LLM-as-a-Judge Failures at Automating the Identification of Poor Quality Outputs in Free-Form Texts

Zongxia Li[1] **Xiyang Wu**[1] **Ishani Mondal**[1]
Alexia Siu[2] **Jordan Boyd-Graber**[2] **Ani Nenkova**[2]
University of Maryland[1] Adobe[2]
{zli12321,wuxiyang,imondal,jbg}@cs.umd.edu asiu,nenkova.adobe.com

Abstract

Large language models (LLM) such as GPT-4, Claude, and LLaMA are routinely used to evaluate long-form text generated by language models. We study the ability of these models to identify low-quality texts, an increasingly rare subset of output that is of great interest to pinpoint during development. We present experiments with a panel of LLM judges, and crowd-sourced approximations of reference judgments. Pinpointing sub-par outputs is a difficult task for both models and crowdworkers, with models doing better overall. Moreover, unlike findings in prior work on factoid QA, panels of cheaper models do not agree as well with high-quality developer judgments of low quality as panels of frontier models. We present both qualitative and quantitative analyses of the relative strengths of models in the panel, gleaning insights into why they yield better results over a single model.

1 Introduction

The outputs of LLMs such as e.g. GPT-4 (Bubeck et al., 2023), Claude (Anthropic, 2024) are used by millions but the evaluation of their long-form text responses pose an outstanding research challenge (Xu et al., 2023). As LLMs produce texts with impressive overall fluency, accuracy, and usefulness, developers of user experiences powered by the models are interested in detecting the rare bad quality generations (Chen et al., 2021).

Aiming to identify poor outputs invites a rethinking of the criteria for automated evaluation. Human agreement rates like accuracy can be misleading since they reflect the abundant good answers rather than the ability to detect poor ones (Section 3.3). We use F_1 measure for the class of bad outputs (i.e. texts that experts would rate as 1, 2, or 3 on a five-point quality scale). We take developer judgments as the GOLD STANDARD for evaluation and study how successfully these can be replaced by LLMs or crowd-sourcing. We expand on the growing body of research on LLMs as evaluators (Yuan et al., 2021; Kamalloo et al., 2023; Li et al., 2024b; Kim et al., 2024; Chiang and yi Lee, 2023) and their accompanying challenges (Abeysinghe and Circi, 2024), and are the first to focus on the detection of bad outputs. We carry out our experiments on SummEval (Fabbri et al., 2021) summaries paired with developer and crowd-sourced quality ratings. These fine-grained scores are important in LLM development to exclude unacceptable outputs and guide the model to ideal responses (Bai et al., 2022).

We build upon prior work from factoid question answering showing that a panel of multiple cheaper models better aligns with all human judgments than a single frontier model does (Verga et al., 2024). Our results on long-text evaluation in contrast show that combining frontier models improves the ability to detect poor outputs, but not necessarily for weaker models (Section 4).

Further, we seek to understand how a panel of judges can resolve some of the challenging cases that a single model cannot handle correctly. Our quantitative and qualitative analysis show that *(i)* different models have distinct weaknesses in evaluating different response aspects but combining multiple strong models helps mitigate individual limitations; *(ii)* a panel of LLMs enhances the detection of challenging unacceptable responses while maintaining a stable and accurate evaluation of common acceptable cases. *(iii)* no evaluation is perfect, even 'high quality' human ratings have errors.

2 From Classical to LLM Evaluation

As modern LLMs handle complex QA tasks beyond entity extraction or short responses (Brown et al., 2020), evaluating their outputs requires reasoning over the responses and is challenging compared to that of earlier extractive QA (Rajpurkar et al., 2016;

Clark et al., 2018; Wang et al., 2019). Classical metrics like F_1 and ROUGE (Lin, 2004), which rely on gold answers and measure word overlap, were commonly used in the past to evaluate longer responses such as summarization (Eyal et al., 2019), long-form QA (Fan et al., 2019), machine translation (Papineni et al., 2002). However, these metrics, despite performing okay on datasets like NarrativeQA (Kočiský et al., 2018), require predefined thresholds and gold answers. They are no longer a good fit for long-form text evaluation (Chen et al., 2019; Zheng et al., 2023).

Using LLMs for automatic evaluations. LLMs align better with human evaluations (Chern et al., 2024; Yuan et al., 2021; Zhong et al., 2022), and a number of studies have identified GPT-4 as the single best automatic evaluator (Vu et al., 2024; Kim et al., 2024). LLM evaluators are more flexible. They accommodate various evaluation tasks via prompt instructions, including reference-free scoring (Chen et al., 2023), chain-of-thought scoring (Chiang, 2023), and pairwise preference assessments (Liu et al., 2024). Although LLM-as-a-judge is gaining popularity (Gao et al., 2024; Laskar et al., 2024), Huang and Zhang (2024) show that most existing benchmarks and evaluations still rely on classical evaluations, such as exact-match, binary *yes/no* responses, or multiple-choice formats, with only one benchmark incorporating long, open-ended answers that use LLMs to evaluate. Further analysis of best practices and concerns are needed to trust reliability for automated LLM evaluations (Li et al., 2024a).

3 Experiments

3.1 Dataset

SummEval is the largest summarization benchmark, covering 1,600 news articles from CNN and DailyMail (Fabbri et al., 2021). The summaries are produced by seven extractive and sixteen abstractive models. The extractive models such as NEUSUM (Zhou et al., 2018) and REFRESH (Narayan et al., 2018) score sentences, then concatenate the highest-scoring sentences into a summary. The dataset provides a source document, a reference summary, and a model-generated summary, along with quality ratings from four crowdworkers and three developers of summarization systems, who had written papers on summarization either for academic conferences or as part of a senior thesis. The ratings use a 5-point scale (1=poor quality, 5=excellent) to evaluate four aspects (Kryscinski et al., 2019; Dang, 2005, Table 1).

Coherence: The summary should be well-structured and organized.
Consistency: The summary only contains statements that are entailed by the source document.
Fluency: The summary should have no formatting problems, capitalization errors, or obviously ungrammatical sentences.
Relevance: The summary should include only important information from the source document without excess redundant information.

Table 1: SummEval adopts the NIST quality aspects (Kryscinski et al., 2019; Dang, 2005). Our goal is to assess how well LLMs and crowdworkers identify problematic responses across these aspects.

3.2 Setting Up Automatic Evaluations

LLM-as-a-Judge prompts an LLM with evaluation instructions and rubrics and a text to evaluate. The quality of model generations and instruction following abilities vary by model sizes, pretraining data, and alignment data and steps (Kaplan et al., 2020). We select a range of models with varying sizes–GPT-3.5-turbo, GPT-4o-mini, GPT-4 (Bubeck et al., 2023), Claude-Sonnet (Anthropic, 2024), Gemini-pro (DeepMind, 2024), LLaMA 3-8B, LLaMA 3-70B (et al, 2024), Mistral 8x7b (Jiang et al., 2024), and Prometheus 2 (Kim et al., 2024).[1] We adopt the *analyze then rate* prompt format proposed by Chiang (2023). This approach describes the evaluation criteria (the same given to the human judges), the source news article, and the summary. We then prompt LLMs to generate evaluation justification and ratings using prompt templates provided in Table 5.

