Evaluating the Evaluator:
Measuring LLMs' Adherence to Task Evaluation Instructions

Bhuvanashree Murugadoss, Nick McKenna, Christian Poelitz,
Ian Drosos, Vu Le, Carina Suzana Negreanu, Chris Parnin, Advait Sarkar

Microsoft

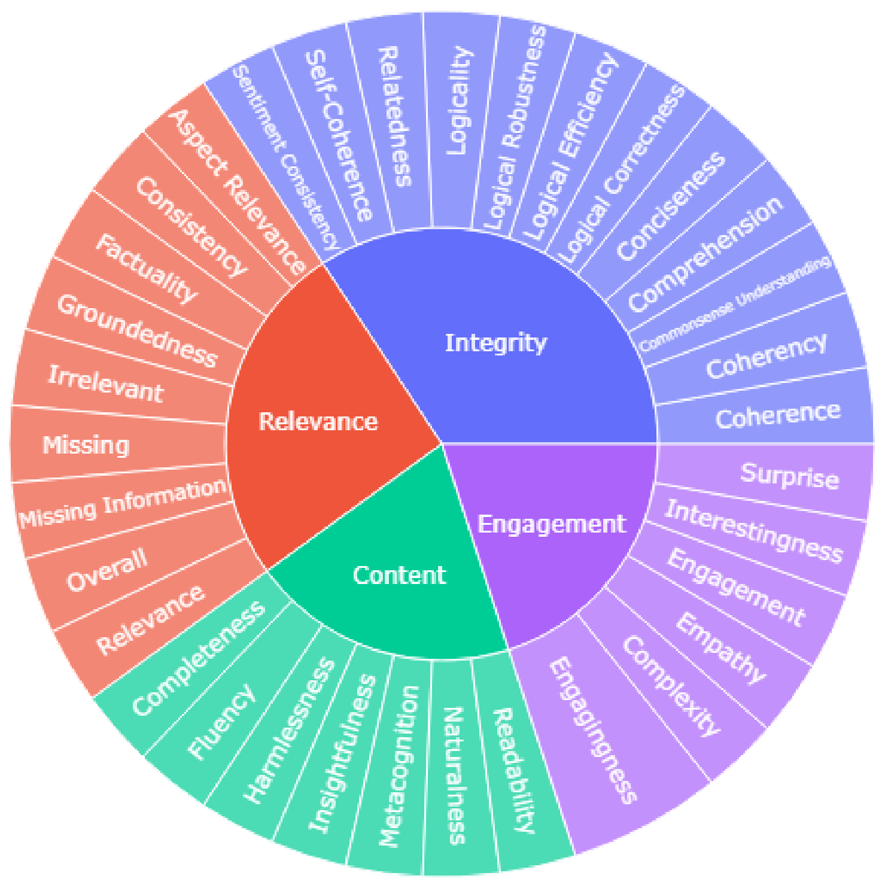
- ❖ Can LLMs replace human judgment in evaluating tasks?
- ❖ Models like GPT-4 and Llama 3, trained with Reinforcement Learning from Human Feedback (RLHF), align closely with human preferences when assessing qualities like text coherence. But are the evaluations truly reflecting the prompts they're given, or are they more influenced by their training on high-quality data?
- ❖ Our findings show the influence of prompting on AI alignment, measured by Pearson correlation, with human judgments and provide valuable resources for enhancing automatic evaluations using LLMs.

LLM-based Evaluation

- ❖ We developed a novel taxonomy categorizing qualitative evaluation criteria into four main groups: **Content, Relevance, Integrity, and Engagement**
  - o 34 specific metrics from 8 SOTA benchmarks
- ❖ We examine how LLMs-as-a-judge respond to prompts with varying levels of instruction detail
  - o These ranged from generic prompts with minimal guidance to detailed rubrics specifying exact evaluation criteria

Criteria taxonomy

- ❖ **Content-based criteria:** Measure how well the solution is presented to the user.
- ❖ **Engagement-based criteria:** Measure how engaging the solution is.
- ❖ **Integrity-based criteria:** Measure how consistent and logical coherent the solution is.
- ❖ **Relevance-based criteria:** Measure how relevant the solution is.

