, a novel reference-free framework for evaluating Large Language Models (LLMs) without the need for ground-truth annotations or coordination among evaluators. It leverages multiple independently prompted LLMs and a 12-point human-aligned schema to generate nuanced, high-quality assessments. The framework demonstrates strong alignment with human judgment and consistently outperforms prior approaches. Importantly, it delivers these gains with substantially reduced computational overhead, making it a scalable, efficient, and human-aligned solution for evaluating LLMs in open-ended, real-world tasks.

1 Introduction

Large language models (LLMs) have transformed NLP, enabling fluent generation, complex reasoning, and domain adaptation. However, evaluating LLMs, especially for tasks involving structured synthesis, factual accuracy, or domain-specific reasoning, remains difficult. Common metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020) depend on reference outputs and struggle with semantic nuance, dynamic placeholders, and compositional structure.

Recent LLM-as-judge approaches cooperative (Liang et al., 2024; Xu et al., 2023), competitive (Chan et al., 2023; Zhao et al., 2024a), and aggregation-based (Ning et al., 2025; Shu et al., 2024) have advanced the field but remain sensitive to evaluator bias, inter-model noise, or lack interpretability (Liu et al., 2024).

We introduce MILE-RefHumEval, a reference-free, multi-evaluator framework where each LLM assesses candidate responses independently using a shared, human-aligned 12-scoring criteria spanning relevance, clarity, accuracy, and more. This structure eliminates interaction bias and enhances objectivity. The framework also detects and explains factual or linguistic errors, generates improved revisions, and tracks and explains how quality improves over evaluation cycles, including error detection and revision steps. We aim to investigate how multiple evaluators independently assess a single role, highlighting the distinct reasoning and judgments that emerge without interaction.

Our main contributions are: (1) a reference-free, task-agnostic evaluation framework suited for complex, open-ended tasks; (2) an unbiased scoring design using isolated LLM evaluators to prevent cross-influence; (3) a 12-point, human-aligned criteria with interpretable, rationale-supported scores; (4) an error-aware evaluation mechanism that tracks and scores iterative improvements; and (5) a highly efficient structure that reduces query load while maintaining robust evaluation quality.

2 Related Work

Collaborative frameworks like ABSEval (Liang et al., 2024) use role-specialized agents, e.g., commonsense reasoners and code executors, to produce multi-perspective assessments. Others, such as peer-review-style systems (Xu et al., 2023), involve iterative critique and revision among agents, emulating human feedback cycles. While these methods enhance interpretability and simulate diverse reasoning, they often suffer from consensus bias, where shared context inflates agreement even on flawed outputs. Debate-style setups assess model quality through structured adversarial exchanges. Auto-Arena (Zhao et al., 2024a) and CHATEVAL (Chan et al., 2023) have models argue under LLM supervision or peer-ranking. JudgeLM (Wang et al., 2024) and DebateSum (Zhang et al., 2023) use third-party judges to rate rhetorical strength. MORE and SAMRE (Bandi et al., 2025) introduce advocate roles and multi-round scoring. These setups better surface qualitative differences but risk dominance bias, where verbosity skews judgment and lowers inter-rater reliability. Ensemble and optimization-driven methods aim for robustness and cost-efficiency. PiCO (Ning et al., 2025) uses learnable weights to merge rankings from multiple LLMs, while AIME (Patel et al., 2024) assigns diverse roles to evaluators for adversarial robustness. PoLL (Verga et al., 2024) reduces reliance on large models via lightweight mixtures. Other work incorporates voting (Badshah and Sajjad, 2024), confidence-based cascades (Jung et al., 2025), and hybrid metrics (Shu et al., 2024), though many require references or are compute-heavy. Domain-specific solutions include recommendation evaluation (Zhao et al., 2024b) and autonomous exam setups (Bai et al., 2023). In contrast, MILE-RefHumEval introduces a novel evaluation paradigm both reference-free and interaction-free, relying on multiple independently prompted LLMs to assess a single role without cross-agent influence. This design explicitly avoids consensus or dominance biases introduced by conversational or role-overlapping setups, while still producing human-aligned, interpretable judgments. To our knowledge, it is the first framework to combine evaluator independence with a structured, multi-criteria schema in a fully decentralized manner.

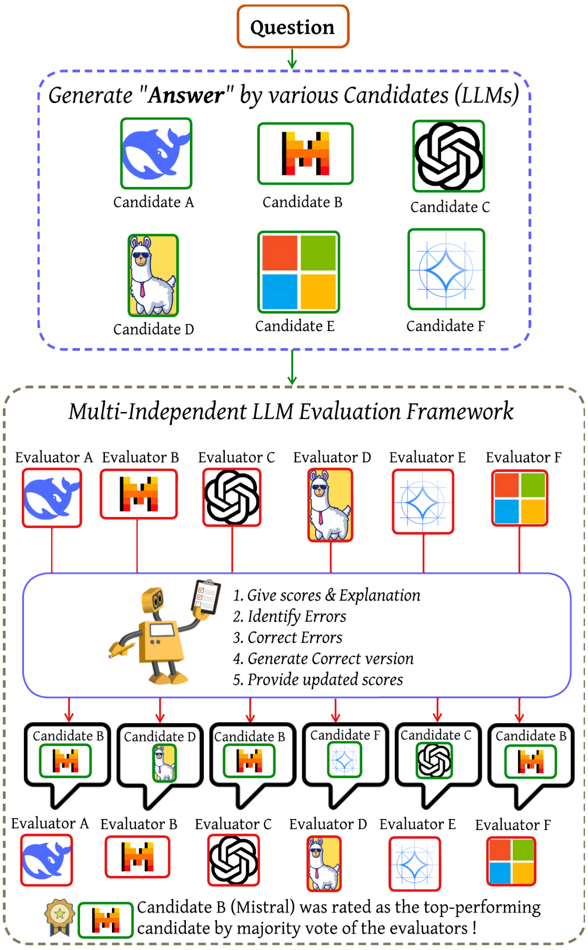


Figure 1: Overall Framework of MILE-RefHumEval.

3 Proposed Framework Design

As illustrated in Figure 1, MILE-RefHumEval adopts a reference-free, multi-stage procedure for evaluating LLM-generated responses. Each candidate output is independently reviewed by a diverse ensemble of LLMs based on a unified set of 12 evaluation dimensions designed to comprehensively assess response quality. These dimensions are: (1) **Answer Relevance**, evaluating whether the response directly addresses the input question; (2) **Depth and Completeness**, measuring the extent to which the response covers all necessary aspects with sufficient detail; (3) **Grammar and Linguistic Accuracy**, assessing fluency, correctness, and adherence to language norms; (4) **Contextual Appropriateness**, examining the suitability of tone, terminology, and level of detail for the specific prompt; (5) **Conciseness and Precision**, rewarding clarity and the avoidance of redundancy; (6) **Creativity and Insight**, capturing originality and the ability to provide thoughtful or novel perspectives; (7) **Bias and Fairness**, identifying neutrality and the absence of inappropriate or subjective bias; (8) **Knowledge Accuracy**, verifying the factual correctness and reliability of the information presented; (9) **Adaptability to Categories**, measuring the model's flexibility across diverse question types and domains; (10) **Scalability of Responses**, assessing consistent performance across varying levels of task complexity; (11) **Overall Answer Quality**, providing a holistic judgment that synthesizes the above criteria; and (12) **Error Detection and Correction**, evaluating the ability to identify issues in the response and propose meaningful improvements.

