Rate, Explain and Cite (REC): Enhanced Explanation and Attribution in Automatic Evaluation by Large Language Models

Aliyah R. Hsu [1 2] **James Zhu** [2] **Zhichao Wang** [2] **Bin Bi** [2] **Shubham Mehrotra** [2] **Shiva K. Pentyala** [2] **Katherine Tan** [2] **Xiang-Bo Mao** [2] **Roshanak Omrani** [2] **Sougata Chaudhuri** [2] **Regunathan Radhakrishnan** [2] **Sitaram Asur** [2] **Claire Na Cheng** [2] **Bin Yu** [1]

Abstract

LLMs have demonstrated impressive proficiency in generating coherent and high-quality text, making them valuable across a range of text-generation tasks. However, rigorous evaluation of this generated content is crucial, as ensuring its quality remains a significant challenge due to persistent issues such as factual inaccuracies and hallucination. This paper introduces three fine-tuned general-purpose LLM auto-evaluators, REC-8B, REC-12B and REC-70B, specifically designed to evaluate generated text across several dimensions: faithfulness, instruction following, coherence, and completeness. These models not only provide ratings for these metrics but also offer detailed explanation and verifiable citation, thereby enhancing trust in the content. Moreover, the models support various citation modes, accommodating different requirements for latency and granularity. Extensive evaluations on diverse benchmarks demonstrate that our general-purpose LLM auto-evaluator, REC-70B, outperforms state-of-the-art LLMs, excelling in content evaluation by delivering better quality explanation and citation with minimal bias. Our REC dataset and models are available at https://github.com/adelaidehsu/REC.

1. Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities in generating high quality coherent text and are deployed in applications for various text-generation tasks (Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2023; Touvron et al., 2023). In order for LLMs to provide up-to-date information or to perform knowledge-intensive tasks (Lewis et al., 2020), Retrieval Augmented Generation (RAG) system (Chen et al., 2017; Borgeaud et al., 2021; Izacard et al., 2022; Guu et al., 2020) has been widely used. RAG involves first retrieving documents that are relevant to the generation task from an external knowledge source, and performing the generation using a knowledge-augmented prompt. However, it is of paramount importance to ensure that the generated text is reliable and can be trusted by a human, as LLMs often suffer from factual inaccuracies and hallucination of knowledge (Ji et al., 2022; Zhang et al., 2023; Shuster et al., 2021). Towards this goal, we propose that the generated text needs to be evaluated along various dimensions shown below:

- **Faithfulness:**Is the LLM generated response factually correct given the context?
- **Instruction Following:** Does the LLM generated response follow the instructions provided in the prompt?
- **Coherence:** Is the LLM generated response coherent?
- **Completeness:** Does the LLM generated response include all the necessary details?
- **Citation:** If a LLM generated response is factually correct, can we provide evidence for where that response came from?

In this paper, we introduce fine-tuned models for the evaluation tasks listed above that can form the basis for all trust-related evaluations for generative AI applications. The models were fine-tuned to not only provide ratings for the metrics but also explanations + citation for why it rated the generation so. An example of how our models operate for content quality evaluation can be found in Figure 1.

Our models are the *first* to enable citation in both automatic evaluation (i.e., content quality citation) and general task output (i.e., general RAG citation) with scalability and efficiency, while prior work often focus on providing solely rating and explanation in automatic evaluation (Zhu et al.,

Work done during an internship at Salesforce. [1]UC Berkeley [2]Salesforce AI Platform. Correspondence to: Aliyah R. Hsu <aliyahhsu@berkeley.edu>, James Zhu <james.zhu@salesforce.com>.

under review.

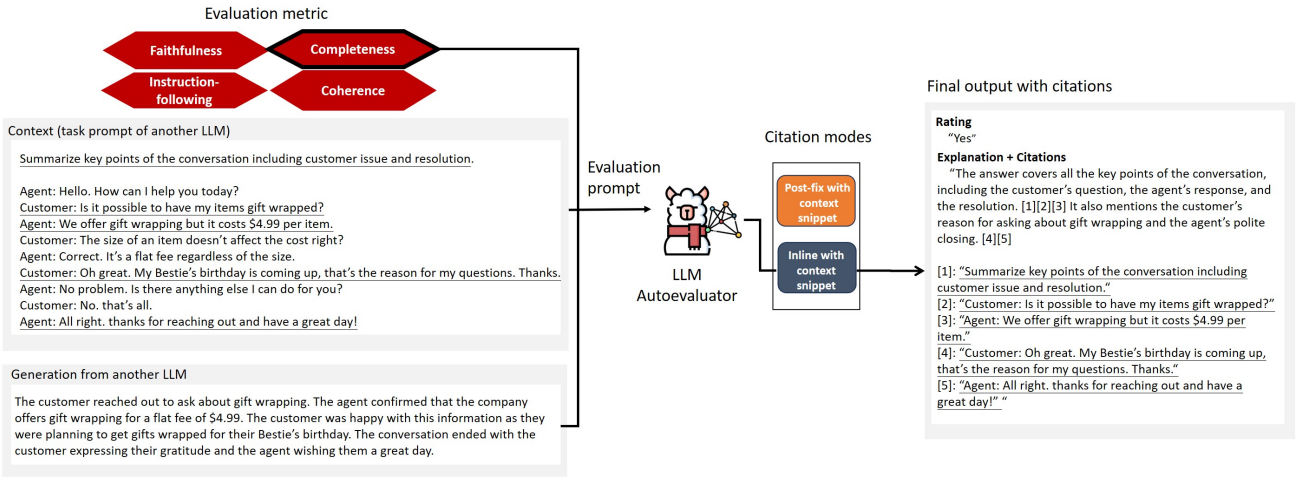

Figure 1. Illustration of the REC framework for content quality evaluation. The auto-evaluator takes in context (a task prompt from another LLM), generation from another LLM, and a user-specified evaluation metric (i.e., completeness). The auto-evaluator outputs rating, explanation with citation according to a user-specified citation mode (i.e., inline with context snippet). Citation are extracted verbatim from context as underlined. For details of the alternative citation mode and the evaluation prompt, see Appendix A.

2023; Li et al., 2023a; Wang et al., 2023b; Vu et al., 2024a), or require iterative prompting (Sun et al., 2023; Cao & Wang, 2024; Ye et al., 2023) or human annotation (Malaviya et al., 2024) for general citation generation. Our contributions include the following:

- A novel general-purpose LLM auto-evaluator that comes in three sizes: 8B, 12B and 70B. Our LLM auto-evaluators can generate improved quality **R**ating, **E**xplanation and **C**itation (**REC**), with little to no trade-off in general instruction task performance. [1]
- A curated dataset for citation and explanation fine-tuning to facilitate future research, which is the *first* public dataset containing both *content quality citation* [2] and *RAG citation*.
- A single model that can perform different modes of citation to cover the trade-off between latency and granularity of citation.

We provide extensive evaluations of our REC models on diverse benchmarks evaluating LLMs' (1) *RAG citation* capability: ALCE (Gao et al., 2023), and ExpertQA (Malaviya et al., 2024); (2) *content quality citation* capability: human evaluation on ABCD summarization (Chen et al., 2021); (3) general capabilities: RewardBench (Lambert et al., 2024), and LLM-AggreFact (Tang et al., 2024a); and (4) cognitive bias: CoBBLEr (Koo et al., 2024), against 7 state-of-the-art (SOTA) LLMs, such as GPTs and Claude. We show that our REC models outperform state-of-the-art baselines on these benchmarks, and that the baselines often struggle to provide accurate citation, undermining their reliability.

[1] All evaluations are limited to English datasets.

[2] Citation to support LLM generated explanation for content quality evaluation. See Section 2 for more details.

2. REC Model Citation Capabilities

In this section, we discuss the two main citation tasks, content quality citation and RAG citation, that the REC models support, with the understanding that the models can also be used for general purposes, such as question answering (QA) and summarization.

