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

Bhuvanashree Murugadoss[1], Christian Poelitz[2], Ian Drosos[2], Vu Le[1], Nick McKenna[2], Carina Suzana Negreanu[1], Chris Parnin[1,3], Advait Sarkar[1]

[1]Microsoft
[2]Microsoft Research Cambridge
[3]North Carolina State University
cpoelitz@microsoft.com

Abstract

LLMs-as-a-judge is a recently popularized method which replaces human judgements in task evaluation with automatic evaluation using LLMs. Due to widespread use of RLHF (Reinforcement Learning from Human Feedback), state-of-the-art LLMs like GPT4 and Llama3 are expected to have strong alignment with human preferences when prompted for a quality judgement, such as the coherence of a text. While this seems beneficial, it is not clear whether the assessments by an LLM-as-a-judge constitute only an evaluation based on the instructions in the prompts, or reflect its preference for high-quality data similar to its fine-tune data. To investigate how much influence prompting the LLMs-as-a-judge has on the alignment of AI judgements to human judgements, we analyze prompts with increasing levels of instructions about the target quality of an evaluation, for several LLMs-as-a-judge. Further, we compare to a prompt-free method using model perplexity as a quality measure instead. We aggregate a taxonomy of quality criteria commonly used across state-of-the-art evaluations with LLMs and provide this as a rigorous benchmark of models as judges. Overall, we show that the LLMs-as-a-judge benefit only little from highly detailed instructions in prompts and that perplexity can sometimes align better with human judgements than prompting, especially on textual quality.

Introduction

Recently, new automatic evaluation approaches that rely on LLMs have been proposed on several NLG tasks, such as summarization (Liu et al. 2023b) and machine translation (Kocmi and Federmann 2023). Previous approaches (Siledar et al. 2024) show that for certain situations, such as when assessing textual consistency or fluency, there is high agreement between human judgements and LLM assessments, even without detailed instructions like for example how to assign specific scores. Most of these approaches prompt an LLM to give a judgement as a Likert score (Likert 1932) with only simple information about the scale, e.g. "give a judgement between 1 (bad) and 5 (good)." More recently, LLM-based evaluations on more fine-grained task-specific criteria (Ye et al. 2024) have also reported high agreement with human judgement, such as assessing the completeness of a solution for a question answering task. For these evaluations, often more detailed instructions are given about when to assign a specific score, similarly to rubric scoring (Andrade 2005).

While these results are promising for the future of automatic evaluation, it is less clear how the models achieve this agreement, and in general, it is a challenge to identify for any given task, which LLM is most appropriate to evaluate it, and with how much information, respectively how much instructions about the evaluation. Alarmingly, recent results show a clear bias in LLMs preferring their own output over others (Panickssery, Bowman, and Feng 2024), and LLMs' perplexity has emerged as a possible quality criteria for filtering (Ankner et al. 2024) on textual quality. This raises the question of whether some of the results on automatic evaluations with LMs reflect a model's preference for data similar to its own (high quality) fine-tuning data instead of following the provided instructions on how to measure the quality of an answer. Especially for fine-grained evaluations with detailed rubric information, we expect the instructions about when to assign a score to be adhered closely.

In this paper we report our findings when using LLMs-as-a-judge (Zheng et al. 2024), where LLMs are used as surrogates for human judgements to evaluate several NLG and LLM-based tasks, in skill-specific settings (e.g. completeness) and skill-unspecific (e.g. textual coherence). We show the annotations for many of these quality criteria in state-of-the-art benchmarks have a high correlation with the perplexity of the LLMs, often higher than prompting the LLM for a score. We identify which evaluation settings can benefit the most from more detailed prompting and for which settings simple generic prompts, or just using models' perplexity as quality score, suffice.

In detail, we make the following three main contributions:

1. We propose a novel taxonomy of qualitative evaluation criteria useful for assessing the competence of automatic evaluation methods by LLMs-as-a-judge. Our taxonomy consists of 4 evaluation categories (Content, Relevance, Integrity, and Engagement) which encapsulate 34 metrics as tested by 8 distinct state-of-the-art benchmarks.
2. We systematically evaluate the effectiveness of LLMs-as-a-judge using the taxonomy with several major LLM families including GPT4, Llama3, Mistral, and Phi3 across 4 levels of increasing prompt instruction. We find that, aggregated across the taxonomy, increasing instruc-

tion by including more granular evaluation rubrics only somewhat improves the Pearson correlation of models with human judgements, by only as much as 4%. However, some individual metrics may benefit.