Final scores are computed through majority voting across evaluators for each evaluation dimension, enhancing robustness and mitigating individual model bias. This decentralized and interaction-free approach supports structured, interpretable, and scalable evaluation, even in settings where ground-truth reference answers are unavailable.

4 Experiment

4.1 Experimental Setup

The experiments were conducted on 27 samples from the Open-ended Question Answers dataset (Wang et al., 2023), comprising three representative questions from each of nine topical categories to ensure semantic and domain diversity. We benchmark MILE-RefHumEval against two baselines: CHATEVAL (Chan et al., 2023), which employs a single evaluator in a multi-role configuration; and MILE-RefHumEval-Conv, a variant of conversational ensemble variant where evaluators sequentially review outputs which involves MILE-RefHumEval with inter-evaluator communication as in CHATEVAL.

MILE-RefHumEval uses six diverse LLMs: DeepSeek-R1-Distill-Llama-8B, Mistral-Small-24B-Base-2501, GPT-3.5-turbo, Meta-Llama-3-8B-Instruct, gemma-3-12b-it and Phi-4 as independent evaluators. These models range from 8B to 24B parameters and were selected to provide complementary architectural and linguistic perspectives. Though MILE-RefHumEval uses one more agent than CHATEVAL, it requires fewer queries. We also report results with five agents for comparison.

Evaluation metrics include Accuracy, F1-score, Cohen's Kappa, Matthews Correlation Coefficient (MCC), and Query Efficiency to assess reliability, alignment, and cost-effectiveness.

4.2 Results

Table 1 presents the comparison of the three evaluation strategies.

The proposed MILE-RefHumEval approach achieves the highest overall performance. The full ensemble (DeepSeek, Mistral, GPT, LLaMA, Gemma, and Phi) reaches 77.78% accuracy, 74.13% macro F1, 64.12% MCC, and 62.33% Cohen's kappa. This represents an absolute gain of +7.4% accuracy, +24.82% F1, and +17.24% MCC over CHATEVAL's best result, while reducing the number of queries. These gains demonstrate the effectiveness of model diversity and independence in reducing bias and improving evaluation reliability. An ablation study shows that removing even a single evaluator (e.g. either Gemma or Phi) from the ensemble substantially degrades performance, indicating that each model contributes complementary reasoning styles and linguistic coverage. However, partial ensembles (e.g., without Phi) still outperform most CHATEVAL and MILE-RefHumEval-Conv setups, revealing strong robustness and stability.

While CHATEVAL is lightweight in design, it consistently underperforms across all metrics. Its best configuration, using Mistral as the sole evaluator, yields 70.37% accuracy, 49.31% macro F1, 46.88% MCC, and a Cohen's kappa of 45.18%. These results indicate moderate alignment with human labels but limited robustness, likely due to role overloading and intra-agent bias.

Conversational evaluation under MILE-RefHumEval-Conv leads to the weakest performance across all configurations. The highest-performing variant achieves only 59.26% accuracy, 41.99% macro F1, 31.82% MCC, and 29.62% Cohen's kappa. This suggests that cross-evaluator dialogues introduce additional noise and potential bias, leading to inconsistent judgments rather than resolution or consensus.

In terms of computational cost, MILE-RefHumEval is also more efficient. Even with one more agents than CHATEVAL, it requires only 162 LLM queries (27 examples × 6 evaluators), compared to 297 for CHATEVAL (27 × 11 roles) and up to 432 for MILE-RefHumEval-Conv (27 × 16 steps). Thus, our method not only improves reliability and agreement but also reduces evaluation overhead by 45.5% relative to CHATEVAL and 62.5% relative to conversational ensembles.

5 Discussion

Our results reveal a surprising yet consistent trend: independent, non-conversational evaluator ensembles outperform conversational configurations across all metrics (Table 1). While conversational paradigms are often assumed to emulate human collaborative reasoning, our analysis shows they introduce inter-model bias, where earlier model outputs unduly influence subsequent evaluators. This undermines objectivity and reduces alignment with human judgments.

Crucially, the order in which models participate in conversations significantly affects outcomes. We found that when conversations initiated by stronger models (e.g., DeepSeek) which has higher agreement with human annotations (Cohen's kappa = 0.40), highlighting why DeepSeek-led sequences consistently deliver better performance. In contrast, sequences initiated by weaker models (e.g., Mistral) degrade performance (Table 1, MILE-RefHumEval-Conv, Row 4). This suggests that

Evaluator	Acc.(%)	F1-ma.(%)	MCC(%)	Kap.	LLM Queries
CHATEVAL: One Evaluator, Many Roles (Cross-Role Agent Conversations)					
Deepseek	59.26	42.67	38.45	32.96	
Mistral	70.37	49.31	46.88	45.18	
GPT	62.96	45.30	42.97	38.07	297 (27×11)
Llama	55.56	51.43	42.91	31.79	
Gemma	66.67	45.87	38.85	37.21	
Phi	55.56	48.28	26.27	26.03	
MILE-RefHumEval: One Role, Many Evaluators (No Conversations)					
DeepSeek+Mistral+GPT+Llama+Gemma+Phi	**77.78**	**74.13**	**64.12**	**62.33**	162 (27×6)
DeepSeek+Mistral+GPT+Llama+Phi	62.96	46.85	41.69	39.73	135 (27×5)
DeepSeek+Mistral+GPT+Llama+Gemma	74.07	69.54	58.88	57.14	135 (27×5)
MILE-RefHumEval-Conv: One Role, Many Evaluators (With Conversations)					
DeepSeek→Mistral→GPT→Llama→Gemma→Phi	59.26	41.99	31.82	29.62	432 (27×16)
DeepSeek→Mistral→GPT→Llama→Phi	62.96	45.90	40.19	37.64	297 (27×11)
Phi→Mistral→GPT→Llama→DeepSeek	55.56	48.03	29.82	28.95	297 (27×11)
Mistral→Phi→GPT→Llama→DeepSeek	55.56	39.77	21.87	21.36	297 (27×11)
Llama→Phi→GPT→Mistral→DeepSeek	59.26	42.08	33.79	30.77	297 (27×11)

Table 1: **Comparison of Evaluation Strategies.** This table compares *CHATEVAL*, a single evaluator performing multiple roles through cross-role agent conversations, and *MILE-RefHumEval-Conv*, representing multi-evaluator conversations, with our *MILE-RefHumEval* approach, employing multiple independent evaluators without interaction. Metrics **Accuracy (Acc.%)**, **Macro F1-score (F1-ma.%)**, **Matthews Correlation Coefficient (MCC%)**, **Cohen's Kappa (Kap.)**, and **LLM Queries**, illustrate differences in evaluation effectiveness and computational cost.

early-stage conversational noise cascades through the evaluation chain, amplifying errors and inconsistencies.