We illustrate the content quality citation task in Figure 1, where the auto-evaluator performs content quality evaluation on another LLM's generation according to a specified metric, and provides rating and explanation with citation for verifiability as an output. When not in an evaluation setting, our models can perform general RAG citation, where we cite from retrieved articles and tag the citation with an answer generated by another LLM, as illustrated in Figure 2.

Our models support various citation modes dependent on the citation tasks: post-fix citation (RAG only), inline citation (RAG only), post-fix citation with context snippet (for both content quality evaluation and RAG), and inline citation with context snippet (for both content quality evaluation and RAG). Output examples of different citation modes can be found in Appendices A and B.

These different modes were designed to accommodate different trade-offs between latency and granularity of citation.

For instance, the post-fix citation mode would be the fastest as there is no need for our models to generate claims or snippets. It simply has to generate the reference to the cited context. On the other hand, the inline citation with context snippet mode has to place the citation inline with the generated response and also point to a snippet within the corresponding cited context. This mode is the most granular, but can increase latency as it needs to generate more output tokens.

3. Related Work

Verifiable Text Generation Providing a reference to the attributable source of information (Liu et al., 2023a; Gao et al., 2023) has been proposed as one of the approaches to mitigate hallucination and improve the factuality of LLM generation. Such work fall into the general *RAG citation* setting defined in our work, where the aim is to generate verifiable content through citation of provided articles. Initial efforts fine-tune LLMs using human-written examples (Nakano et al., 2021) or machine-generated examples verified by humans (Menick et al., 2022), but their privately maintained training data limits further research. With the advent of more capable LLMs (OpenAI, 2023; Jiang et al., 2023a), most existing work rely on zero-shot prompting or few-shot prompting to cite articles during generation (Kamalloo et al., 2023; Gao et al., 2023; Liu et al., 2023a), although the quality of their generated citation leaves much room for improvement (Malaviya et al., 2024). Other work utilize an additional natural language inference (NLI) model to add citation (Gao et al., 2022; Chen et al., 2023). To improve upon prior work, we propose a fine-tuned LLM that can generate better-grounded responses supported with citation of various modes, with the fine-tuning dataset released to the public to facilitate future research and no additional NLI models needed.

General-purpose auto-evaluators Collecting human annotations to evaluate LLM is not only costly and time consuming, but hard to replicate (Ouyang et al., 2022; Zheng et al., 2023; Chiang & Lee, 2023), and as a result, LLMs become a natural automated proxy to evaluate LLM capabilities on various benchmarks (Bai et al., 2024; Bubeck et al., 2023; Chiang et al., 2023; Fu et al., 2023; Liu et al., 2023b; Wang et al., 2023a). Existing LLM auto-evaluators often judge LLM outputs by expressing "preference" over outputs from a reference model (Dubois et al., 2024; Li et al., 2023b; Yuchen Lin et al., 2024), or by providing rating and explanation according to a user-defined metric as a direct assessment (Li et al., 2023a; Wang et al., 2023b; 2024b). Closer to our work, Jiang et al. (2024) and Xu et al. (2023) fine-tune LLMs to generate rating, explanation, and a detailed analysis to pinpoint errors in a response evaluated. Unlike prior work, our fine-tuned model is a more generalizable auto-evaluator since it provides *content quality citation* for both good and bad evaluated responses, in addition to rating and explanation. This not only allows users to be able to diagnose where errors are from but provides supporting evidence when a generation is good, which are both important to establish trust in a model.

4. Data Collection

To fine-tune our general-purpose LLM auto-evaluator, we meticulously collect a mixture of data encompassing a broad spectrum of LLM capabilities (Section 4.1), and we denote our collected dataset as REC-Data. Due to a lack of publicly available content quality citation datasets, we leverage synthetic data generation using an LLM to curate such data (Section 4.2) to facilitate the research community. We perform rule-based automatic quality check followed by a unified task formatting to post-process our data for instruction fine-tuning (Section 4.3).

4.1. Task Types

To enhance the explainability in the automatic evaluation of our LLM auto-evaluator, while maintaining its general instruction-following capability, we gather datasets from a diverse range of task types (See detailed REC-Data distribution in Appendix C). These tasks represent essential capabilities that an advanced auto-evaluator LLM should possess:

- Pairwise Evaluation: Compare two responses at the same time and express a preference according to evaluation criteria.
- Pointwise Evaluation: Evaluate specific aspects of a response independently according to evaluation criteria and provide a rating.
- Open-ended Evaluation: Evaluate a response independently and provide a free-form explanation, usually to support either pairwise or pointwise evaluation.
- Citation: Evaluate a response alongside the context, and provide citation for verifiable answer attribution.
- General Instruction: Generate a response as instructed (no evaluation tasks in this type), such as summarization and QA.

4.2. Synthetic Data Generation

We leverage synthetic data generation for some of the tasks, including pointwise evaluation and citation. To ensure the quality of synthetic data, we leverage a powerful instruction-tuned model based on Llama-3.1-70B to generate both pointwise evaluation data and citation data for RAG and content

quality. [3] For pointwise evaluation, the model rates every response on each of the 4 metrics listed in Section 1. Example prompts used to generate the synthetic data are provided in Appendix D.

4.3. Post-processing and Unified Task Format

Knowing that LLM-generated outputs are susceptible to hallucination, we further post-process the synthetic data to ensure the quality. Specifically, we filter synthetic LLM outputs that are either in wrong JSON output formats, or with invalid citation by checking whether the citation is extracted verbatim from the context. Our final REC-Data has in total around 140k datapoints, each containing a task prompt and a completion.

5. Model Training

Our general-purpose LLM auto-evaluator is available in three sizes, REC-8B, REC-12B, and REC-70B. REC-12B is instruction fine-tuned from Mistral-Nemo[4], and REC-8B and REC-70B are instruction fine-tuned from Llama-3.1-8B and Llama-3.1-70B (Team, 2024), respectively. We trained REC-8B and REC-12B to assess the generalizability of the REC approach to different model architectures in similar model sizes. We adopted supervised fine-tuning (SFT) to optimize the models. To accommodate GPU memory constraints: we filtered examples where the combined length of a prompt and a response exceeded 6,144 tokens. We used Low-Rank Adaptation (LoRA) during fine-tuning (Hu et al., 2021), with a rank of $r = 256$ for REC-8B and REC-12B, and $r = 64$ for REC-70B. We applied 4-bit quantization for REC-70B during initialization. All models were trained using a learning rate of 1×10^{-4}, with data shuffled and for a single epoch. REC-8B and REC-12B were trained with a batch size of 2 and a gradient accumulation factor of 8, whereas for REC-70B we used a batch size of 1 with the same gradient accumulation factor. All models were trained on eight NVIDIA H100 GPUs, each with 80GB of memory. The total training time was approximately 4.5 hours for REC-8B and REC-12B, and 5.5 hours for REC-70B.

6. Evaluation

In this section, we evaluate our REC models on diverse benchmarks assessing LLMs' (1) *RAG citation* capability: ALCE (Gao et al., 2023), and ExpertQA (Malaviya et al., 2024); (2) *content quality citation* capability: human evaluation on ABCD summarization (Chen et al., 2021); (3) general capabilities: RewardBench (Lambert et al., 2024), and LLM-AggreFact (Tang et al., 2024a); and (4) cognitive bias: CoBBLEr (Koo et al., 2024), each of which is an important measure to a general-purpose LLM auto-evaluator. We compare our results against 7 SOTA LLMs, including Misral-7B (`Mistral-7B-Instruct-v0.2`) (Jiang et al., 2023b), Mistral-Nemo (`Mistral-Nemo-Instruct-2407`), Llama-3.1-70B (Grattafiori et al., 2024), Claude-3-Opus (`claude-3-opus-20240229`), GPT-3.5 (`gpt-3.5-turbo`), GPT-4 (`gpt-4-turbo`) (OpenAI et al., 2024), and GPT-4o (`gpt-4o`).