Automation Success Criteria. We distinguish two output types: *acceptable* (summaries rated 4 or 5) and *unacceptable* (rated 1 to 3). Across evaluations by developers, crowdworkers, or LLMs, bad-quality outputs are the minority (Figure 1).[2]

We turn to a comparison of other LLMs as judges (Table 2). Using *any* LLMs results in overall better F_1 than crowdsourcing. GPT-4 is always among the best models but depending on the quality aspect, it is not consistently better than others. LLaMA-70B is better or equivalent to GPT-4 in spotting problems with coherence, consistency, and relevance. Prometeus-2, which has been fine-tuned on synthetic data specifically for evaluation does very similarly to crowdsourcing. As GPT-4 represents

[1]Curiously, Gemini appears not to have been aligned for evaluation and produces invalid fraction ratings.

[2]Exceptions exist for weak models, where low ratings dominate (Appendix A.2).

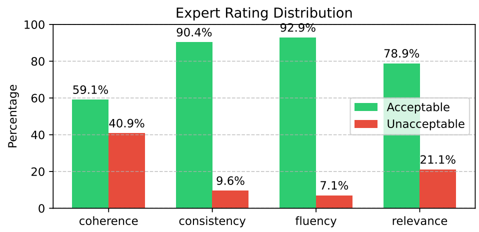


Figure 1: Poor quality outputs are rare compared to good ones. Identifying the problematic outputs can facilitate error analysis and speed up development (Distribution for crowdworkers and all models in Appendix A.2).

one of the best evaluators, we use it as an example to analyze failure modes in LLM evaluation.

Identifying rare unacceptable outputs is crucial for error analysis, yet both GPT-4 and crowdworkers show significantly lower precision and recall for unacceptable cases compared to acceptable ones (on average 0.4 and 0.6 lower respectively; Appendix 5). Using crowdsourcing to approximate developer judgments is not worth the effort; fully replacing human judgments with GPT-4 better reproduces the developer's identification of poor outputs. However, despite GPT-4 superiority over crowdsourcing, GPT-4 is still not good at identifying poor outputs, especially for short responses or responses directly copied from the original text (Section 4; Appendix 7 with examples).

GPT-4's precision in identifying summaries lacking coherence is fairly good, 80% but its recall is just under 50%. If used for error analysis during development, such recall will not be helpful as a human will have to look at the rest of the data to find all problematic outputs. Not doing that additional human inspection leads to the potential risk of underestimating the degree of the problem, as only half of the incoherent summaries will be surfaced.[3] Precision of GPT-4's ability to evaluate summary relevance, i.e. if it only contain major, summary worthy details, is under 44%, though its recall is almost 70%. GPT-4's ability to identify problems with consistency, also referred as factuality (Min et al., 2023) or hallucinations (Huang et al., 2025), is overall the best and most balanced (Figure 3), with precision of 65% and recall of 71% for the unacceptable class.

3.3 Single Model vs. Ensemble Evaluation

Verga et al. (2024) show that replacing a single best LLM (GPT-4) with multiple cheaper LLMs can improve automatic evaluations for short-form factoid question answering. However, the underlying reason why multiple models are better than a single model is not explored in that work, nor has the approach been tested on evaluation of long-form text. No single model wins on all evaluation aspects, but the top three evaluators are considerably better than other on at least one aspect (Table 2). In the next section, we analyze ensemble evaluations with all possible combinations of these models using our automation success criteria. We test ensemble size two to eight and determine their judgments using average scores, matching the expert and crowdworker rating method.

[3]The same pattern applies to all other models and crowdsource (Appendix 5).

4 LLMs Complement Each Other

We now study the effectiveness of a panel of judges compared to single-model evaluations, and the optimal number of models in the panel. We then conduct a qualitative analysis of common failure modes of the best three LLM and crowd evaluations.

Combining multiple models has higher alignment with developer judgments. We compute the Macro F_1 score between automatic and developer evaluations. Figure 2 shows the median and maximum scores for panels of different compositions. For the median score, when model selection is unknown, the shift from a single model to a three-model panel shows notable improvement, while adding more models provides minimal benefit. Adding more models increases cost but yields little improvement in evaluation.

For the maximum score, while using a single strong model might be adequate for specific evaluation tasks, Figure 2 shows that different models excel in different aspects. Different LLMs show strengths in different evaluation criteria, suggesting that relying on a single LLM for all aspects does not yield optimal results. For all size panels, GPT-4, LLaMA3-70B, and Claude were part of the panel achieving the maximum score. Combining these individually strong models yield the better results (coherence: 0.76; consistency: 0.84; fluency: 0.73; relevance: 0.76) than any of its constituents (individual best in Table 2). On the other hand, combining three weaker models can yield worse results than its best constituents (Result details in Appendix 5).

Qualitative Analysis. Next the authors analyze the source text, reference summary, model sum-

Aspect	Crowdworker	GPT-3-turbo	GPT4o-mini	GPT-4	Claude-sonnet	Mistral-8x7b	LLaMA3-8B	LLaMA3-70B	Prometheus
Coherence	0.50	0.70	0.60	0.70	0.68	0.45	0.58	**0.74**	0.64
Consistency	0.47	0.51	0.63	**0.82**	**0.82**	0.42	0.66	**0.82**	0.47
Fluency	0.49	0.47	0.61	**0.72**	0.62	0.57	0.60	0.51	0.45
Relevance	0.51	0.59	0.54	0.68	**0.71**	0.53	0.67	0.68	0.51

Table 2: The overall LLM F_1 scores are higher than crowdworkers. LLMs are generally more aligned with developers than crowdsource workers at evaluation.

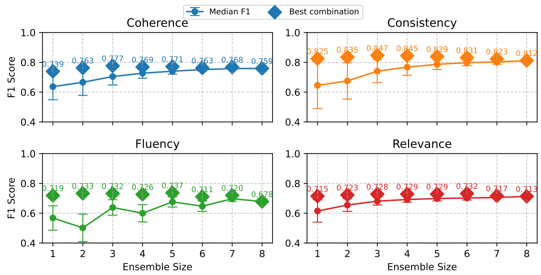


Figure 2: Increasing the ensemble size increases the overall median and best F_1 scores.