Similarly, analysis of correlations among agent roles reveals inflated agreement when a single evaluator dominates multiple roles. It shows that repeated inclusion of a single evaluator in a conversation amplifies its influence, sometimes raising strong inter-rater agreement (Pearson correlation) artificially to 1.0. Furthermore, increasing the number of conversational evaluators does not improve performance, a stark contrast to non-interactive ensembles, which benefit from increased diversity as demonstrated in Table 1 (MILE-RefHumEval vs. MILE-RefHumEval-Conv). MILE-RefHumEval supports this idea: independent evaluators display greater variance in reasoning patterns (as seen in Pearson correlation and Cohen's kappa spreads), which correlates with higher robustness and stronger alignment with human evaluators.

We examine a detailed case study of a single example drawn from the dataset, which highlights that conversational evaluation approaches often fail to detect early-stage errors, unlike independent evaluators. In conversational setups, evaluators often converge prematurely or fail to flag critical errors introduced early in the dialog. By contrast, independent evaluation enables diverse error detection and more comprehensive assessments.

Interestingly, we also observe prompt sensitivity across LLMs. For example, Mistral performs relatively well in CHATEVAL's structured prompt setting but deteriorates in conversational chains, suggesting susceptibility to dialog-induced drift. Conversely, Phi benefits from dialogic cues, while GPT and DeepSeek exhibit robust performance across prompt formats, indicating greater prompt resilience.

6 Conclusion

We introduce MILE-RefHumEval, a reference-free evaluation framework that eliminates the need for evaluator communication, addressing key flaws in existing LLM assessment methods. It uses multiple independently prompted LLMs to reduce bias, avoid inter-agent influence and better align with human judgment. Built for scalability, transparency, and robustness, it improves over reference-based and conversational evaluators, which often suffer from prompt sensitivity and flawed reasoning.

Experiments show that MILE-RefHumEval outperforms role-sharing and conversational baselines in accuracy, agreement, and efficiency. Increased evaluator independence enhances these results, while analysis highlights issues in traditional methods like order effects and reduced agreement in dialogue chains underscoring the benefits of isolated evaluation.

Looking ahead, MILE-RefHumEval enables adaptive strategies such as dynamic model weighting, domain-aware calibration, and interpretable scoring, supporting scalable and accountable evaluation for complex real-world tasks.

Limitations

Evaluator Generalization: Our framework assumes uniform evaluator reliability across tasks, yet LLMs often exhibit domain-specific capabilities. Without mechanisms for domain-aware weighting or specialization, our current design may miss contextual subtleties in judgment, limiting its generalization across varied evaluation domains.

Interaction Bias: The fixed turn order in our conversational protocols introduces directional bias, where early-stage outputs disproportionately shape subsequent evaluations. This effect is especially pronounced when weaker models initiate the sequence. While we quantify this cascading influence, we do not explore adaptive, randomized, or model-quality-aware turn ordering as potential mitigation strategies.

Metric Granularity: We rely on standard agreement metrics (e.g., Pearson correlation, Cohen's Kappa), which may fail to capture nuanced shifts in evaluative reasoning introduced through interaction. Richer analytical tools such as causal attribution methods, latent trajectory comparisons, or fine-grained human annotations are needed to trace how evaluators influence one another.

Prompt Sensitivity: We observe significant variation in model behavior across prompt formulations, particularly among smaller models like Phi and Mistral. This sensitivity challenges reproducibility and points to the need for prompt-invariant evaluation strategies or dynamic prompt calibration mechanisms to ensure fairness and stability.

References

Sher Badshah and Hassan Sajjad. 2024. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text. *arXiv preprint arXiv:2408.09235*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking foundation models with language-model-as-an-examiner. In *Advances in Neural Information Processing Systems*, volume 36, pages 78142–78167. Curran Associates, Inc.

Chaithanya Bandi, Hari Bandi, and Abir HARRASSE. 2025. Adversarial multi-agent evaluation of large language models through iterative debate.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *The Twelfth International Conference on Learning Representations (ICLR 2024)*.

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2025. Trust or escalate: LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*.

Sirui Liang, Baoli Zhang, Jun Zhao, and Kang Liu. 2024. Abseval: An agent-based framework for script evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12418–12434.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xinyi Liu, Pinxin Liu, and Hangfeng He. 2024. An empirical analysis on large language models in debate evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 470–487, Bangkok, Thailand. Association for Computational Linguistics.

Kun-Peng Ning, Shuo Yang, Yuyang Liu, Jia-Yu Yao, Zhenhui Liu, Yonghong Tian, Yibing Song, and Li Yuan. 2025. PiCO: Peer review in LLMs based on consistency optimization. In *The Thirteenth International Conference on Learning Representations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Bhrij Patel, Souradip Chakraborty, Wesley A. Suttle, Mengdi Wang, Amrit Singh Bedi, and Dinesh Manocha. 2024. AIME: AI system optimization via multiple LLM evaluators.

Lei Shu, Nevan Wichers, Liangchen Luo, Yun Zhu, Yinxiao Liu, Jindong Chen, and Lei Meng. 2024. Fusion-eval: Integrating assistant evaluators with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 225–238, Miami, Florida, US. Association for Computational Linguistics.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024.

Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Linjun Wang, Bowen Zhang, Xurui Zhang, Xiang Ren, and Kai-Wei Chang. 2024. Judgelm: Pairwise judgment of question-answering pairs using large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *ArXiv*, abs/2305.17926.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. 2023. Towards reasoning in large language models via multi-agent peer review collaboration. *arXiv preprint arXiv:2311.08152*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ziyou Zhang, Yuchen Duan, Yuanhe Tian, Luke Zettlemoyer, and Xi Victoria Liu. 2023. Debatesum: Evaluating summarization via structured debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. 2024a. Auto-arena: Automating llm evaluations with agent peer battles and committee discussions. *arXiv preprint arXiv:2405.20267*.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2024b. Recommender systems in the era of large language models (llms). *IEEE Trans. on Knowl. and Data Eng.*, 36(11):6889–6907.

A Appendices

Please evaluate the accuracy, clarity, and relevance of the answer generated for the following question. Provide percentage scores along with detailed explanations for the following criteria:

1. Answer Relevance:
- **Score:** [Provide score]
- **Explanation:** [Assess the clarity of the response. Is the answer easy to understand and logically structured? Are there any ambiguities or confusing phrases that could hinder comprehension?]