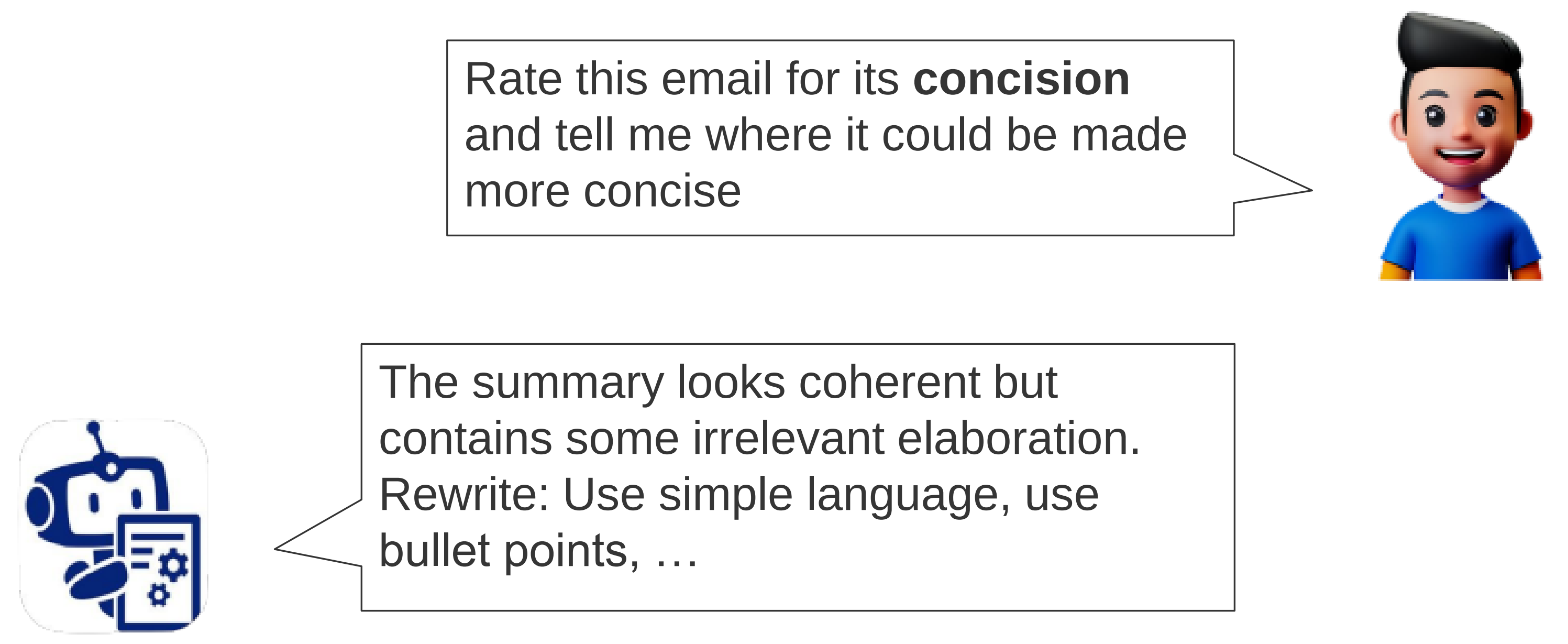


Experiments: Prompting Level of Detail

We evaluate the LLMs as a judge using four different prompting strategies, reflecting different amounts of instructions.

1. **Perplexity (no prompt):** We score each task solution by the perplexity under the corresponding LLM, as measure unbiased by any prompting.
2. **Generic quality prompt:** We prompt each LLM with a basic instruction to measure the quality of the task solution but give no specific criteria or instructions.
3. **Criteria specific prompt:** We prompt each LLM with an instruction to measure the quality for a specific criteria but with no instructions how.
4. **Full rubric prompt:** We prompt each LLM with scoring instructions to measure the quality for a specific quality criterion.