	AutoAIS	FActscore
Mistral-7B	26.39	84.03
Mistral-Nemo	41.39	85.78
Llama-3.1-70B	52.11	86.42
Claude-3-Opus	59.84	87.16
GPT-3.5	57.65	86.27
GPT-4	60.18	87.31
GPT-4o	60.86	87.03
REC-8B	51.36	85.81
REC-12B	<u>62.90</u>	<u>88.22</u>
REC-70B	**66.52**	**89.05**

Table 1. General RAG citation quality and answer correctness results on ExpertQA. The best and the second best performance for each metric are in bold and underlined separately.

	Fluency	Correct	Citation F1
Mistral-7B	64.49	15.12	42.94
Mistral-Nemo	61.35	18.03	41.32
Llama-3.1-70B	<u>72.01</u>	19.09	40.18
Claude-3-Opus	67.53	12.11	35.30
GPT-3.5	51.20	19.02	**50.87**
GPT-4	57.98	20.08	<u>47.09</u>
GPT-4o	56.29	**22.47**	40.50
REC-8B	68.20	18.01	38.77
REC-12B	64.94	19.60	43.63
REC-70B	**73.08**	<u>20.52</u>	41.73

Table 2. Average fluency, answer correctness and citation quality across datasets on ALCE. The best and the second best performance for each metric are in bold and underlined separately.

6.1. RAG citation & Correctness

The two benchmarks discussed in this section evaluate LLMs' ability to generate RAG citation in the *inline* mode.

ExpertQA ExpertQA (Malaviya et al., 2024) is designed to evaluate LLMs' ability to generate accurate and well-attributed responses in technical and high-stakes fields, such as medicine and law. It was developed by involving experts from 32 different fields, who contributed 2,177 domain-specific questions based on their knowledge. In total, 484 participants helped curate these questions. LLMs were then used to generate responses to the questions, followed by human experts evaluating the quality of these responses on sev-

[3] Agreement between this model's judgment and human labelers' judgment is 86%, on par with the agreement between human labelers.

[4] https://mistral.ai/news/mistral-nemo

eral criteria, including factual correctness, completeness of attribution, source reliability, and informativeness. We adopt the metrics used in its paper: AutoAIS (Gao et al., 2022), which is similar to citation recall, and FActscore (Min et al., 2023) which measures the percentage of generated claims that are factual.

As shown by the zero-shot results in Table 1, our REC-12B and REC-70B achieve the highest AutoAIS and FActscore scores, and our REC-8B is also able to outperform other similar-sized models (Mistral-7B and Mistral-Nemo), which not only indicates their better citation quality but also the generalizability of their attribution capability across domains.

ALCE The ALCE (Automatic LLMs' Citation Evaluation) benchmark (Gao et al., 2023) is designed to assess the ability of LLMs to generate text with accurate and relevant citation, focusing on three key dimensions: fluency, answer correctness, and citation quality. The ALCE benchmark is built on three datasets: ASQA (a short-answer question dataset), QAMPARI (which provides lists of correct answers), and ELI5 (a long-form question-answer dataset). Note that fluency is not measured in QAMPARI because of the nature of its answers being in lists. For each dataset, ALCE evaluates how well a model generates text that is not only fluent and accurate but also properly supported by citation. The citation quality evaluation includes citation precision (ensuring all cited sources are relevant) and citation recall (ensuring all necessary sources are cited). [5]

From Table 2, we see that our REC-70B model ranks highly in fluency and answer correctness. Although the GPT models excel in citation quality in these datasets, the much smaller REC-12B and REC-70B models outperform most of the competitors. When taking the three metrics together and considering the overall average performance, REC-70B achieves an average score of 45.11, the highest among the compared models. Since all criteria are important in generating a good response with accurate citations [6], this highlights the robustness and balance of REC-70B across fluency, correctness, and citation quality. REC offers a well-rounded solution for different types of questions, making it highly reliable for real-world applications where both citation quality and answer correctness are essential.

6.2. Rating & Explanation & Citation

The benchmark discussed in this section evaluates LLMs' ability to generate content quality citation in the *inline with context snippet* mode.

[5]For a detailed definition of the fluency and answer correctness metrics, please refer to Gao et al. (2023).

[6]For why the three metrics are all important to a verifiable LLM generation, see Appendix E.

ABCD To evaluate our LLM auto-evaluator's citation capability more comprehensively, besides examining its general *RAG citation* capability on QA tasks as reported in Section 6.1, we evaluate the *content quality citation* capability on a summarization task, the ABCD dataset (Chen et al., 2021), in this section.

The ABCD dataset (Chen et al., 2021) contains customer support conversation transcripts. We generate summaries by prompting an open-source LLM, Mistral-7B (Jiang et al., 2023a), with: "Summarize knowledge from transcripts after they've ended, including the customer issue and resolution." to form (dialogue, summary) pairs for evaluation. LLM auto-evaluators then take as input the summarization task prompt, and a (dialogue, summary) pair, to generate rating, explanation, and citation according to the metrics defined in Section 1. Note that we use a held-out set of 50 examples in this experiment so there exists no data overlap with the ABCD subsets used during training.

Unlike the general *RAG citation* benchmarks evaluated in Section 6.1, which contain human annotated citation groundtruth, there exists no well-established benchmarks to evaluate *content quality citation*. Hence, we seek human annotations to serve as a reference standard in this task. The machine outputs were labeled by 12 machine learning experts, each of whom has a graduate degree in computer science or related fields. They were asked to evaluate the correctness of each of the rating, explanation, and citation generated by the machine. For rating and explanation evaluation, the annotators were asked to provide "Yes/No" labels to determine the correctness of the response. For citation evaluation, they were asked to to provide manually-written citation for the claims in the generated explanation to serve as the groudtruth citation. (Details of the human labeling guideline can be found in Appendix F). Human evaluation on the Coherence metric was not performed because it is conceptually challenging to define what should be cited in a coherent response (e.g., citing an entire paragraph adds little interpretive value). Citation for coherence are more meaningful in negative cases where parts of the response are incoherent. To ensure a balanced evaluation with both positive and negative cases across metrics, we decided to exclude Coherence from the ABCD summarization content quality task.

We assign two labelers for each model and find the average inter-rater agreement is around 95%. We report the average accuracy for rating and explanation across the two sets of labels, and we compute F1 score between machine-generated citation and the intersection of the two sets of manual citation. From Table 3, we see that most LLM auto-evaluators obtain high rating and explanation accuracies, indicating that rating in "Yes/No" and free-form explanation generation are relatively easy tasks for the LLMs. For citation,

	Instruction-following			Completeness			Faithfulness		
	Rate Acc	Explain Acc	Citation (F1)	Rate Acc	Explain Acc	Citation (F1)	Rate Acc	Explain Acc	Citation (F1)
Mistral-7B	**1.00**	**1.00**	0.05	0.92	0.97	0.04	0.97	0.96	0.08
Mistral-Nemo	0.96	0.96	0.51	0.93	0.95	0.11	0.95	0.95	0.29
Llama-3.1-70B	0.98	0.98	**0.73**	0.96	0.97	0.54	0.97	0.96	0.37
Claude3-Opus	**1.00**	**1.00**	0.67	**0.98**	**0.99**	0.46	**0.98**	0.97	0.59
GPT-3.5	0.88	0.99	0.48	0.88	0.58	0.21	0.78	**0.99**	0.29
GPT-4	0.97	0.97	0.38	0.96	0.96	0.38	0.96	0.96	0.28
GPT-4o	0.95	0.95	0.35	0.93	0.93	0.56	0.96	0.71	0.59
REC-8B	**1.00**	0.98	0.48	0.94	0.95	0.60	0.95	0.98	0.59
REC-12B	**1.00**	**1.00**	0.50	0.94	0.95	0.62	0.94	0.98	0.60
REC-70B	**1.00**	**1.00**	0.50	0.96	0.96	**0.67**	0.97	0.98	**0.63**

Table 3. Human evaluation results on ABCD summarization content quality evaluation task.

our REC models perform the best in both completeness and faithfulness metrics, while in instruction-following, they are ranked second to llama-3.1-70B and Claude3-Opus. Specifically, for the instruction-following metric, we find REC models often cite slightly more sentences to ensure a good coverage of the entire conversation between a customer and an agent. In contrast, Llama-70B and Claude3 seem to be more selective.