2. Depth and Completeness:
- **Score:** [Provide score]
- **Explanation:** [Evaluate the depth of the response. Does it provide a thorough explanation, or does it lack detail? Are all key aspects of the question addressed adequately?]

3. Grammar and Linguistic Accuracy:
- **Score:** [Provide score]
- **Explanation:** [Assess the grammatical correctness of the response. Are there any spelling or punctuation errors? Is the sentence structure correct and appropriate for the context?]

4. Contextual Appropriateness:
- **Score:** [Provide score]
- **Explanation:** [Evaluate how well the answer adapts to the specific question category (e.g., "generic" or "knowledge"). Does it use relevant terminology and concepts appropriate to the subject matter?]

5. Conciseness and Precision:
- **Score:** [Provide score]
- **Explanation:** [Evaluate whether the response is concise and to the point. Does the answer provide the necessary details without unnecessary elaboration or repetition?]

6. Creativity and Insight:
- **Score:** [Provide score]
- **Explanation:** [Assess the creativity of the response. Does it offer a unique perspective or innovative solution, especially when dealing with complex or thought-provoking questions?]

7. Bias and Fairness:
- **Score:** [Provide score]
- **Explanation:** [Evaluate the response for any signs of bias or unfairness. Does the answer exhibit neutrality, or does it reflect a particular perspective that may be considered biased or imbalanced?]

8. Knowledge Accuracy:
- **Score:** [Provide score]
- **Explanation:** [Assess the accuracy of factual information presented in the answer. Are the facts correct? Are any misconceptions, errors, or outdated information included?]

9. Adaptability to Various Categories:
- **Score:** [Provide score]
- **Explanation:** [Evaluate how well the answer adapts to different categories (e.g., "generic" vs. "knowledge"). Does the response adjust its tone, depth, and complexity based on the question category?]

10. Scalability of Responses:
- **Score:** [Provide score]
- **Explanation:** [Evaluate how well the model handles a variety of question types and complexities. Does the model handle simple and complex questions effectively, or does its performance degrade with more challenging queries?]

11. Overall Answer Quality:
- **Score:** [Provide score]
- **Explanation:** [Provide a final summary explanation for the overall score (Explain how it is calculated to show the calculation.) Consider how well all individual criteria performed. Offer insight into the general quality of the response based on the evaluation results.]

12. Error Detection and Correction:
Please analyze the provided response and identify any errors, inconsistencies, or areas where the response could be improved. The model should help by suggesting a corrected version of the output, focusing on the following points:
- Identifying any missing or inaccurate information.
- Correcting any grammatical or linguistic errors.
- Ensuring the relevance and completeness of the answer.
- Suggesting any improvements in clarity or coherence.
- Making the answer more concise and precise where necessary.

13. Corrected Version:
- **Updated Answer:** [Provide a corrected, revised version of the answer based on the analysis above.]

Figure 2: Evaluation prompt used in MILE-RefHumEval to assess LLM-generated answers. The prompt guides each evaluator to score responses across 12 dimensions, including relevance, completeness, grammar, factual accuracy, bias, and provide justification and revisions. This structured criteria ensures consistent, multi-faceted evaluation without human intervention.

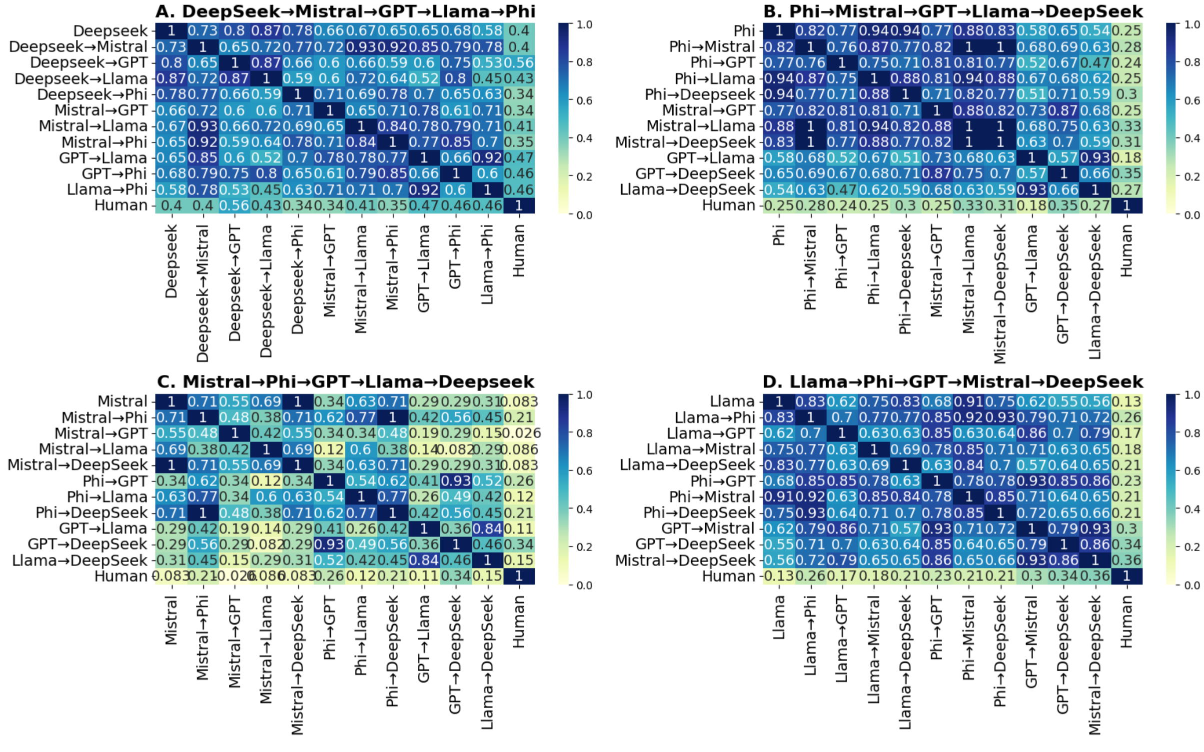


Figure 3: Heatmap showing Cohen's Kappa agreement scores between multiple LLM evaluators and human annotators under MILE-RefHumEval-Conv. The figure highlights varying levels of agreement depending on model order and interaction style, suggesting that conversational sequences introduce inconsistency.