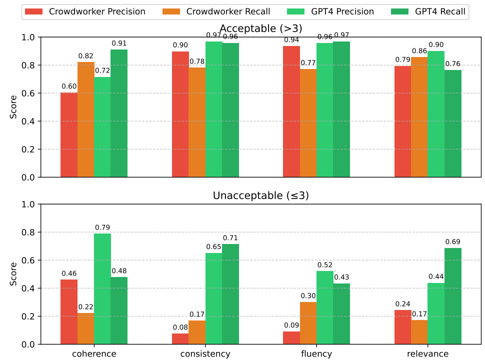


Figure 3: Being able to catch the unacceptable responses is more challenging than evaluating acceptable responses. Crowdsource workers and GPT-4 have much lower precision and recall at unacceptable responses than acceptable responses.

mary, and LLM score justification for scores where a strong ensemble (GPT-4, LLaMA3-70B, Claude) is better than the best individual model. These account for 21% coherence, 5% consistency, 30% fluency, and 18% relevance. We outline cases in which individual LLMs are likely to erroneously disagree with developers (fine-grained categorizations in Appendix 6).[4]

The two major errors for coherence appear to be confounding of length and a given property, and inability to focus only on the assigned quality criterion. Specifically, in the cases where a model (mostly GPT-4) in the panel disagrees with developers, the LLM justifications criticize capitalization and consistency, but developers judge them as coherent responses by the overall flow of the sentences and are less concerned on minor punctuation errors (11%).

If the response contains statements contradicting the source text, even though the response is coherent, LLMs penalize the response's coherence and fluency (8%), while neither experts nor crowdworkers do. LLMs often evaluate factors beyond the assigned criteria, such as assessing faithfulness when only fluency evaluation is required (4%). On the other hand, 5% of the examples of fluency that have spelling errors were judged as acceptable by the developers, while models accurately flag them as potentially problematic when shown to a reader (See Appendix 13 for examples). Given these observations, holistic evaluations of the overall quality of the output are likely to be more successful than attempts to evaluate fixed aspects.

Panels correct individual model weaknesses. LLaMA3-70B tends to rate longer texts as irrelevant: 9% of the reviewed examples rated by experts and other LLMs as relevant but LLaMA rated as poor are longer summaries, 30 tokens longer than an average summary. Claude on the other hand evaluates beyond its assigned criteria for consistency. Often summaries are consistent with the source article, but Claude rates them low, noting that the summary lacks a central theme sentence (6%). GPT-4 diverges from experts and other models by heavily penalizing shorter responses (one to two-sentence summaries) that omit key details (8%). In sum, different LLMs have different strengths and weaknesses when evaluating different aspects. Combining them together reduces judgment errors individual models make, especially for the unacceptable responses (Section 4, Appendix 5).

5 Ensemble Combination Analysis

Section 4 shows that ensemble of size three has the best alignment with developer judgments. We pick out the best model combination for sizes 1, 3, 5, and 7 for each aspect and show that for all aspects,

[4]We also analyzed when panel of LLM evaluation fails but crowdworkers succeed.These cases are rare: 3% coherence, 2% consistency, 0.8% fluency, 4% relevance. None of our results support the use of low quality crowdsourcing in evaluation.

Aspect	Size	Best F1	Median F1	Std Dev	Best Combination
Coherence	1	0.739	0.606	±0.059	Llama3-1-70b
Coherence	3	0.777	0.650	±0.032	Llama3-1-70b + gpt-3.5-turbo + Claude-sonnet
Coherence	5	0.771	0.674	±0.017	Llama3-1-70b + gpt-3.5-turbo + GPT-4o-mini + GPT-4 + Prometheus
Coherence	7	0.768	0.678	±0.008	Llama3-1-70b + gpt-3.5-turbo + GPT-4o-mini + GPT-4 + Claude-sonnet + Prometheus + Llama3-2-8b
Consistency	1	0.825	0.653	±0.152	Claude-sonnet
Consistency	3	0.847	0.715	±0.088	Claude-sonnet + GPT-4 + Llama3-1-70b
Consistency	5	0.839	0.747	±0.053	Claude-sonnet + GPT-4 + Llama3-1-70b + Llama3-2-8b + GPT-4o-mini
Consistency	7	0.823	0.763	±0.026	Claude-sonnet + GPT-4 + Llama3-1-70b + Llama3-2-8b + GPT-4o-mini + gpt-3.5-turbo + Prometheus
Fluency	1	0.719	0.579	±0.054	GPT-4
Fluency	3	0.732	0.648	±0.030	GPT-4 + GPT-4o-mini + Prometheus
Fluency	5	0.737	0.679	±0.018	GPT-4 + GPT-4o-mini + Prometheus + Llama3-2-8b + gpt-3.5-turbo
Fluency	7	0.720	0.695	±0.009	GPT-4 + GPT-4o-mini + Prometheus + Llama3-2-8b + gpt-3.5-turbo + Claude-sonnet + Llama3-1-70b
Relevance	1	0.715	0.598	±0.046	Llama3-1-70b
Relevance	3	0.728	0.631	±0.019	Llama3-1-70b + GPT-4 + Mistral-8x7b
Relevance	5	0.729	0.650	±0.008	Llama3-1-70b + GPT-4 + Mistral-8x7b + Claude-sonnet + Prometheus
Relevance	7	0.717	0.657	±0.004	Llama3-1-70b + GPT-4 + Mistral-8x7b + Claude-sonnet + Prometheus + GPT-4o-mini + gpt-3.5-turbo

Table 3: Best performing model ensembles for each aspect and ensemble size in SummEval. The best single models always appear to be in the best model combination no matter of ensemble size.

ASPECTS	Best Combination	Best F_1	Ensemble F1	Best F_1 Accept	Ensemble F_1 Accept	Best F_1 Unaccept	Ensemble F_1 Unaccept
Coherence	A	0.74	0.76↑	0.80	0.80	0.74	0.76 ↑
	B	0.70	0.63↓	0.72	0.66↓	0.69	0.66↓
	Crowd	0.51	0.50↓	0.58	0.60↑	0.43	0.40↓
Consistency	A	0.82	0.84 ↑	0.96	0.97 ↑	0.69	0.72↑
	B	0.51	0.49 ↓	0.73	0.70 ↓	0.29	0.29 ↓
	Crowd	0.46	0.46	0.76	0.77↑	0.16	0.15↓
Fluency	A	0.72	0.73 ↑	0.97	0.97	0.47	0.49 ↑
	B	0.57	0.54	0.86	0.82	0.27	0.27
	Crowd	0.47	0.48↑	0.75	0.77↑	0.19	0.19
Relevance	A	0.72	0.73 ↑	0.88	0.87↓	0.54	0.57 ↑
	B	0.59	0.61 ↑	0.83	0.77 ↓	0.44	0.45 ↑
	Crowd	0.52	0.52	0.73	0.75↑	0.30	0.29↓

Table 4: The combinations for A are three strong models– GPT-4, LLaMA-3-70B, and Cluade-Sonnet, and the combinations for B are three weaker models– GPT-3.5-turbo, Prometheus, and Mistral-8x7. Most of the aspects show an improvement above the Best F_1 score in the model combination list for strong model groups A, but a decrease below the best scores for the combination of weaker models. Crowd refers to crowdworker results. Since crowdworkers themselves are weak evaluators, combining them does not necessarily improve or even decrease the evaluation alignment with developers.