6.3. General Capabilities

RewardBench The RewardBench dataset assesses LLMs' performance across several critical abilities using curated chosen-rejected response pairs (Lambert et al., 2024). The evaluation involves determining the model's win percentage based on its ability to score the "chosen" response higher than the "rejected" one. Specifically, it's focused on testing chat performance, safety measures, and reasoning (both in terms of code and math skills), while controlling for potential biases such as overfitting to prior datasets.

We first create prompts following the RewardBench evaluation format, where each prompt consisting of a query and two potential responses. The task for LLMs is to evaluate the two responses and determine which is better. The performance of the reward models is quantified by calculating the win percentage for chosen-rejected pairs associated with each prompt. A "win" is defined by a scenario where an LLM presents a preference for the selected answer. This quantitative measure provides a clear benchmark for evaluating model performance across tasks. By averaging the results across the 4 core sections: Chat, Chat Hard, Safety, and Reasoning, a comprehensive measure of each model's win percentage is derived. As shown in Table 4, our REC models are able to achieve SOTA performance among all generative LLMs on the RewardBench Leaderboard, demonstrating their outstanding ability to serve as an auto-evaluator for pairwise comparisons.

LLM-AggreFact This is a benchmark for measuring the grounding capabilities of auto-evaluators. Given a reference document and a claim, the auto-evaluator determines if the claim is fully supported by the document. This holistic benchmark combines 10 attribution datasets used in recent studies on LLM factuality. Given some of the models such as GPT-3.5 have smaller context windows, we removed examples longer than 16K tokens from our evaluation. We use the same prompt instructions across categories, as shown in Appendix G. To reduce model API costs, we randomly sampled 256 examples per evaluation task similar to the approach used in the FLAMe (Vu et al., 2024b).

Table 5 presents our attribution results on LLM-AggreFact (Tang et al., 2024b), categorized into four common use-cases: (1) LLM-FactVerify: fact verification of LLM-generated responses, (2) Wiki-FactVerify: evaluating correctness of Wikipedia claims, (3) Summarization: assessing faithfulness of summaries, and (4) Long-form QA: evaluating long-form answers to questions. REC-12B and REC-70B outperform all other models in all four categories, and REC-8B outperform all the other similar-sized models (Mistral-7B and Mistral-Nemo), demonstrating their strong performance in attribution on diverse datasets.

6.4. Bias Testing

CoBBLEr Recent studies have found that LLM-as-a-Judge often exhibits cognitive biases, such as preferences for verbosity, egocentrism, bandwagon, and an overly authoritative tone (Wang et al., 2024a; Koo et al., 2024; Zheng et al., 2024; Chen et al., 2024). To investigate the biases of the compared models, we evaluate them on the CoBBLEr benchmark (Cognitive Bias Benchmark for LLMs as EvaluatoRs) (Koo et al., 2024). This dataset is designed to evaluate the quality and reliability of LLMs when used as automated evaluators in a question-answering (QA) setting. It assesses the presence of six cognitive biases, both implicit and induced, when LLMs are tasked with ranking responses generated by various other models. CoBBLEr's core objective is to identify the extent of bias in LLM evaluation outputs.

The CoBBLEr dataset includes 50 QA instructions, randomly selected from two well-established benchmarks: BIG-bench (bench authors, 2023) and ELI5 (Fan et al., 2019).

	Chat	Chat Hard	Safety	Reasoning	AVERAGE
Mistral-7B	76.10	54.30	81.50	53.90	66.50
Mistral-Nemo	86.60	59.60	74.80	70.30	72.80
Llama-3.1-70B	97.60	58.90	73.00	78.50	77.00
Claude-3-Opus	94.70	60.30	86.60	78.70	80.10
GPT-3.5	92.20	44.50	65.50	59.10	65.30
GPT-4	95.30	74.30	87.60	86.90	86.00
GPT-4o	96.10	76.10	88.10	86.60	86.70
REC-8B	92.18	85.96	88.24	91.66	89.51
REC-12B	91.90	86.60	92.00	93.60	91.00
REC-70B	94.10	90.10	93.20	96.40	**93.50**

Table 4. General LLM capabilities evaluation on RewardBench. Scores for all the compared LLMs are obtained from the RewardBench Leaderboard (Lambert et al., 2024).

	LLM-FactVerify	Wiki-FactVerify	Summarization	Long-form QA
Mistral-7B	68.04	56.82	55.80	62.80
Mistral-Nemo	70.18	65.27	62.14	60.88
Llama-3.1-70B	75.25	68.84	69.56	76.98
Claude-3-Opus	76.82	76.52	70.65	76.82
GPT-3.5	72.81	69.53	68.60	70.40
GPT-4	77.44	77.22	72.45	76.92
GPT-4o	76.31	79.48	72.18	75.22
REC-8B	78.04	72.46	66.27	72.73
REC-12B	<u>78.78</u>	<u>85.34</u>	<u>72.51</u>	<u>77.23</u>
REC-70B	**79.67**	**87.48**	**74.04**	**79.12**

Table 5. LLM-AggreFact performance across four common use-cases: LLM-FactVerify (ClaimVerify + FactCheck + Reveal), Wiki-FactVerify (WiCE), Summarization (AggreFact + TofuEval), and Long-form QA (ExpertQA + LFQA). The best and the second best performance for each metric are in bold and underlined separately.

16 LLMs, both open and closed-source models, generate responses to these instructions. The evaluations involve pairwise comparisons between the responses of two models, wherein each model also acts as an evaluator to rank its own and others' outputs. The biases tested are categorized into two groups: (1) Implicit biases, such as egocentric bias (where a model tends to prefer its own outputs), and (2) Induced biases, such as order bias, where the ranking of responses is influenced by their order in the evaluation.

Table 6 presents the performance of the compared models on the CoBBLEr benchmark with metrics of order bias, bandwagon effect, compassion, selective bias, salience, distraction, and frequency. Note that lower scores indicates better performance (i.e., fewer biases) on CoBBLEr. Overall, our REC-70B model has the best average performance with a score of 0.2141, followed closely by GPT-4 (0.2279) and GPT-4o (0.2349). REC-12B and REC-8B, on the other hand, outperforms all the other same-sized and smaller models (Mistral-Nemo and Mistral-7B), and even some larger models (Llama-3-70B and GPT-3.5). The results suggest that training with REC produces an LLM judge with fairer evaluation capabilities, independent of the cognitive biases tested, such as verbosity, order, or an overly authoritative tone. In contrast, off-the-shelf LLMs can be more influenced by factors such as the order of responses or the length of the text.

7. Conclusions

LLMs excel in generating coherent and high-quality text for various tasks, but evaluating content quality is challenging due to issues like factual inaccuracies and hallucination. This paper introduces three fine-tuned models—REC-8B, REC-12B, and REC-70B—designed to evaluate generated text across several dimensions: faithfulness, instruction following, coherence, completeness. These models not only provide ratings for these metrics but also offer detailed explanations and verifiable citation, i.e., Ratings, Explanations, and citation (REC), thereby enhancing trust in the content. To achieve this goal, a curated dataset for explanations and citation fine-tuning is proposed to facilitate future research, which is the first public dataset containing both content quality citation and RAG citation. The models support various citation modes, including Post-fix citation, Post-fix citation mode with snippet, Inline citation, and Inline citation mode with snippet to balance the trade-off between latency and granularity of citation. Extensive evaluations on diverse benchmarks (i.e., ALCE, ExpertQA, ABCD summarization, RewardBench, LLM-AggreFact, and CoBBLEr) demonstrate that the fine-tuned LLM auto-evaluator, REC-70B, outperforms state-of-the-art models. It excels in content evaluation, delivering clear explanations and ensuring citation accuracy.