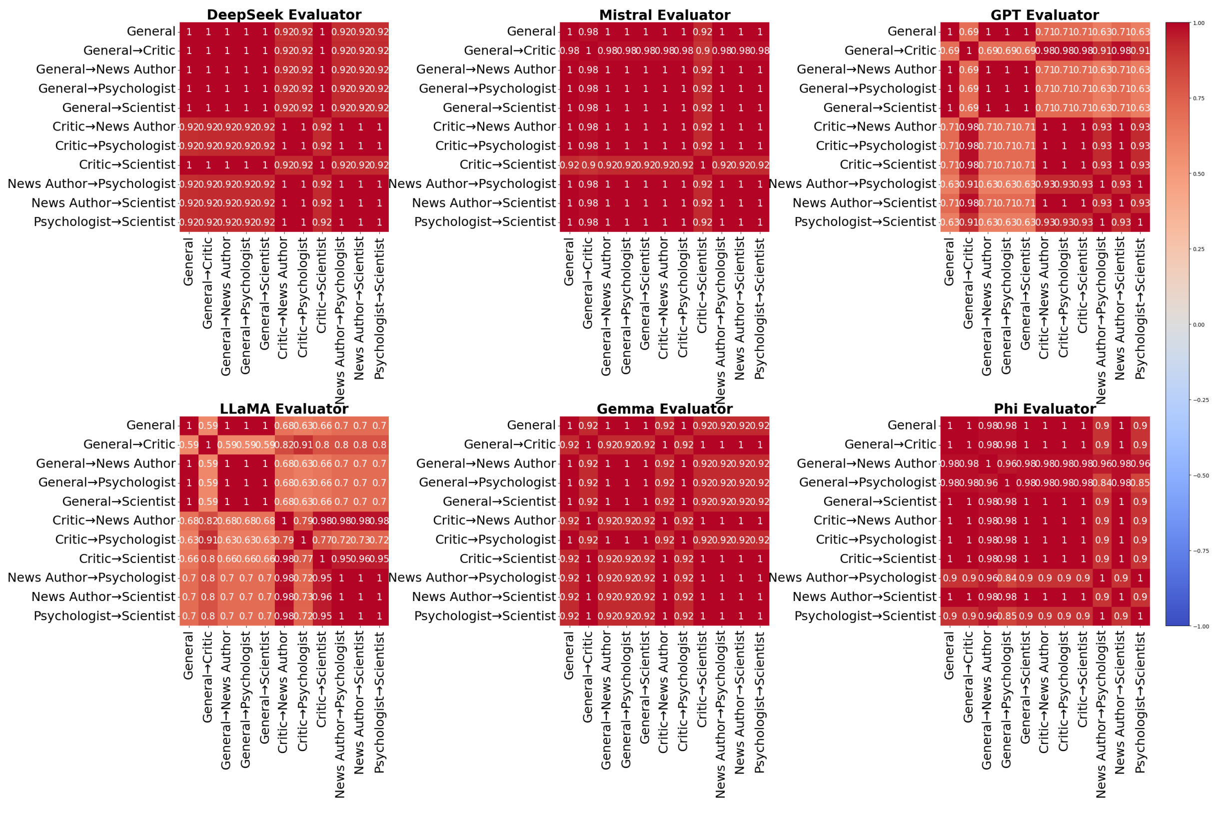


Figure 4: Pearson correlation of scores produced by a single evaluator (CHATEVAL) across different role-played personas. This figure demonstrates that repeating the same model across multiple roles can inflate agreement artificially, reducing diversity of judgment and interpretive value.

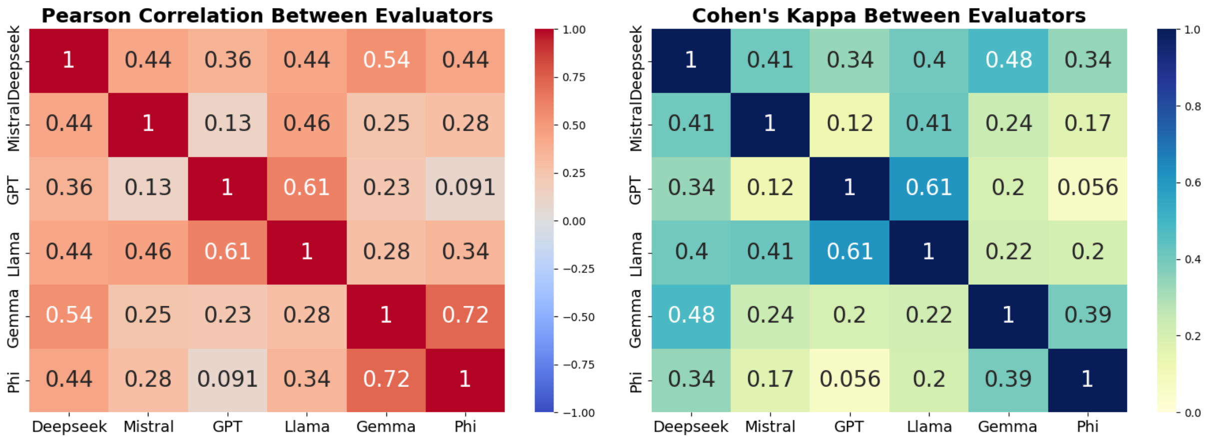
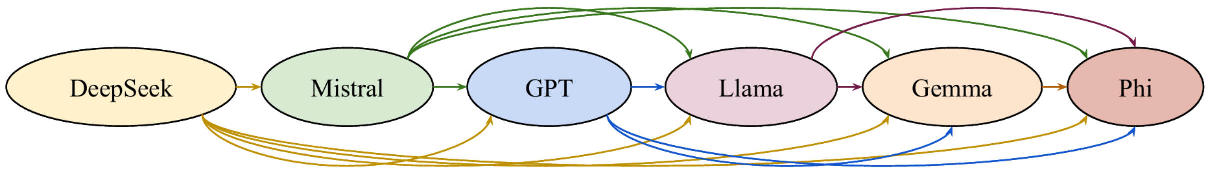


Figure 5: Inter-evaluator agreement analysis in MILE-RefHumEval, showing variance and correlation among independently operating LLMs. The figure emphasizes that non-conversational, diverse evaluators lead to broader reasoning coverage and stronger alignment with human assessments compared to role-based or interactive setups.


```
agent_name = """Critic Role"""
critic_role = """You are now Critic, one of the referees in this task. You will check fluent writing, clear
sentences, and good wording in summary writing. Your job is to question others judgment to make
sure their judgment is well-considered and offer an alternative solution if two responses are at the
same level."""

"""
[Question]
{question}
[The Start of Assistant 1's Answer]
{answer_gpt}
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
{answer_vicuna}
[The End of Assistant 2's Answer]
[System]
We would like to request your feedback on the performance of two AI assistants in response
to the user question displayed above.
Please consider the helpfulness, relevance, accuracy, and level of detail of their responses.
Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall
Performance. There are a few other referees assigned the same task, it's your responsibility to discuss with
them and think critically before you make your final judgment.
Here is your discussion history: \n
[The Start of General Public Role Feedback]
{chat_history_multiagent_general_public_role}
[The End of General Public Role Feedback] \n
{critic_role} \n
Now it's your time to talk, please make your talk short and clear, {agent_name} !
"""
```