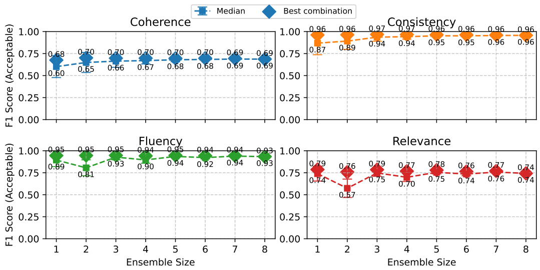
GPT-4, LLaMA3-70B, or Claude always appear in one of the best combinations, showing that a strong model in the mix is necessary to improve reliable evaluations (Table 3).

Strong Model Combination vs. Weak Model Combination We show that combining strong models increases evaluation reliability (Section 4). We pick out the three best models as a combination, and three worse models as a combination, as well as crowdworkers and show that combining strong models can lead to an increase of overall judgment alignment with experts while combing weak models or low-quality crowdsource ratings do not necessarily increase judgment quality, and may sometimes decrease it than their best constituents (Table 4).

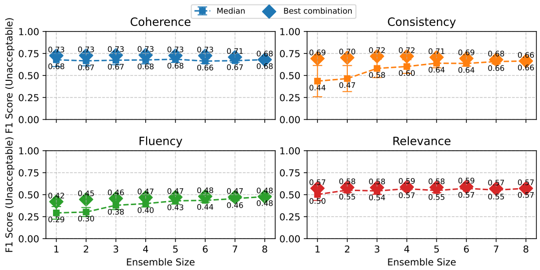
Ensemble Helps with the Rare Unacceptable Class Single models are already good at identifying the acceptable class, which is the more common class, thus combining more models does not increase the alignment with experts. However, combining models increases the median and maximum F_1 score alignment with experts for the rare unacceptable class (Figure 4).

We present the result on the precision and recall of acceptable classes compared with ensemble size in Figure 5. Results show that precision remains stable as ensemble size increases, while recall improves compared to single-model evaluation.

Identifying Problematic Cases Is Challenging We compute the precision and recall of LLMs agreements with gold ratings, which shows the precision and recall for the acceptable (upper panel) and unacceptable (lower panel) class from crowdsourced and GPT4-as-a-judge, with developer ratings as the ground truth. GPT-4 is better at the task than crowdsourcing, with reasonable precision and recall for the acceptable class and lackluster numbers for the unacceptable class (Figure 3). In addition, regardless of models or crowdworkers, the preci-

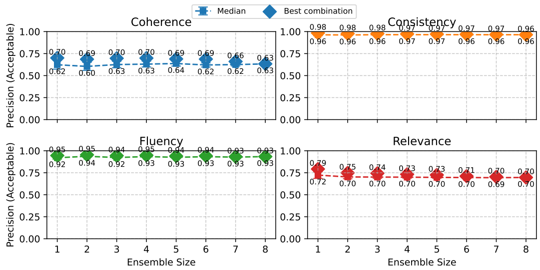


(a) F_1 score for acceptable class.

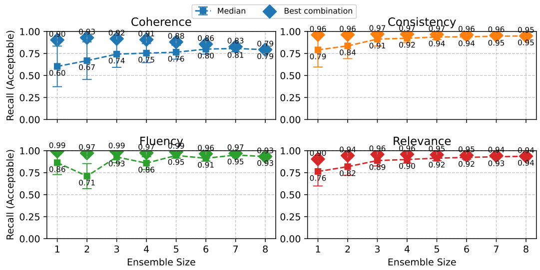


(b) F_1 score for unacceptable class.

Figure 4: Ensemble models and single models perform similarly for the majority acceptable class, but ensembles show notable improvement in identifying the rarer unacceptable responses.



(a) Precision for acceptable classes.



(b) Recall for acceptable classes.

Figure 5: The precision and recall of acceptable classes verses the ensemble size of models. As we can see, the precision for the acceptable class remains stable in regard with increasing ensemble sizes, but the recall increases compared to using one model for evaluation.

sion and recall for the unacceptable class are highly imbalanced.

6 Conclusion

Automating evaluation is easier for the common class than the rare class. However, identifying the rare class is crucial for detecting low quality responses. A failure to identify low quality responses can result in extra human annotations to accurately find the low-quality responses. While combining multiple strong models improves detection, it does not help weak models and crowdworkers. The percentage of examples crowdworkers identify correctly but LLMs cannot is rare, suggesting LLM evaluations are more reliable than low-quality crowdsourced evaluations. Even high-quality human annotations (developers, while original work refers to experts) can make mistakes. LLMs excel in aspects requiring detailed text analysis, such as consistency, but no evaluation is perfect. Future work can explore semi-automated evaluations that combine human expertise with multiple LLMs, leveraging each of their strengths while minimizing human effort and cost to make evaluations reliable.

7 Limitation

Recent years have seen a surge in LLMs with increasingly sophisticated capabilities across diverse tasks. This advancement has made reliable model evaluation more challenging during both the development and deployment phases, whether conducted by humans or other models. Our work conducts quantitative and qualitative analysis and shows that even high-quality human annotations have errors. LLMs and crowdworkers struggle to identify the problematic generations accurately, and combining multiple strong models increases the ability to identify those cases. However, most existing available datasets only contain crowdsource ratings, and it is challenging for us to find more suitable datasets that both include high-quality human ratings and crowdworker ratings. While we find that crowdsource ratings are less reliable than LLM evaluations, simply adding more datasets and comparing more crowdworker and LLM ratings will only reveal high disagreements without indicating which is correct. Resolving these disagreements requires high-quality human ratings, which are often unavailable. Additional available datasets with the available developer and crowdworker ratings in the future would strengthen our claim.

8 Ethics

All of our annotations are done by authors using aspect definitions in Table 1. The total number of annotations is 1,156 examples, where each example includes a source document, a reference summary, a model summary, three LLM justifications, and scores, as well as developer and crowdworker scores. The Institutional Review Board (IRB) has reviewed and exempted our annotation protocol. Our study does not present potential risks for the annotators.

References

Bhashithe Abeysinghe and Ruhan Circi. 2024. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches.

Anthropic. 2024. Claude: An ai assistant by anthropic. Accessed: 2024-12-23.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, Suchir Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study.

Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate.

Cheng-Han Chiang. 2023. A closer look into automatic evaluation using large language models.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations?

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Hoa Trang Dang. 2005. Overview of duc 2005.

DeepMind. 2024. Gemini pro. `https://deepmind.google/technologies/gemini/pro/`. Accessed: 2025-01-02.

Aaron Grattafiori et al. 2024. The llama 3 herd of models.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering.

Fan Gao, Hang Jiang, Rui Yang, Qingcheng Zeng, Jinghui Lu, Moritz Blum, Dairui Liu, Tianwei She, Yuang Jiang, and Irene Li. 2024. Large language models on wikipedia-style survey generation: an evaluation in nlp concepts.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

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.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024a. A survey on fairness in large language models.

Zongxia Li, Ishani Mondal, Yijun Liang, Huy Nghiem, and Jordan Lee Boyd-Graber. 2024b. Pedants: Cheap but effective and interpretable answer equivalence.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

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.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational autoraters: Taming large language models for better automatic evaluation.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. *ArXiv*, abs/2305.18201.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation.