	Order	Bandwagon	Compassion	Selective	Salience	Distraction	Frequency	AVERAGE
Mistral-7B	0.163	0.518	0.346	0.671	0.603	0.0143	0.1406	0.3508
Mistral-Nemo	0.126	0.543	0.238	0.725	0.595	0.0196	0.0059	0.3218
Llama-3.1-70B	0.115	0.251	0.318	0.545	0.599	0.0116	0.0705	0.2729
Claude-3-Opus	0.078	0.081	0.380	0.483	0.589	0.0093	0.0087	0.2327
GPT-3.5	0.164	0.877	0.371	0.851	0.644	0.0029	0.0183	0.4183
GPT-4	0.092	0.098	0.399	0.431	0.569	0.0015	0.0046	0.2279
GPT-4o	0.075	0.086	0.417	0.468	0.578	0.0089	0.0112	0.2349
REC-8B	0.105	0.273	0.311	0.582	0.590	0.0119	0.0052	0.2683
REC-12B	0.089	0.225	0.291	0.526	0.583	0.0108	0.0048	0.2471
REC-70B	0.073	0.095	0.258	0.487	0.572	0.0091	0.0043	**0.2141**

Table 6. Bias evaluation on CoBBLEr (lower is better).

Acknowledgements

We appreciate Yilun Zhou, Shafiq Rayhan Joty, Peifeng Wang, and Austin Xu from Salesforce AI research for insightful discussions and help to benchmark model performance. We gratefully acknowledge partial support from NSF grant DMS-2413265, NSF grant 2023505 on Collaborative Research: Foundations of Data Science Institute (FODSI), the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639, NSF grant MC2378 to the Institute for Artificial CyberThreat Intelligence and OperatioN (ACTION), and a Berkeley Deep Drive (BDD) grant from BAIR and a Dean's fund from CoE, both at UC Berkeley.

Limitations

Our work introduces the innovative **R**ate, **E**xplain, and **C**ite (**REC**) model for trusted automated AI evaluations, offering significant benefits but with a few limitations that future research will address. First, our work is limited to the baselines we compare, as non-LLM citation methods, for instance, approaches that leverage sentence embedding (Reimers & Gurevych, 2019) similarity to identify and cite the most semantically relevant sources were not included. Including such techniques could serve as valuable baselines to further validate our model's generalization and effectiveness. Second, the need for post-generation explanations and citation may introduce latency. Future work will compare latencies across models and explore optimizations to improve speed. We are also considering fine-tuning the snippet output using starting and ending character numbers to optimize for latency. Finally, even though the 12B base model is capable of generating multi-lingual output, our current evaluations are limited to English datasets. Expanding to multilingual evaluations is a key goal for future research.

Impact Statement

Our contributions present a novel general-purpose auto-evaluator that supports various modes of citation, as well as a dataset with citation and explanations to support further research. There are a few ethical considerations for users of such technology. As with the use of fine-tuning and RAG, it is important to be aware of potential propagating biases and feedback loops. Biased data used for training and/or retrieval could lead to model outputs that reinforces such biases, impacting model quality and fairness. Although we extensively evaluated our LLM auto-evaluator and found it to be safe, unbiased, and factual, users should consider whether such considerations hold true for the input source itself. Of note, even established sources such as knowledge banks may contain inaccurate or inconsistent information. Finally, though we aimed to be transparent and consistent for our data labeling tasks, using clear instructions and validations, note that data labeling tasks are inherently subject to the bias of labeler background and experiences.

References

Bai, Y., Ying, J., Cao, Y., Lv, X., He, Y., Wang, X., Yu, J., Zeng, K., Xiao, Y., Lyu, H., Zhang, J., Li, J., and Hou, L. Benchmarking foundation models with language-model-as-an-examiner. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

bench authors, B. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T. W., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Si-

monyan, K., Rae, J. W., Elsen, E., and Sifre, L. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:244954723.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., teusz Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. URL https://api.semanticscholar.org/CorpusID:218971783.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J. A., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y.-F., Lundberg, S. M., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712, 2023. URL https://api.semanticscholar.org/CorpusID:257663729.

Cao, S. and Wang, L. Verifiable generation with subsentence-level fine-grained citations. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 15584–15596, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.920. URL https://aclanthology.org/2024.findings-acl.920.

Chen, A., Pasupat, P., Singh, S., Lee, H., and Guu, K. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *ArXiv*, abs/2305.14908, 2023. URL https://api.semanticscholar.org/CorpusID:258865339.

Chen, D., Fisch, A., Weston, J., and Bordes, A. Reading Wikipedia to answer open-domain questions. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL https://aclanthology.org/P17-1171.

Chen, D., Chen, H., Yang, Y., Lin, A., and Yu, Z. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3002–3017, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.239. URL https://aclanthology.org/2021.naacl-main.239.

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. Humans or llms as the judge? a study on judgement biases. *ArXiv*, abs/2402.10669, 2024.

Chiang, C.-H. and Lee, H.-y. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.870.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N. M., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., García, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Díaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K. S., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022. URL https://api.semanticscholar.org/CorpusID:247951931.

Dubois, Y., Liang, P., and Hashimoto, T. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. ELI5: Long form question answering. In Korhonen, A., Traum, D., and Màrquez, L. (eds.),

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL https://aclanthology.org/P19-1346.

Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V., Lao, N., Lee, H., Juan, D.-C., and Guu, K. Rarr: Researching and revising what language models say, using language models. *ArXiv*, abs/2210.08726, 2022.

Gao, T., Yen, H., Yu, J., and Chen, D. Enabling large language models to generate text with citations. *ArXiv*, abs/2305.14627, 2023. URL https://api.semanticscholar.org/CorpusID:258865710.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey,

M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pre-training. *ArXiv*, abs/2002.08909, 2020. URL https://api.semanticscholar.org/CorpusID:211204736.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Yu, J. A., Joulin, A., Riedel, S., and Grave, E. Few-shot learning with retrieval augmented language models. *ArXiv*, abs/2208.03299, 2022. URL https://api.semanticscholar.org/CorpusID:251371732.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1 – 38, 2022. URL https://api.semanticscholar.org/CorpusID:246652372.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b. *ArXiv*, abs/2310.06825, 2023a. URL https://api.semanticscholar.org/CorpusID:263830494.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023b. URL https://arxiv.org/abs/2310.06825.

Jiang, D., Li, Y., Zhang, G., Huang, W., Lin, B. Y., and Chen, W. TIGERScore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=EE1CBKC0SZ.

Kamalloo, E., Jafari, A., Zhang, X. C., Thakur, N., and Lin, J. J. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *ArXiv*, abs/2307.16883, 2023. URL https://api.semanticscholar.org/CorpusID:260334522.

Koo, R., Lee, M., Raheja, V., Park, J. I., Kim, Z. M., and Kang, D. Benchmarking cognitive biases in large language models as evaluators. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 517–545, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.29. URL https://aclanthology.org/2024.findings-acl.29.

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., Smith, N. A., and Hajishirzi, H. Rewardbench: Evaluating reward models for language modeling. https://huggingface.co/spaces/allenai/reward-bench, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401, 2020. URL https://api.semanticscholar.org/CorpusID:218869575.

Li, J., Sun, S., Yuan, W., Fan, R.-Z., Zhao, H., and Liu, P. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*, 2023a.

Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023b.

Liu, N. F., Zhang, T., and Liang, P. Evaluating verifiability in generative search engines. *ArXiv*, abs/2304.09848, 2023a. URL https://api.semanticscholar.org/CorpusID:258212854.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153.