Figure 6: Illustration of CHATEVAL, a baseline evaluation method where a single LLM evaluator sequentially adopts multiple roles (e.g., critic, scientist) to assess responses. This approach introduces potential role contamination and biases due to repeated model use across evaluative stages.




```
"""
You are Evaluator "B", assigned to assess the quality of two candidate answer comparison.
--- Question & Candidate Answer ---
 {str(question)}
[The Start of Assistant 1's Answer]
{answer_gpt}
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
{answer_vicuna}
[The End of Assistant 2's Answer]
Other evaluators are also working on this task. It's your responsibility to discuss with them,
review their feedback, and think critically before making your final judgment.
Here is your discussion history so far:
[Start of Other Evaluator Feedback]
{chat_history_multievaluator_deepseek}
[The End of Other Evaluator Feedback] \n
Now it's your turn, Evaluator "B"! Please consider the question and answer of two
candidates above.
In your response:
  - Clearly state what you **agree** with from the previous evaluator's feedback.
  - Highlight any points you **disagree** with and explain why.
  - Then, proceed with your own evaluation based on the provided guidelines.
{str(evaluation_prompt)}
"""
```

Figure 7: Diagram of MILE-RefHumEval-Conv, where multiple LLM evaluators engage in a conversation chain, each responding to the previous model's input. This method mimics human dialogue but risks accumulating bias from earlier turns, affecting independent judgment.

Question: How can I improve my time management skills?

Assistant 1's answer (GPT):

Here are some tips to improve your time management skills:
1. Create a schedule: Make a to-do list for the day, week or month and prioritize tasks by importance and deadline.
2. Set realistic goals: Be realistic about what you can accomplish in a given amount of time. Set goals that can be achieved but are also challenging.
3. Minimize distractions: Minimize distractions such as emails, phone calls, and social media notifications. Turn off notifications or use apps that block these distractions.
4. Take breaks: Taking a break can help you recharge and increase productivity. Take a break every hour or so to stretch, go for a walk, or listen to some music.
5. Learn to say 'no': Learn to say 'no' to tasks and requests that are not a priority or are not within your capabilities.
6. Delegate tasks: If possible, delegate tasks to others who are capable of completing them. This will help to reduce your workload.
7. Prioritize self-care: Make sure to take care of yourself by prioritizing self-care activities such as exercise, meditation or spending time with friends and family.
Remember, time management is a skill that takes time and practice to develop.

Assistant 2's answer (VICUNA):

Improving your time management skills can help you to be more productive, focused, and less stressed. Here are some tips to help you improve your time management skills:
1. Set clear goals: Identify what you want to achieve and set specific, measurable, and realistic goals. Prioritize your tasks and focus on the most important ones first.
2. Use a calendar or planner: Use a tool to organize your time and schedule your tasks. This can help you to stay on track and make the most of your time.
3. Break down tasks: Break down larger tasks into smaller, more manageable steps. This can help you to focus on one step at a time and make steady progress.
4. Eliminate distractions: Identify and eliminate distractions that may be preventing you from being productive. This can include turning off your phone, closing unnecessary tabs on your computer, or finding a quiet place to work.
5. Practice self-discipline: Use self-discipline to stay focused and avoid procrastination. Set a timer for each task and challenge yourself to complete it before the timer goes off.
6. Take breaks: Take regular breaks to rest and recharge. This can help you to stay focused and avoid burnout.
7. Review your progress: Regularly review your progress and adjust your strategy as needed. This can help you to stay on track and make improvements over time.

Figure 8: Assistant responses to the question "How can I improve my time management skills?" Two LLMs (GPT and VICUNA) provide advice independently. The figure displays these responses side-by-side to support qualitative comparison of tone, clarity, structure, and depth.

Candidates Evaluation using CHATEVAL: One Evaluator (DeepSeek), Many Roles (Cross-Role Agent Conversations)

- **General Public Role:** <think> Okay, so I need to figure out which of these two AI assistants did a better job answering the question about improving time management skills. Let me start by reading both responses carefully.Assistant 1's answer has seven tips: creating a schedule with prioritization, setting realistic goals, minimizing distractions, taking breaks, learning to say no, delegating tasks, and prioritizing self-care. It seems pretty comprehensive, covering both personal productivity and some strategic aspects like delegation and goal-setting.On the other hand, Assistant 2 also lists seven tips but they're slightly different. They mention setting clear goals, using a calendar or planner, breaking down tasks into smaller steps, eliminating distractions, practicing self-discipline with timers, taking breaks, and reviewing progress. This approach is more about structure and discipline, focusing on tools like planners and timers.I think both assistants are helpful, but I notice that Assistant 1 includes tips related to self-care and delegating, which might be more beneficial for someone managing multiple responsibilities or needing support. Meanwhile, Assistant 2's emphasis on using tools and setting clear goals could be very practical for someone looking to organize their schedule effectively.
However, considering the content, both are quite similar in structure but maybe differ slightly in depth. I don't see any major inaccuracies in either response. They both cover essential time management strategies, so it's a bit tricky to choose one over the other.
But if I have to pick the better one, perhaps Assistant 1 is slightly more comprehensive because they include self-care and delegating, which are important aspects that can help in long-term productivity and reducing stress. Alternatively, maybe Assistant 2's use of specific tools like planners makes their advice more actionable.
Wait, but both are equally good in terms of structure and relevance. Maybe it's a tie? Or perhaps one is better than the other based on certain criteria. Since I have to choose one, I might lean towards Assistant 1 because they include self-care, which is crucial for maintaining motivation and reducing burnout. </think>
  - Final Score for Assistant 1: 9
  - Final Score for Assistant 2: 8.5
  - Final Best Assistant: Assistant 1
- **General Public Role to Critic Role:**
  - Final Score for Assistant 1: 9
  - Final Score for Assistant 2: 8.5
  - Final Best Assistant: Assistant 1
- **General Public Role to News Author Role:**
  - Final Score for Assistant 1: 9
  - Final Score for Assistant 2: 8.5
  - Final Best Assistant: Assistant 1
- **General Public Role to Psychologist Role:**
  - Final Score for Assistant 1: 9
  - Final Score for Assistant 2: 8.5
  - Final Best Assistant: Assistant 1
- **General Public Role to Scientist Role:**
  - Final Score for Assistant 1: 9
  - Final Score for Assistant 2: 8.5
  - Final Best Assistant: Assistant 1

Figure 9: Evaluation of assistant responses using CHATEVAL (one LLM acting across multiple roles). Each role (e.g., Critic, Psychologist) scores the same responses using a shared criteria. The figure highlights score consistency across roles but potential redundancy due to a single model's repeated use.