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. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

Task Description: You will be given one summary written for a news article. Rate the summary on one metric. **Evaluation Criteria:** *ASPECTS (1-5)* - DESCRIPTION OF THE ASPECTS **Input Format:** *Source Text:* {Document} *Summary:* {Summary} **Output Format:** *ASPECTS:* Analysis: [analysis of ASPECTS] Rating: [1-5]

Table 5: Prompt template for SummEval rating. Analysis-before-rating approach yields better evaluation than chain-of-thought or direct rating.

A Evaluation Methodology

A.1 Prompt Templates for Evaluation

We prompt an LLM with the template provided in Table 5 to first let LLMs provide justifications for the rating and then output the rating. Following Chiang (2023), we find that prompting an LLM to first analyze the response based on the criteria then rate gives better evaluation than chain-of-thought or direct rating.

A.2 Rating Distribution Analysis

We show the distribution of ratings for developers, crowdworkers, and all selected LLMs. The unacceptable class is the rare class in most of the cases (Figure 6), which presents challenges for accurate evaluation.

B Error Analysis

B.1 Quantitative Error Analysis

We analyze 1,156 examples where a single model fails but the ensemble model corrects them, and where crowdworkers are correct but LLMs fail. Table 6 presents our systematic categorization of error patterns for each model and evaluation aspect.

B.2 Error Categories by Evaluation Aspect

The categorizations reveal that different LLMs have distinct weakness patterns across different aspects, but combining them in an ensemble corrects many of these errors. Key failure modes include:

- **Coherence**: Models struggle with direct source copying detection and penalize short responses excessively
- **Consistency**: Difficulty detecting non-factual information and handling coreference resolution
- **Fluency**: Confusion between grammatical errors and formatting issues

Error	Coherence	Consistency	Fluency	Relevance
GPT-4 Wrong	Direct source copying of the beginning of the document; short length answers (Overall average token length of the response is 200 but the average length of the mistaken short length answers is around 40 tokens)	—	Rating factual errors rather than fluency	Punishes for omitting details
LLaMA-3-70B	Direct source copying	Fails to detect non-factual information	Points out grammatical errors rather than formatting (ignoring the provided definition)	Errors on longer responses (The average length of LLaMA-3 makes errors on this aspects is 230 tokens, 30 tokens more than the overall average response length)
Claude-Sonnet	Punishes for repetition and factual errors	Punishes for redundancy and omitting details	Very lenient on grammar	Pays little attention to redundant information
Crowdworkers Right & LLM Wrong	Short length response with less details	Coreference; responses that include multiple subjects	Factual mistakes and grammar errors that are beyond fluency	Coreference: identifying the subject him/her is referring to

Table 6: Qualitative analysis of why individual LLMs might fail on specific quality aspects. The last row shows common errors that all three LLMs make but crowdworkers mark correctly.

- **Relevance**: Over-penalization for omitted details rather than assessing actual relevance

C Case Study Examples

We provide illustrative examples covering three categories of cases: (1) where a single model fails but an ensemble succeeds, (2) where ensemble evaluations conflict with human judgments, and (3) where models are correct but human experts are wrong. These examples address all four evaluation aspects and demonstrate both the strengths and limitations of ensemble approaches.

C.1 Single Model Failures Corrected by Ensemble

Figures 7–10 highlight cases where a single model fails while an ensemble succeeds. Single models tend to over-penalize minor errors—grammar, capitalization, punctuation, and omissions—without fully considering evaluation criteria. Ensembles mitigate this by leveraging diverse failure modes, allowing majority judgment to reduce blind spots.

C.2 Ensemble-Human Disagreements

Figures 11–14 present cases where ensemble evaluations conflict with human experts and crowdworkers. These examples highlight failures where all models struggle to detect direct copying or excessively penalize omitted details and grammar errors instead of adhering strictly to the evaluation criteria.

C.3 Cases Where Ensemble Outperforms Human Experts

Figure 15 presents a case where ensemble models successfully identify fluency errors in the model summary, while both human experts and crowdworkers overlook them. This example highlights the potential of ensemble models for large-scale evaluations and their ability to maintain consistent evaluation criteria.