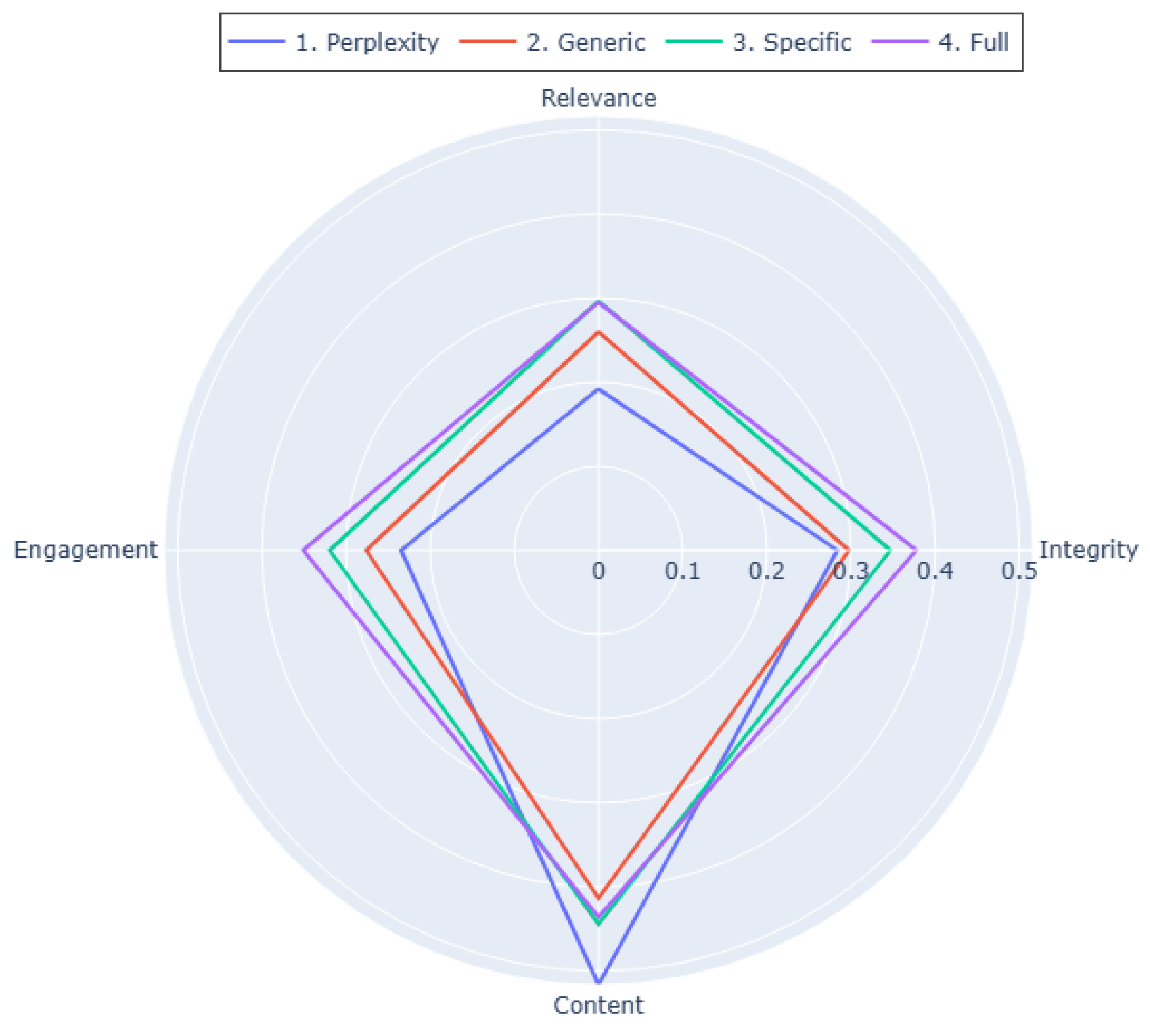
Judging the quality of AI generated responses in human-AI interactions is challenging. Using another LLM to "judge" the quality and propose corrections has potential for greater automatic resource development


Rate this email for its **concision** and tell me where it could be made more concise

The summary looks coherent but contains some irrelevant elaboration. Rewrite: Use simple language, use bullet points, …