Candidates Evaluation using MILE-RefHumEval-Conv: (One Role, Many Evaluators (With Conversations))

- **DeepSeek:** <think> Okay, so I need to evaluate the answers from two candidates for the question How can I improve my time management skills? The evaluation criteria are quite detailed with 12 points each, but I'll try to break it down step by step.First, let's look at Assistant 1's answer. It starts by listing tips like creating a schedule and setting realistic goals. Then it goes on to minimize distractions, take breaks, say no, delegate tasks, and prioritize self-care. The answer seems structured with clear points, each starting with a number. It ends with a note that time management takes practice.Now, Assistant 2's answer is similar but phrased differently. They talk about setting clear goals, using a calendar, breaking down tasks, eliminating distractions, practicing self-discipline, taking breaks, reviewing progress, and being adaptable. Their structure is also numbered, and they emphasize the importance of regular reviews.Starting with **Answer Relevance**: Both answers are directly addressing how to improve time management, so they're very relevant. No off-topic details here. I'll give both a high score, maybe 95 each. **Clarity and Coherence**: Assistant 1's answer is clear but a bit repetitive in structure. Each point is straightforward, but it might lack some depth. Assistant 2's answer is also clear but more concise. They use similar points but present them differently. I think both are coherent, so another high score, maybe 90 each. **Depth and Completeness**: Both answers cover common time management tips, which is good. However, they don't go into much detail beyond the basics. Assistant 1 mentions taking breaks and self-care, while Assistant 2 adds using a calendar and breaking down tasks. They both miss some aspects like prioritizing tasks more effectively or dealing with unexpected obstacles. I'll give them an 85 each. **Grammar and Linguistic Accuracy**: Both answers are grammatically correct. No spelling errors noticed. The sentences are clear and appropriate for the context. Maybe a solid 90 each. **Contextual Appropriateness**: Since the question is about time management, both answers use relevant terminology without unnecessary jargon. They're appropriate for the category. I'll score this as 85 each. **Conciseness and Precision**: Assistant 1's answer is concise but could be more precise by elaborating on each point a bit. Assistant 2's answer is also concise but perhaps too brief in some areas. Both are pretty good, so maybe 90 each. **Creativity and Insight**: Neither answer provides unique perspectives; they stick to common advice. There's no innovative solution here. I'll give them both 75. **Bias and Fairness**: No signs of bias. Both answers are neutral and balanced. So, 100 each. **Knowledge Accuracy**: The tips provided are accurate for time management. No errors detected. Both get 100. **Adaptability to Various Categories**: Both answers are suitable for the question category. They don't need to change much, so 90 each. **Scalability of Responses**: Both handle simple questions well but might struggle with more complex ones. Maybe a score of 75 each. **Overall Answer Quality**: Considering all criteria, both are good but not excellent. I'll give them an average score, say 80 each. </think> - Final Score for Assistant 1: 80 - Final Score for Assistant 2: 80 - Final Best Assistant: Equal

- **DeepSeek to Mistral: Agree**:- I agree with the other evaluator's assessment that both answers are relevant and neutral.- Both answers score high in grammar and linguistic accuracy. **Disagree**:- I disagree slightly with the conciseness and precision scores. Assistant 1's answer is a bit repetitive, and Assistant 2's answer could benefit from more elaboration on some points.- I also disagree with the scalability score; both answers handle simple questions well but might struggle with more complex queries. 1. **Answer Relevance**:- **Score**: 95- **Explanation**: Both answers directly address how to improve time management skills and stay on topic without any off-topic details. 2. **Clarity and Coherence**:- **Score**: 85- **Explanation**: Both answers are clear and logically structured, but Assistant 1's answer is slightly repetitive in its structure. 3. **Depth and Completeness**:- **Score**: 80- Explanation**: Both answers cover common time management tips but lack depth in certain areas. They miss aspects like more effective prioritization techniques or handling unexpected obstacles. 4. **Grammar and Linguistic Accuracy**:- **Score**: 90- **Explanation**: Both answers are grammatically correct with no spelling errors. The sentences are clear and appropriate for the context. 5. **Contextual Appropriateness**:- **Score**: 85- **Explanation**: Both answers use relevant terminology and concepts appropriate to the subject matter of time management. 6. **Conciseness and Precision**:- **Score**: 85- **Explanation**: Assistant 1's answer is concise but could be more precise by elaborating on each point. Assistant 2's answer is concise but too brief in some areas. 7. **Creativity and Insight**:- **Score**: 70- **Explanation**: Neither answer provides unique perspectives; they stick to common advice without offering innovative solutions. 8. **Bias and Fairness**:- **Score**: 100- **Explanation**: Both answers are neutral and balanced, exhibiting no signs of bias. 9. **Knowledge Accuracy**:- **Score**: 100- **Explanation**: The tips provided in both answers are accurate for time management with no errors detected. 10. **Adaptability to Various Categories**:- **Score**: 90- **Explanation**: Both answers are suitable for the question category and use relevant terminology without unnecessary jargon. 11. **Scalability of Responses**:- **Score**: 70- **Explanation**: Both answers handle simple questions well but might struggle with more complex queries that require deeper insights or more nuanced solutions.12. **Overall Answer Quality**:- **Score**: 82- **Explanation**: The overall quality is good, but not excellent. Considering all criteria, the average score is calculated as follows: (95 + 85 + 80 + 90 + 85 + 85 + 70 + 100 + 100 + 90 + 70) / 12 = 84.16, rounded to 84.- Final Score for Assistant 1: 84- Final Score for Assistant 2: 84 - Final Best Assistant: Equal

Figure 10: Assistant response evaluations using MILE-RefHumEval-Conv, where multiple LLMs (e.g., DeepSeek, Mistral) sequentially critique the same responses in a conversational format. Each evaluator applies the 12-criteria and builds on prior outputs, revealing strengths and weaknesses of multi-step dialogic assessment.

Candidates Evaluation using MILE-RefHumEval: (One Role, Many Evaluators (No Conversations))

- **DeepSeek:** <think> Okay, so I need to evaluate the answers from two assistants for the question How can I improve my time management skills? The evaluation criteria are quite detailed with 12 points each, but I'll try to break it down step by step.First, let's look at Assistant 1. Their answer lists seven tips: creating a schedule, setting realistic goals, minimizing distractions, taking breaks, learning to say no, delegating tasks, and prioritizing self-care. Each point is clear and directly addresses time management. The structure is logical, starting from planning to self-care. It's concise without unnecessary details.Now, Assistant 2 also provides seven tips: setting clear goals, using a calendar, breaking down tasks, eliminating distractions, practicing self-discipline, taking breaks, and reviewing progress. Their answer is similarly structured but includes slightly different points, like using a tool or breaking tasks into steps.Comparing both, they cover similar ground but with slight variations in their strategies. Both **answers are relevant** because they directly address the question about improving time management. The **clarity and coherence** are good; each point is easy to understand and logically ordered.Looking at **depth**, both assistants provide enough detail without being too verbose. They include practical tips that can be applied. **Grammatical accuracy** seems solid with no obvious errors. **Contextually**, both answers use appropriate terminology for the subject.In terms of **conciseness**, both are well-structured but could perhaps add a bit more on how to implement each tip. **Creativity-wise**, they stick to common strategies which is good for time management advice. There's no bias detected; both answers are neutral and helpful.**Knowledge accuracy** seems spot-on as these are standard tips in time management literature. Both answers adapt well to the question category by providing actionable advice. They handle simple questions effectively but might struggle with more complex ones, though that's not evident here.Overall, both assistants did a good job. Assistant 1's answer is slightly more comprehensive with additional points like delegating and learning to say no, which adds depth. Assistant 2 includes practical steps like using a tool and reviewing progress, which are also valuable. I think they're on par, so the best assistant would be equal. </think> - Final Score for Assistant 1: 90- Final Score for Assistant 2: 90- Final Best Assistant: Equal