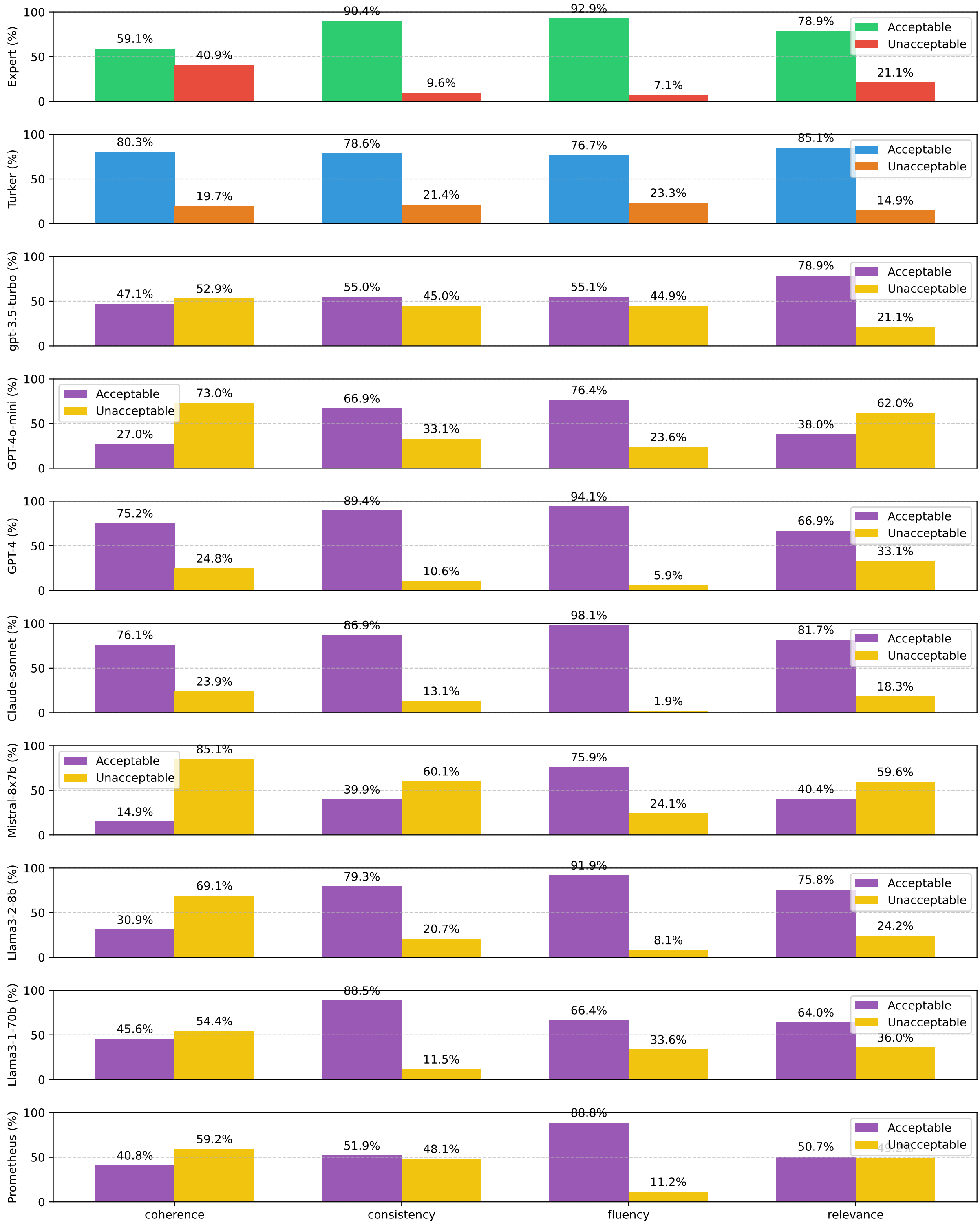


Figure 6: The rating distributions for all models are skewed towards one class. However, most of the good evaluators have more acceptable ratings than unacceptable ratings. The worst model Mistral-8x7b is highly skewed towards the unacceptable class.

Source: Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with Burnley on Sunday . 'Just been watching the game , did you miss the coach ? # RubberDub # 7minutes , ' Merson put on Twitter .

Reference: Andros Townsend an 83rd minute sub in Tottenham 's draw with Burnley . He was unable to find a winner as the game ended without a goal . Townsend had clashed with Paul Merson last week over England call-up . **Model**

Summary: Paul merson has restarted his row with andros townsend . The tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with burnley . Townsend was brought on in the 83rd minute for tottenham as they drew 0-0 with burnley .

Human Developer Rating: 2, 1, 2
GPT-4 (Rating: 4)
Analysis: The summary is coherent as it presents a clear sequence of events regarding the interaction between Paul Merson and Andros Townsend. It starts by stating that Merson restarted his row with Townsend, then describes Townsend's late entry into a game, and reiterates the timing of his entry, which supports the initial point about the row being related to his late game appearance. The summary maintains a logical flow and sticks to the topic without introducing unrelated information.

Comments on Model's Mistakes: The model fails to detect direct source copying without some context and transitions between sentences.

Reference: Naoki Ogane says that Chelsea have made an offer for Yoshinori Muto . The 22-year-old forward has one goal in 11 games for Japan . Muto admits that it is an 'honour ' to receive an offer from the Blues . Chelsea have signed a £200m sponsorship deal with Yokohama Rubber . Muto graduated from university with an economics degree two weeks ago . He would become the first Japanese player to sign for Chelsea . **Model Summary:**

Chelsea have made an offer for FC Tokyo 's 22-year-old forward Yoshinori Muto .

Human Developer Rating: 5, 5, 4
LLaMA-3-70B (Rating: 2)
Analysis: The summary is a single sentence that provides a clear and concise statement about Chelsea's offer for Yoshinori Muto. However, it lacks context and does not build upon any previous information, making it a standalone statement rather than a coherent body of information.

Comments on Model's Mistakes: The ensemble models overly punish for short but precise and coherent answers.

Figure 7: **Single Model Failure And Ensemble Success Case: Coherence.** We emphasize the key aspects of LLM failure modes, providing detailed commentary on these failures and supporting them with relevant references throughout the text.

Reference: Researchers claimed antibiotics could be contributing to 'obesity epidemic . ' Their large-scale study was published in the respected Pediatrics journal . It found one third of 10 to 11-year-olds in England are overweight or obese . Children who took antibiotics as babies were more likely to be overweight .

Model Summary: A third of 10-11 year olds and more than a fifth of 4-5 year olds in england are overweight or obese , leading to fears that today 's generation will be the first to die at an earlier age than their parents . Writing in the respected medical journal pediatrics , they said that the widely-prescribed drugs could be contributing to the obesity epidemic . The researchers said say that just as antibiotics are used to make farm animals put on weight , the may also be fattening our children .

Human Developer Rating: 5, 5, 5
LLaMA-3-70B (Rating: 3)
Analysis: The summary is partially consistent with the source text. It accurately reports the statistics on childhood obesity in England and the potential link between antibiotics and weight gain. However, it omits key details from the study, such as the specific age range of the children studied (0-6 months) and the findings on repeated prescriptions and gut bacteria. Additionally, the summary quotes the researchers out of context, making it seem like they are making a direct causal link between antibiotics and obesity, when in fact the study only suggests an association.

Comments on Model's Mistakes: The model overly punishes omitting details in the summary rather than key facts.

Reference: The Italian coast guard says 8,480 migrants were rescued from Friday to Monday .Save the Children said Tuesday 400 migrants could be missing from a boat . The Italian coast guard can not confirm that report .

Model Summary: Italian coast guard fears 400 migrants could be missing . A migrant boat carrying 550 people capsized in the mediterranean sea . At least 480 migrants have died in 2015 .

Human Developer Rating: 2, 2, 3
LLaMA-3-70B (Rating: 4)
Analysis: The summary is partially consistent with the source text. The first two sentences of the summary are factually aligned with the source, as they mention the fear of 400 migrants missing and the capsizing of a migrant boat carrying 550 people. However, the third sentence is inconsistent, as it states that at least 480 migrants have died in 2015, whereas the source text mentions that this number refers to the number of migrants who have died while crossing the Mediterranean since the beginning of the year, not just in 2015.

Comments on Model's Mistakes: The model fails to detect and properly lower the evaluation when detecting non-factual information and to set up the correspondence over multiple facts.

Figure 8: **Single Model Failure And Ensemble Success Case: Consistency.** We emphasize the key aspects of LLM failure modes, providing detailed commentary on these failures and supporting them with relevant references throughout the text.

Reference: Andros Townsend an 83rd minute sub in Tottenham 's draw with Burnley . He was unable to find a winner as the game ended without a goal . Townsend had clashed with Paul Merson last week over England call-up .

Model Summary: Paul merson has restarted his row with andros townsend . The tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with burnley . Andros townsend scores england 's equaliser in their 1-1 friendly draw with italy in turin .

Human Developer Rating: 5, 5, 5
LLaMA-3-70B (Rating: 1)
Analysis: The summary has several errors in grammar, spelling, and punctuation, such as missing articles, incorrect capitalization, and lack of proper sentence structure. The text is not smooth to read and has an unnatural tone.

Comments on Model's Mistakes: The model points out grammatical errors rather than formatting (ignoring the provided definition).

Reference: The A & E networks are remaking the blockbuster " Roots " miniseries , to air in 2016 . The epic 1977 miniseries about an African-American slave had 100 million viewers .

Model Summary: A & e , lifetime and history announced thursday that the three networks would simulcast a remake of the saga of kunta kinte , an african who was captured , shipped to america and sold into slavery . Producers will consult scholars in african and african-american history for added authenticity ..

Human Developer Rating: 5, 5, 5
LLaMA-3-70B (Rating: 1)
Analysis: The summary has several errors that affect its clarity and smoothness. The text lacks proper capitalization, punctuation, and sentence structure. The sentences are also quite short and lack variation.