Malaviya, C., Lee, S., Chen, S., Sieber, E., Yatskar, M., and Roth, D. ExpertQA: Expert-curated questions and attributed answers. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2024. URL https://openreview.net/forum?id=hhC3nTgfOv.

Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G., and McAleese, N. Teaching language models to support answers with verified quotes. *ArXiv*, abs/2203.11147, 2022. URL https://api.semanticscholar.org/CorpusID:247594830.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Long, O., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332, 2021. URL https://api.semanticscholar.org/CorpusID:245329531.

OpenAI. Gpt-4 technical report. In *GPT-4 Technical Report*, 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N.,

Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L. E., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. J. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. URL https://api.semanticscholar.org/CorpusID:246426909.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. Retrieval augmentation reduces hallucination in conversation. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://api.semanticscholar.org/CorpusID:233240939.

Sun, H., Cai, H., Wang, B., Hou, Y., Wei, X., Wang, S., Zhang, Y., and Yin, D. Towards verifiable text generation with evolving memory and self-reflection. *ArXiv*, abs/2312.09075, 2023. URL https://api.semanticscholar.org/CorpusID:266210156.

Tang, L., Laban, P., and Durrett, G. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024a. URL https://arxiv.org/pdf/2404.10774.

Tang, L., Laban, P., and Durrett, G. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*, 2024b.

Team, L. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. M., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A. S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I. M., Korenev, A. V., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. URL https://api.semanticscholar.org/CorpusID:259950998.

Vu, T., Krishna, K., Alzubi, S., Tar, C., Faruqui, M., and Sung, Y.-H. Foundational autoraters: Taming large language models for better automatic evaluation. *ArXiv*, abs/2407.10817, 2024a. URL https://api.semanticscholar.org/CorpusID:271213398.

Vu, T., Krishna, K., Alzubi, S., Tar, C., Faruqui, M., and Sung, Y.-H. Foundational autoraters: Taming large language models for better automatic evaluation. *arXiv preprint arXiv:2407.10817*, 2024b.

Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J., and Zhou, J. Is ChatGPT a good NLG evaluator? a preliminary study. In Dong, Y., Xiao, W., Wang, L., Liu, F., and Carenini, G. (eds.), *Proceedings of the 4th New Frontiers in Summarization Workshop*, pp. 1–11, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.newsum-1.1. URL https://aclanthology.org/2023.newsum-1.1.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., and Sui, Z. Large language models are not fair evaluators. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511.

Wang, T., Yu, P., Tan, X. E., O'Brien, S., Pasunuru, R., Dwivedi-Yu, J., Golovneva, O., Zettlemoyer, L., Fazel-Zarandi, M., and Celikyilmaz, A. Shepherd: A critic for language model generation, 2023b.

Wang, Y., Yu, Z., Zeng, Z., Yang, L., Wang, C., Chen, H., Jiang, C., Xie, R., Wang, J., Xie, X., Ye, W., Zhang,

S., and Zhang, Y. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *International Conference on Learning Representations (ICLR)*, 2024b.

Wang, Z., Dong, Y., Delalleau, O., Zeng, J., Shen, G., Egert, D., Zhang, J. J., Sreedhar, M. N., and Kuchaiev, O. Helpsteer2: Open-source dataset for training top-performing reward models. *CoRR*, abs/2406.08673, 2024c. URL http://dblp.uni-trier.de/db/journals/corr/corr2406.html#abs-2406-08673.

Xu, W., Wang, D., Pan, L., Song, Z., Freitag, M., Wang, W., and Li, L. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5967–5994, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.365. URL https://aclanthology.org/2023.emnlp-main.365.

Ye, X., Sun, R., Arik, S. Ö., and Pfister, T. Effective large language model adaptation for improved grounding and citation generation. In *North American Chapter of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:265220884.

Yuchen Lin, B., Deng, Y., Chandu, K., Brahman, F., Ravichander, A., Pyatkin, V., Dziri, N., Le Bras, R., and Choi, Y. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv e-prints*, pp. arXiv–2406, 2024.

Zhang, J., Li, Z., Das, K., Malin, B., and Sricharan, K. Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusID:265019095.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023. URL https://api.semanticscholar.org/CorpusID:259129398.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Zhu, L., Wang, X., and Wang, X. Judgelm: Fine-tuned large language models are scalable judges. *arXiv*, 2023.

A. Content Quality Citation Modes and the Evaluation Prompt

Here we provide the evaluation prompt template used in Figure 1, the raw auto-evaluator output, and a final output example in an alternative citation mode for content quality evaluation.

A.1. Content Quality Evaluation Prompt

```
You will be asked to provide
feedback on one quality metric for a task
prompt given to another LLM and that
LLM's output. Given the task prompt as a
context, your job is to give feedback on
whether there are any errors in the output,
based on the quality metric. If there
are no errors in the output, please
also provide your feedback that there are
no errors. Remember your main goal is to
provide feedback so that another LLM can
use your feedback to improve its
generation.
You need to support statements in your
feedback that can be linked back to ANOTHER
LLM'S TASK PROMPT with citations.
Please make sure you read and understand
these instructions carefully. Please keep
this document open while reviewing, and
refer to it as needed.
Please only generate your answer following
the Output JSON format. Do not generate
anything else.

Evaluation criteria:
{metric_name} ({metric_scale})
- {metric_description}

Respond using the following JSON format
based on the provided ANOTHER LLM'S
TASK PROMPT and ANOTHER LLM'S OUTPUT:
Desired format:
{
   "answer":<answer>,
   "feedback":<generated feedback>,
   "statements":[
   {
       "statement_string":<one statement
       extracted from feedback>,
       "citations":[
       {
            "snippet":<snippet>
       }]
    }]
}

### INSTRUCTIONS:
Here are the instructions to follow to
generate the response.

1. Generate a feedback based on ANOTHER
LLM'S TASK PROMPT and ANOTHER LLM'S
OUTPUT. Explain whether the evaluation
criteria is met.
    - Set "answer" to "Yes" if the
    evaluation criteria is met,
    and "No" otherwise.
    - Set "feedback" to the
    generated feedback.
2. For each statement in your feedback,
extract relevant sentences that support
the statement from ANOTHER LLM'S
TASK PROMPT.
3. For each extracted sentence:
    - Set "snippet" to be the extracted
    sentence.
4. Compile relevant snippets for the
statement:
    - Set "statement_string" to be
    a statement in the generated feedback.
    - Set "citations" to be a list of JSON
    objects containing all relevant
    snippets (from step 3) supporting the
    "statement_string".
5. Compile the statements into a list of
JSON objects.
    - Set "statements" to be a list of
    JSON objects containing metadata of
    all statements (from step 4) in the
    feedback.

[BEGIN DATA]
### ANOTHER LLM'S TASK PROMPT:
Summarize knowledge from transcripts
after they've ended. Summarizing the key
points of the conversation including
customer issue and resolution.

### Conversation:
{convo}

### ANOTHER LLM'S OUTPUT:
{answer}
[END DATA]

### Response(JSON only):
```

A.2. Content Quality Citation Modes

Using the same input example as in Figure 1, we show the raw auto-evaluator output that allows users to further process according to either of the two citation modes, "post-fix with context snippet" and "inline with context snippet", to fit their needs. With the final output in "inline with context snippet" mode already presented in Figure 1, here we show another possible final output format in "post-fix with context snippet" mode.