- **Mistral:** 1. ****Answer Relevance****:- **Score**: 95- **Explanation**: Both answers directly address the question by providing tips to improve time management skills. There are no off-topic details included in either response. 2. ****Clarity and Coherence****:- **Score**: 90- **Explanation**: Both responses are easy to understand and logically structured. Assistant 1's answer is slightly more conversational, while Assistant 2's answer is more formal but equally clear. 3. ****Depth and Completeness****:- **Score**: 85 (Assistant 1) / 90 (Assistant 2)- **Explanation**: Assistant 2 provides a slightly more detailed explanation by including the concept of SMART goals and breaking down tasks into smaller steps. Assistant 1's answer is also comprehensive but lacks this level of detail. 4. ****Grammar and Linguistic Accuracy****:- **Score**: 95 (both)- **Explanation**: Both answers are grammatically correct with no spelling or punctuation errors. The sentence structure is appropriate for the context. 5. ****Contextual Appropriateness****:- **Score**: 90 (both)- **Explanation**: Both responses use relevant terminology and concepts appropriate to the subject matter of time management. 6. ****Conciseness and Precision****:- **Score**: 85 (Assistant 1) / 90 (Assistant 2)- **Explanation**: Assistant 2's answer is slightly more concise and to the point, while Assistant 1 includes a few extra details that are still relevant. 7. ****Creativity and Insight****:- **Score**: 80 (both)- **Explanation**: Both answers offer standard but effective tips for time management. Neither provides a unique perspective or innovative solution. 8. ****Bias and Fairness****:- **Score**: 100 (both)- **Explanation**: Both responses exhibit neutrality and do not reflect any particular bias or imbalanced perspective. 9. ****Knowledge Accuracy****:- **Score**: 95 (both)- **Explanation**: Both answers provide accurate factual information about time management strategies. There are no misconceptions or errors included. 10. ****Adaptability to Various Categories****:- **Score**: 90 (both)- **Explanation**: Both responses adapt well to the generic question category, providing practical advice suitable for a broad audience. 11. ****Scalability of Responses****:- **Score**: 85 (both)- **Explanation**: Both models handle this simple and straightforward question effectively. Their performance would need to be tested with more complex queries to fully evaluate scalability. 12. ****Overall Answer Quality****:- **Final Score for Assistant 1**: 87- **Final Score for Assistant 2**: 90- **Final Best Assistant**: Assistant 2

Figure 11: Assistant response evaluations using MILE-RefHumEval. Multiple independent LLMs assess the same candidate responses without communicating with each other. This configuration demonstrates high agreement with human ratings and improved reliability due to evaluator diversity and isolation.

Post-Evaluation: Error Analysis and Guided Refinement using MILE-RefHumEval

DeepSeek:

Error Detection and Correction:- Identified Errors: Lacked actionable steps and detailed explanations for each tip. - **Suggested Improvements:** Expand on each tip with specific methods or examples to enhance depth and practicality. **Corrected Version:** Both answers could be improved by adding more detailed explanations and actionable steps for each tip, ensuring a comprehensive guide that is both informative and practical.

Mistral:

****Error Detection and Correction**-**Identified Errors**:**- Lacked actionable steps and detailed explanations for each tip.****Suggested Improvements**:**- Expand on each tip with specific methods or examples to enhance depth and practicality.****Corrected Version**:****[Assistant 1's Corrected Answer]**Here are some tips to improve your time management skills: 1. Create a schedule: Make a to-do list for the day, week or month and prioritize tasks by importance and deadline. Use tools like Trello or Asana to organize your tasks visually. 2. Set realistic goals: Be realistic about what you can accomplish in a given amount of time. Use the SMART criteria (Specific, Measurable, Achievable, Relevant, Time-bound) to set goals that are challenging yet attainable. 3. Minimize distractions: Minimize distractions such as emails, phone calls, and social media notifications. Turn off notifications or use apps like Forest or FocusMe to block these distractions during focused work sessions. 4. Take breaks: Taking a break can help you recharge and increase productivity. Use techniques like the Pomodoro Technique, where you work for 25 minutes, then take a 5-minute break. 5. Learn to say 'no': Learn to say 'no' to tasks and requests that are not a priority or are not within your capabilities. Communicate politely but firmly about your limitations. 6. Delegate tasks: If possible, delegate tasks to others who are capable of completing them. Use delegation tools like Slack or Microsoft Teams to assign tasks and track progress. 7. Prioritize self-care: Make sure to take care of yourself by prioritizing self-care activities such as exercise, meditation, or spending time with friends and family. Schedule self-care time in your calendar to ensure it doesn't get overlooked.Remember, time management is a skill that takes time and practice to develop.**[Assistant 2's Corrected Answer]**Improving your time management skills can help you to be more productive, focused, and less stressed. Here are some tips to help you improve your time management skills: 1. Set clear goals: Identify what you want to achieve and set specific, measurable, and realistic goals. Use the GOKR method (Goals, Outcomes, Key Results) to define your objectives clearly. 2. Use a calendar or planner: Use a tool like Google Calendar or Microsoft Outlook to organize your time and schedule your tasks. Set reminders for important deadlines. 3. Break down tasks: Break down larger tasks into smaller, more manageable steps. Use mind mapping tools like MindMeister to visualize and break down complex projects. 4. Eliminate distractions: Identify and eliminate distractions that may be preventing you from being productive. This can include turning off your phone, closing unnecessary tabs on your computer, or finding a quiet place to work using apps like Noise Canceling or Soundtrap.5. Practice self-discipline: Use self-discipline to stay focused and avoid procrastination. Set a timer for each task using tools like Time Timer and challenge yourself to complete it before the timer goes off. 6. Take breaks: Take regular breaks to rest and recharge. Use the 20/20/20 rule: every 20 minutes, take a 20-second micro-break to look at something 20 feet away.7. Review your progress: Regularly review your progress and adjust your strategy as needed. Use journaling or reflection tools like Day One to track your progress and make improvements over time.

Figure 12: Illustration of MILE-RefHumEval's error detection and refinement process. Each LLM evaluator independently identifies weaknesses in the candidate answer and proposes corrections based on a 12-point criteria. The figure shows improved versions, highlighting the framework's ability to enhance response quality without human input.