Comments on Model's Mistakes: The model is overly strict on minimal errors, like capitalization, punctuation, and sentence structure, ignoring the provided definition. Also, the model is limited in understanding non-English context in the text when evaluating fluency.

Figure 9: **Single Model Failure And Ensemble Success Case: Fluency.** We emphasize the key aspects of LLM failure modes, providing detailed commentary on these failures and supporting them with relevant references throughout the text.

Reference: Researchers claimed antibiotics could be contributing to 'obesity epidemic . ' Their large-scale study was published in the respected Pediatrics journal . It found one third of 10 to 11-year-olds in England are overweight or obese . Children who took antibiotics as babies were more likely to be overweight .

Model Summary: Finnish researchers compared the weight and height of more than 12,000 healthy two year old with records on antibiotic prescription . By two years-old , one in five boys and one in ten girls was overweight or obese .

Human Developer Rating: 3, 3, 3
Claude-sonnet (Rating: 5)
Analysis: The summary captures the key finding from the source text that a significant portion of 2-year-old children were overweight or obese, and that there was a correlation between antibiotic use in infancy and being overweight at age 2. It focuses on the important details from the study without including extraneous information.

Comments on Model's Mistakes: The model fails to pay much attention to redundant information and does not punish for omitting key points

---

Reference: Jesse and Melissa Meek revealed they 're expecting a child in a rap video they made set to theme from 'The Fresh Prince of Bel-Air . ' The clip , which features Jesse and Melissa Meek rapping as they drive in a car , has been viewed over 1.7 million times on YouTube . It took five takes to film the Happy Valley , Oregon , couple 's video .

Model Summary: The clip , which features jesse and melissa meek rapping as they drive in a car , has been viewed over 1.7 million times on youtube . ' in happy valley , oregon , livin ' it wed , bought our first house and started to build our homestead .

Human Developer Rating: 4, 2, 2
Claude-sonnet (Rating: 5)
Analysis: The summary captures the key details about the couple announcing their pregnancy through a rap video set to the Fresh Prince of Bel-Air theme song, including their names, location, and that they bought their first house. It focuses on the relevant information from the source text without including extraneous details.

Comments on Model's Mistakes: The model is does not punish for omitting key points and fails to detect direct copy from the source text.

Figure 10: **Single Model Failure And Ensemble Success Case: Relevance.** We emphasize the key aspects of LLM failure modes, providing detailed commentary on these failures and supporting them with relevant references throughout the text.

Source: Barcelona club president Josep Maria Bartomeu has insisted that the La Liga leaders have no plans to replace Luis Enrique and they 're 'very happy ' with him . The 44-year-old took only took charge of the club last summer , signing a two-year-deal , and is only six games away from winning the title . Despite speculation this season that Enrique will be replaced in the summer , Bartomeu refuted these claims and says he 's impressed with how the manager has performed . Luis Enrique only took charge at the club last summer and has impressed during his tenure Barcelona president Josep Maria Bartemou says the club are 'very happy ' with Enrique 's performance 'There 's a lot of time to talk about his contract in the future , ' the Barca president told bySPORT . 'Luis Enrique is concentrating on La Liga and on everything he needs to do in relation to the team . '...

Reference: Barcelona president Josep Bartomeu says the club are happy with Enrique . Barca are currently top of La Liga and closing in on the league title . Enrique 's future at the club has been speculated over the season .

Model Summary: Barcelona club president josep maria bartomeu has insisted that the la liga leaders have no plans to replace luis enrique and they 're ' very happy ' with him . The 44-year-old took charge of the club last summer , signing a two-year-deal , and is only six games away from winning the title . Barcelona president josep maria bartemou says the club are ' very happy ' with enrique 's performance .

Human Developer Rating: 4, 2, 2
Crowdworker Rating: 3, 3, 3, 3

GPT-4 (Rating: 5)
Analysis: The summary is well-structured and maintains a logical flow of information. It starts by stating the club president's stance on not replacing the manager, Luis Enrique, and their satisfaction with his performance. It then provides context about Enrique's tenure at the club, mentioning his contract and proximity to winning the title. The repetition of the club's happiness with Enrique in the last sentence reinforces the main point, maintaining coherence throughout the summary.

LLaMA-3-70B (Rating: 5)
Analysis: The summary is well-structured and coherent, with each sentence building on the previous one to provide a clear picture of Barcelona's stance on Luis Enrique's position. The sentences are logically connected and flow smoothly, providing a cohesive narrative.

Claude-Sonnet (Rating: 4)
Analysis: The summary maintains a good level of coherence by presenting the key points in a logical flow. It starts by stating the main idea that Barcelona has no plans to replace Luis Enrique and is happy with him. It then provides relevant details about Enrique's tenure, contract duration, and the club's potential to win multiple trophies this season. The sentences are well-connected and build upon each other to form a coherent body of information.

Comments on Models' Mistakes: The models fail to detect direct copy from the source text.

Figure 11: **Conflict with Crowdworker: Coherence.** We emphasize the key aspects of LLM failure modes, providing detailed commentary on these failures and supporting them with relevant references throughout the text.

Reference: Sir Bradley Wiggins will bid to break cycling 's hour record this year . He will race at London 's Olympic Velodrome in front of 6,000 fans in June . Wiggins will ride in next month 's Tour de Yorkshire . He is also targeting his eighth Olympic medal at the Rio 2016 games .

Model Summary: Sir bradley wiggins will bid for cycling 's hour record at london 's olympic velodrome . The 2012 tour de france winner is 35 on april 28 . He is 35 , will attempt to add to his achievements by riding the furthest distance in 60 minutes at the lee valley velopark . He has four gold , one silver and two bronze - at the 2016 rio olympics . The four-time olympic champion and luke rowe are expected to set a mark which will be broadcast live on sky sports .

Human Developer Rating: 5, 5, 5
Crowdworker Rating: 5, 5, 5, 5

GPT-4 (Rating: 3)
Analysis: The summary is mostly consistent with the source text, but there are a few inaccuracies: (1) The summary states that Wiggins and Luke Rowe are expected to set a mark, but the source text does not mention Luke Rowe in this context. It only mentions Wiggins' attempt to break the hour record. (2) The summary mentions Wiggins' medals at the 2016 Rio Olympics, but the source text discusses his target for a British record eighth Olympic medal at the 2016 Rio Olympics, not that he already has these medals from that event. (3) The summary implies that the event will be broadcast live on Sky Sports due to Wiggins and Rowe, but the source text only mentions the broadcast in relation to Wiggins' attempt.

LLaMA-3-70B (Rating: 2)
Analysis: The summary contains some factual inconsistencies and hallucinations. For example, the summary states that Sir Bradley Wiggins is 35, which is correct, but it also mentions that he has four gold, one silver, and two bronze medals at the 2016 Rio Olympics, which is incorrect as the article only mentions his current medal count and his target for the 2016 Olympics. Additionally, the summary mentions that Luke Rowe is expected to set a mark, which is not mentioned in the original article.