A.2.1. Raw auto-evaluator Output

Citations are extracted verbatim from the input context to support statements in the auto-evaluator's explanation.

```
"statements": [
 {
```

```
   "statement_string": "The answer covers
   all the key points of the conversation,
   including the customer's question, the
   agent's response, and the resolution.",
   "citations": ["Summarizing key points
   of the conversation including customer
   issue and resolution.", "Customer: Is it
   possible to have my items gift wrapped?"
   ,"Agent: We offer gift wrapping but it
   cost $4.99 per item"]
 },
 {
   "statement_string": "It also mentions the
   customer's reason for asking about gift
   wrapping and the agent's polite closing."
   ,"citations": ["Customer: Oh great. My
   Bestie's birthday is coming up, that's
   the reason for my questions. Thanks",
   "Agent: All right. thanks for
   reaching out and have a great day!"]
  }
]
```

A.2.2. Final Output Example in "post-fix with context snippet" Citation Mode

```
The answer covers all the key points of the
conversation, including the customer's
question, the agent's response, and the
resolution. It also mentions the customer's
reason for asking about gift wrapping and
the agent's polite closing.
[1][2][3][4][5]

[1]: "Summarize key points of the
conversation including customer issue
and resolution."
[2]: "Customer: Is it possible to have
my items gift wrapped?"
[3]: "Agent: We offer gift wrapping but
it costs $4.99 per item."
[4]: "Customer: Oh great. My Bestie's
birthday is coming up, that's the reason
for my questions. Thanks."
[5]: "Agent: All right. thanks for
reaching out and have a great day!"
```

B. Details of using REC Models for General RAG citations

We demonstrate how REC models can also be used for general RAG citations in Figure 2. In the example shown, we present this question "What is photosynthesis?" to another LLM, and collect retrieved articles with ids as the context, and the generated answer from another LLM to serve as inputs. The goal is for our LLM auto-evaluator to provide citations extracted verbatim from the context to support the generation from another LLM. REC models support four different citation modes, and the details can also be found in this section.

- **Retrieved Context**:

  ID 1233

  Photosynthesis is the process by which green plants and some other organisms utilize sunlight to synthesize their food. This remarkable process involves the conversion of carbon dioxide and water into glucose and oxygen, facilitated by chlorophyll. It serves as the foundation for energy production in these organisms and plays a crucial role in maintaining the balance of oxygen in our atmosphere.

  ID 1422

  Cellular respiration refers to a series of metabolic reactions that occur within the cells of organisms. This process transforms biochemical energy from nutrients into adenosine triphosphate (ATP), which is essential for cellular functions. During cellular respiration, waste products are released, and oxygen is a vital element needed for the process to function efficiently. This energy conversion is critical for sustaining life.

  ID 4431

  DNA replication is the mechanism by which a cell duplicates its DNA in preparation for cell division. This process ensures that genetic information is accurately copied and passed on to daughter cells. The enzyme DNA polymerase is instrumental in DNA replication, playing a crucial role in maintaining genetic fidelity and continuity across generations.

- **LLM Generated Response to Cite**:

  Photosynthesis is a process that converts carbon dioxide and water into glucose and oxygen using sunlight and chlorophyll. This process occurs in green plants and certain other organisms.

B.1. RAG Citation Prompt

As an example, below is the prompt that we used to generate inline citations. This can be modified to generate responses in each of the different RAG citation modes shown in the following sections.

```
Your task is to provide
citations for a generated response
from an LLM for a RAG
(Retrieval Augmented Generation)
application. You will be provided two
sections below. The RETRIEVED CHUNKS
section is the context from
where you will generate citations.
The LLM GENERATED ANSWER section
is for the LLM generated answer
from where you will generate claims.

###
Respond using the following JSON format:
```

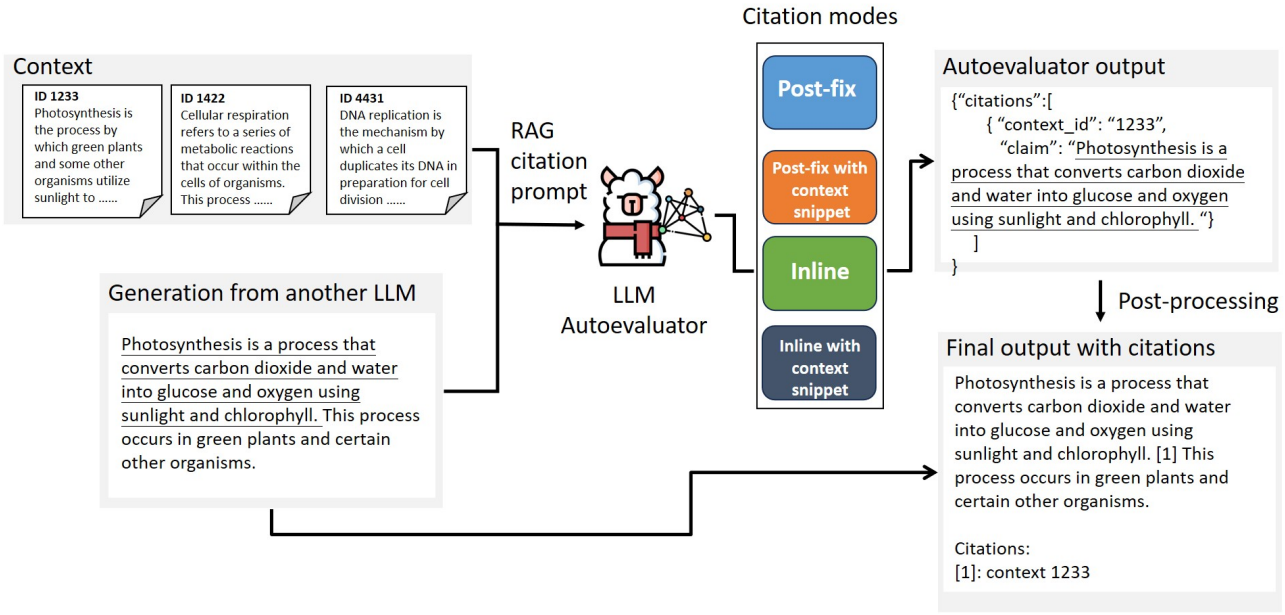

Figure 2. Illustration of using REC models for general RAG citations.

```
Desired format:

{
   "citations":[
      {
         "context_id":"<context_id>",
         "claim": "<claims identified from
         LLM GENERATED ANSWER>"
      }
   ]
}

###INSTRUCTIONS
In order to generate citations,
follow these steps below:

1. Review the "LLM GENERATED ANSWER"
section below and extract a set of claims,
VERBATIM and set the "claim" in the output
json.
2. Your task is to generate  a citation
for each of the claims identified
in Step 1, by following the steps below:
    i. For each claim, identify the
    most relevant chunk and the
    corresponding "context_id" from
    the "RETRIEVED CHUNKS" section that
    supports the claim.
    ii. For the identified chunk,
    set the "context_id" in the output
    to the corresponding context_id.
    Set "context_id" to "None" if none
    of the chunks in the
    "RETRIEVED CHUNKS" section supports
    the claim.
3. Ensure that you structure your
response to adhere to the desired
output JSON format.

###
RETRIEVED CHUNKS:
{retrieved_chunks}

###
LLM GENERATED ANSWER:
{answer}

###
Response (JSON only):
```

B.2. Post-fix Citation

```
{
  "citations": [
    {
      "context_id": "1233"
    }
  ]
}
```

B.3. Post-fix Citation with Context Snippet

```
{
  "citations": [
    {
      "context_id": "1233",
         "snippet": "Photosynthesis is
         the process by which green
         plants and some other organisms
         utilize sunlight to synthesize
```

```
        their food. This remarkable
        process involves the
        conversion of carbon
        dioxide and water into
        glucose and oxygen,
        facilitated by chlorophyll"
    }
  ]
}
```

B.4. Inline Citation

```
{
  "citations": [
    {
      "context_id": "1233",
         "claim": "Photosynthesis is a
         process that converts carbon
         dioxide and water into glucose
         and oxygen using sunlight and
         chlorophyll."
    }
  ]
}
```