Main findings

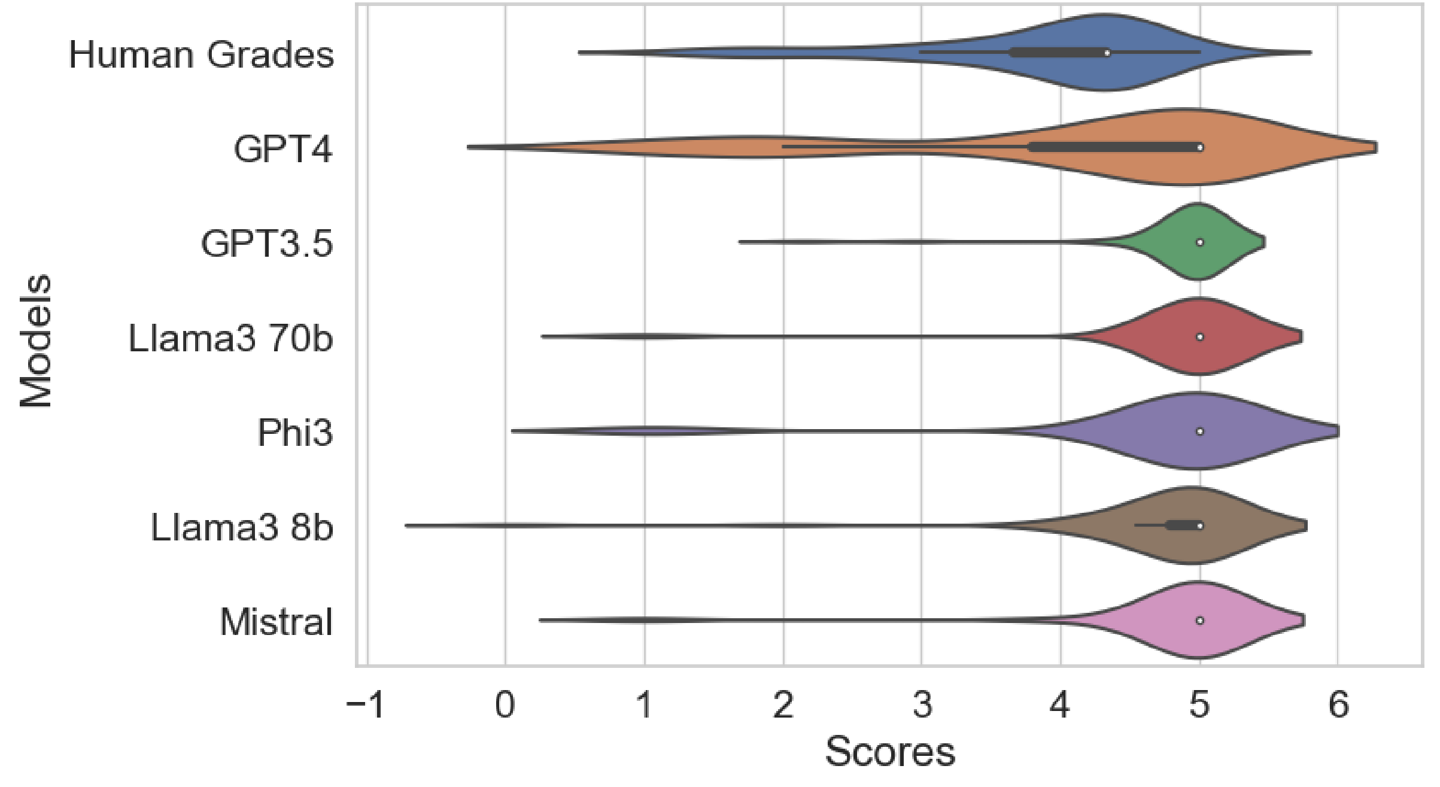
1. Detailed quality criteria information might not be necessary in the most powerful models; for instance, GPT-4shows a high level of agreement with human judgements even without detailed instruction.



2. Simple perplexity values are very effective at estimating textual quality, often outperforming the results of prompting the LLMs-as-a-judge with basic instructions.

	Perplexity	Generic	Specific	Full
harmlessness	**1.000**	0.486	0.749	0.928
completeness	**0.764**	0.708	0.707	0.707
readability	**0.731**	0.256	0.329	0.341
fluency	**0.430**	0.310	0.314	0.312
metacognition	0.374	**0.454**	0.436	0.330
insightfulness	0.266	0.327	**0.444**	0.339
naturalness	0.274	0.454	0.458	**0.466**

3. Judging task-specific quality criteria like relevance or logical correctness requires more capable, larger models, aligning with previous research on necessary model capabilities for critiquing.



Conclusion

- ❖ LLMs-as-a-judge do not always adhere to task evaluation instructions.
- ❖ While LLMs strongly align with general human preferences, adding detailed evaluation instructions in prompts offers limited advantages.
- ❖ Furthermore, model perplexity is an alternative evaluation method for text quality criteria without the need for any prompt engineering.