Claude-Sonnet (Rating: 2)
Analysis: The summary contains several factual inconsistencies with the source text. It incorrectly states that Wiggins and Luke Rowe are expected to set a record together, when the source only mentions Rowe as Wiggins' teammate. It also incorrectly claims that Wiggins has won medals at the 2016 Rio Olympics, which had not happened at the time of the article.

Comments on Models' Mistakes: The models fail to emphasize essential facts and overly punish redundancy and omitted details.

Figure 12: **Conflict with Crowdworker: Consistency.** We emphasize the key aspects of LLM failure modes, providing detailed commentary on these failures and supporting them with relevant references throughout the text.

Reference: Neymar helped Brazil beat Chile 1-0 at the Emirates stadium last weekend . Barcelona won the El Clasico to go four points clear at the top of La Liga . Luis Enrique 's side take on Celta Vigo in La Liga on Sunday .

Model Summary: Real madrid beat chile 1-0 in el clasico on thursday . Dani alves and dani alves scored the goals in the euro league contest . Neymar and alves joined the club 's sister rafaella on sunday . Real madrid remain top of their euro league rivals at the emirates .

Human Developer Rating: 5, 5, 5
Crowdworker Rating: 4, 5, 4, 4

GPT-4 (Rating: 1)
Analysis: The summary contains multiple errors and inconsistencies that affect its fluency. The use of lowercase for proper nouns like "Real Madrid," "Chile," "El Clasico," "Dani Alves," and "Euro League" is incorrect. The summary also incorrectly states that "Real Madrid beat Chile 1-0 in El Clasico," which is a mix-up of different events and teams. Additionally, the phrase "Dani Alves and Dani Alves scored the goals" is repetitive and incorrect. The overall sentence structure and coherence are poor, making the summary difficult to understand and follow.

LLaMA-3-70B (Rating: 1)
Analysis: The summary has several errors that affect its clarity and accuracy. The first sentence is incorrect, as it was Barcelona that beat Real Madrid 85-80, not the other way around. Additionally, the sentence "Dani alves and dani alves scored the goals in the euro league contest" is nonsensical, as Dani Alves is a single person and cannot score goals for both teams. The sentence "Neymar and alves joined the clubś sister rafaella on sunday" is also unclear, as Rafaella is Neymarś sister, not the clubś. Finally, the last sentence is incorrect, as Real Madrid remain top of their Euro League division due to points difference, not because they beat Chile at the Emirates.

Claude-Sonnet (Rating: 1)
Analysis: The summary contains multiple factual errors, poor grammar, and awkward phrasing that make it difficult to understand the main points. It incorrectly states that Real Madrid beat Chile, confuses the basketball match with a soccer game, and has issues with subject-verb agreement and word choice.

Comments on Models' Mistakes: The models insist on pointing out factual, grammatical, and pronunciation errors rather than formatting (ignoring the provided definition).

Figure 13: **Conflict with Turker: Fluency.** We emphasize the key aspects of LLM failure modes, providing detailed commentary on these failures and supporting them with relevant references throughout the text.

Reference: Ben Stokes returns to the ground where he broke his wrist hitting a locker . He clashed with West Indies batsman Marlon Samuel in the second Test . But his aggression , when controlled , is important to England .

Model Summary: Nothing is sacred in the team . When ben stokes first turned up at the kensington oval this week someone would have said : ‘ mind that locker , stokesey . It was here , of course , that stokes broke his wrist punching a locker after being dismissed last year and that will be a reminder to him of the need to channel his aggression in the right way now he is back in the england team . Stokes (left) clashed with marlon samuels during the second test between west indies and england

Human Developer Rating: 4, 5, 3
Crowdworker Rating: 4, 4, 4, 4

GPT-4 (Rating: 3)
Analysis: The summary captures the essence of Ben Stokes' aggressive nature and the incident of him breaking his wrist, which is relevant to the source text. It also mentions his clash with Marlon Samuels, which is an important event discussed in the source. However, the summary omits other significant details such as the advice from coach Peter Moores, Stokes' need for careful management, and his potential contribution to the team in terms of runs and wickets. These omissions result in a summary that, while focused, misses out on some key aspects of the source text that are important for a complete understanding of Stokes' situation and impact on the team.

LLaMA-3-70B (Rating: 2)
Analysis: The summary includes some important details about Ben Stokes, such as his reputation as a fiery character, his past incident of breaking his wrist punching a locker, and his clash with Marlon Samuels. However, the summary lacks context and does not provide any information about the main point of the article, which is the need for Stokes to channel his aggression in the right way and the importance of careful man-managing by the England team.

Claude-Sonnet (Rating: 3)
Analysis: The summary captures some relevant details about Ben Stokes' aggressive behavior and the incident where he punched a locker and broke his wrist. However, it omits important context about Stokes' role in the England cricket team, his value as an exciting player, and the need to manage his aggression constructively. The summary focuses narrowly on the negative aspects of his behavior without providing a balanced perspective.

Comments on Models' Mistakes: The models overly punish for omitting details.

Figure 14: **Conflict with Crowdworker: Relevance.** We emphasize the key aspects of LLM failure modes, providing detailed commentary on these failures and supporting them with relevant references throughout the text.

Reference: Bayern Munich boss Pep Guardiola ripped a hole in his trouser leg . His underwear were on show during the European match in Germany . The German giants booked place in the semi-finals thanks to 6-1 rout .

Model Summary: Bayern munich manager pep guardiola 's underwear was on show . Porto beat 5-0 at the allianz arena on tuesday night . The spanish tactician 's underwear is on show after the incident . The bayern beat porto in the champions league on tuesday .

Human Developer Rating: 4, 4, 5
Crowdworker Rating: 3, 4, 4, 5

GPT-4 (Rating: 1)
Analysis: The summary has several grammatical errors and awkward phrasing that affect its clarity. The repetition of "underwear was on show" is redundant, and the sentence "Porto beat 5-0 at the allianz arena on tuesday night" is grammatically incorrect and misleading as it suggests Porto won, which is incorrect based on the source text. The summary also lacks smooth transitions between sentences, making it choppy and hard to follow.

LLaMA-3-70B (Rating: 1)
Analysis: The summary has many errors that make it hard to understand or sound unnatural. The sentences are short and lack proper grammar, punctuation, and capitalization. The text also lacks coherence and clarity, making it difficult to follow.

Claude-Sonnet (Rating: 1)
Analysis: The summary contains several grammatical errors, awkward phrasing, and lacks coherence. There are issues with subject-verb agreement ("Porto beat 5-0"), incorrect verb tenses ("The bayern beat porto"), and repetitive statements about Guardiolaś underwear being on show. The sentences do not flow smoothly, making the summary difficult to follow.

Comments on Models' Analysis: The models point out many capitalization, grammatical, and pronunciation errors, along with some confusing sentences, but the Human Developers fail to detect these mistakes.

Figure 15: **When Models are Right but Developer are Wrong: Fluency.** We emphasize the key aspects of LLM failure modes, providing detailed commentary on these failures and supporting them with relevant references throughout the text.