B.5. Inline Citation with Context Snippet

```
{
  "citations": [
    {
      "context_id": "1233",
       "claim": "Photosynthesis is a
         process that converts carbon
         dioxide and water into glucose
         and oxygen using sunlight and
         chlorophyll.",
       "snippet": "Photosynthesis is the
         process by which green
         plants and some other
         organisms utilize sunlight
         to synthesize their food. This
         remarkable process involves
         the conversion of carbon
         dioxide and water into
         glucose and oxygen,
         facilitated by chlorophyll"
    }
  ]
}
```

C. REC-Data Details

For general instruction tasks, we include summarization on ABCD (Chen et al., 2021), a dataset containing customer support dialogue transcripts, in five languages. For pointwise evaluation, we leverage HelpSteer2 (Wang et al., 2024c) and perform synthetic data generation (Section 4.2) on the (dialogue, summary) pairs from ABCD, and QA pairs from FloDial. For pairwise and open-ended evaluation tasks, we use Auto-j (Li et al., 2023a), Shepherd (Wang et al., 2023b), Skywork [7], HelpSteer2[8], OffsetBias[9], and Code Preference[10]. For citations, we perform synthetic data generation (Section 4.2) on FloDial for both RAG citations and content quality citations, and on ABCD, Auto-j, Shepherd, and Email Thread Summary [11] for content quality citations.

Note that it's possible to derive different tasks from the same dataset when different task prompts are assigned to an LLM. We provide a detailed breakdown of REC-Data in Figure 3.

D. Synthetic Data Generation Prompt Example

The prompts used to generate synthetic content quality and RAG citation data are the same as the ones shown in Appendices A and B.

D.1. Pointwise Evaluation ("Faithfulness metric")

```
You will be given a source text given to
another LLM, and that LLM's answer.
Your task is to rate that LLM's answer
on one quality metric.
Please make sure you read and understand
these instructions carefully. Please
keep this document open while reviewing,
and refer to it as needed.


Evaluation criteria:
metric_name = 'Faithfulness'
metric_scale = 'Yes/No'
metric_description = ""
The generated answer only contains truthful
content, and does not contain invented
or misleading facts that are
not supported by the context.
""


Evaluation steps:
1. Read the source text.
2. Read the other LLM's answer and compare
it to the source text. Check if the
evaluation criteria is met.
3. If the evaluation criteria is met,
answer Yes; if the evaluation criteria
is not met, answer No.
```

[7] https://huggingface.co/datasets/Skywork/Skywork-Reward-Preference-80K-v0.1

[8] https://huggingface.co/datasets/nvidia/HelpSteer2

[9] https://huggingface.co/datasets/NCSOFT/offsetbias

[10] https://huggingface.co/datasets/Vezora/Code-Preference-Pairs

[11] https://www.kaggle.com/datasets/marawanxmamdouh/email-thread-summary-dataset/data

```
4. Create a JSON response
{
    "metriclabel": <Yes|No only>,
    "justification":
    <explanation for the score>
}

Example:
Source text given to another LLM:
{query_with_context}

The other LLM's answer:
{answer}

### Response(JSON Only):
```

E. Remark on Why The Three Metrics in ALCE are All Important

Regarding why the three criteria are *equally important*, we encourage the reader to consider from an *user's perspective*. An LLM's citation response cannot be deemed good if it is lacking in any of the three criteria: fluency, answer correctness, and citation accuracy. To illustrate this, we aim to provide an example of a good citation response that scores well on all three criteria, as well as three examples that each fail in one of the criteria (assuming the same RAG QA setting as in Appendix B).

- **Example 1 (scoring high in all criteria):** Photosynthesis is the process by which green plants and some other organisms utilize sunlight to synthesize their food. This remarkable process involves the conversion of carbon dioxide and water into glucose and oxygen, facilitated by chlorophyll [1233].
- **Example 2 (lacks fluency):** Photosynthesis, it's the way that green plants and also some other organisms they use (*not fluent) sunlight for making their food. This process, it's interesting, converts carbon dioxide with water into glucose and oxygen, and chlorophyll helps with it [1233].
- **Example 3 (lacks correctness):** Photosynthesis, it's when green plants and other organisms use water (*wrong) to directly create sunlight (*wrong) as food. This process, it's interesting, turns oxygen into glucose and carbon dioxide, with chlorophyll doing nothing important.
- **Example 4 (lacks citation accuracy):** Photosynthesis is the process by which green plants and some other organisms utilize sunlight to synthesize their food. This remarkable process involves the conversion of carbon dioxide and water into glucose and oxygen, facilitated by chlorophyll [1422] (*wrong citation).

It is highly likely that a user would prefer the first example, which is well-balanced across all three criteria, as a good response. We hope this illustration better conveys our point.

F. ABCD Content Quality Citation Human Evaluation Instruction

Annotation instruction: You will be evaluating attributed judgments made by an LLM auto-evaluator for answers generated by another LLM against three metrics: instruction-following, completeness, and faithfulness. Your task is to evaluate LLM auto-evaluator's answer by completing the following THREE tasks: *you can skip if the output shows "invalid json".

- **Metric answer correctness:** Please label "yes"/"no" by evaluating whether the metric "answer" in auto-evaluator's output is correct or not, based on the provided conversation, task prompt, and summary.
- **Metric explanation correctness:** Please label "yes"/"no' by evaluating whether the "feedback" in auto-evaluator's output is correct or not, based on the provided conversation, task prompt, and summary. Only proceed to citation annotation (next task) if you label "yes" in this task, namely the explanation is correct.
- **Citation annotation:** In "statements", you will find different "statement_string" extracted from the "feedback". Please provide citation for each "statement_string" by directly copying relevant sentences (Please cite full sentences, not substrings) from "task prompt" and "chat" columns (Please DON'T cite from "summary"). If the "statement_string" is hallucinated and doesn't make sense in the first place, please put "[¡halu¿]" as the citation to signal.

G. Prompt for LLM-AggreFact Eval

```
Determine whether the provided claim is
consistent with the corresponding document.
Consistency in this context implies that
all information presented in the claim is
substantiated by the document. If not,
it should be considered inconsistent.

Document: {doc}
Claim: {claim}

Please assess the claim's consistency
with the document by responding with
either "yes" or "no".

Answer (yes|no only):
```

Category	Dataset	Amount	Task Types	Modes
General Instruction	**Multi-lang ABCD**	3000	summarization	languages: de, es, it, ja, pt
Evaluation	**Shepherd**	1317	Explanation	
	Auto-J	4396	Explanation + rating / Pairwise response preference explanation	single (960), pairwise (3436)
	Skywork	82000	Pairwise response preference	
	HelpSteer2	10161	Pairwise response preference	
	OffsetBias	8504	Pairwise response preference	
	Code Preference	10000	Pairwise response preference	
	ABCD	2000	Metric labels	
	FloDial QA	2000	Metric labels	
	HelpSteer2	5000	Metric labels	
Citations	**FloDial QA**	2166	RAG citations	post-fix, post-fix with snippet, inline, inline with snippet
	Email thread synthetic citations	2206	Content quality citations: Metric labels + explanation + citations	coherence (579), completeness (412), faithfulness (524), instruction-following (691)
	FloDial RAG citations	1720	Content quality citations: Metric labels + explanation + citations	coherence (109), completeness (385), faithfulness (607), instruction-following (619)
	FloDial QA	428	Content quality citations: Metric labels + explanation + citations	coherence (119), completeness (83), faithfulness (27), instruction-following (199)
	ABCD	564	Content quality citations: Metric labels + explanation + citations	coherence (293), completeness (38), faithfulness (87), instruction-following (146)
	Auto-J	1871	Content quality citations: Metric labels + explanation + citations	coherence (123), completeness (418), faithfulness (509), instruction-following (821)
	Shepherd	2413	Content quality citations: Metric labels + explanation + citations	coherence (193), completeness (461), faithfulness (748), instruction-following (1011)
	Total	139,746		

Figure 3. Detailed breakdown of REC-Data distribution.