3. We evaluate the potential of simple model perplexity as an alternative automatic evaluation to LLMs-as-a-judge. While perplexity often outperforms minimal prompting in terms of correlation with human judgements when detailed rubrics are not available, textual content-related metrics are the closest aligned category. For these metrics perplexity achieves a Pearson correlation of 0.51 in contrast to 0.44 when prompting the LLM-as-a-judge, suggesting it is the better choice for simple scenarios.

Related Work

LLMs as evaluators for general NLG (Liu et al. 2023b), as well as for knowledge and problem-solving tasks (Ye et al. 2024) have been widely studied recently (Chiang and Lee 2023; Li et al. 2024; Gao et al. 2024). Most previous approaches, either perform pair-wise evaluations (Ji et al. 2023; Chen et al. 2023), measuring the preference of one of two examples for a given criterion, or perform direct assessments for a single given example and a evaluation criterion (Liu et al. 2023b; Ye et al. 2024). Additionally, they distinguish between reference-free evaluations, where the LLM is presented only an example and the criterion for evaluation, and reference-based evaluations with given annotated examples of different qualities or ground truth for each example.

Generally, previous works use LLMs as evaluators by using simple prompting strategies (Siledar et al. 2024), only few fine-tuned models are available (Kim et al. 2023, 2024) for measuring for specific quality criteria. Most other fine-tuning approaches concentrate on scenario-specific quality feedbacks (Li et al. 2023; Wang et al. 2023) or on specific use-cases (McAleese et al. 2024).

Recently, there are several approaches (Liu, Moosavi, and Lin 2024; Liu et al. 2024, 2023c; Stureborg, Alikaniotis, and Suhara 2024; Doddapaneni et al. 2024) reporting biases and mismatches with human annotations, unreliability, as well as the need for task and dataset specific calibrations (Bavaresco et al. 2024).

Besides the ample prior works, our work is the first to study whether the models' perplexity can be a better surrogate for quality then prompting the corresponding model and whether instructions in the prompts are impacting the results across a number of different LLMs-as-a-judge.

Evaluating LLMs-as-a-judge

In this section we give a short definition of LLMs-as-a-judge and automatic evaluation using AI. We define different settings of prompting the LLMs-as-a-judge to measure the impact on the alignment of LLM judgments with human judgements. We also evaluate a prompt-free metric using simple model perplexity. This alternative approach requires no prompt engineering and transparently measures alignment with training data without bias from a prompt, so it is a compelling alternative for evaluation. Finally, we introduce a new taxonomy, aggregating the quality criteria most frequently used in state-of-the-art benchmarks for automatic evaluation with LLMs. We categorize these into 4 groups representing the major aspects of evaluating AI generated responses.

LLMs-as-a-judge

As LLM-as-a-judge we refer to the definition introduced by (Zheng et al. 2024) as potential replacement for human annotations by prompting an LLM for a judgement of an AI assistant response. We concentrate on judging textual examples only e.g., AI generated summaries for news articles (Fabbri et al. 2021) or step-by-step solution to mathematical reasoning questions (Golovneva et al. 2023). We phrase the task to judge an AI generated response as the following: Given a task A and an AI generated solution B, judge the quality of the solution B considering only the task A. In contrast to other previous approaches, we perform a reference-free evaluation where we do not provide a possible correct reference solution. We solely rely on the models' ability to judge the solution given only the task.

To measure the impact of prompting the LLM-as-a-judge, we study the LLMs' performance in **4** different settings:

1. **Perplexity**: We score each task solution by its perplexity under the corresponding LLM, given only the task description. This approach is unbiased by prompts, so it transparently measures alignment with model training data, providing a good comparison and alternative to prompt-based approaches.
2. **Generic quality prompt**: We prompt each LLM-as-a-judge with a basic instruction to measure the quality of the task solution, but give no specific criteria or instructions. In this case, we rely solely on the models' prior knowledge about the quality for the task solution from the examples generated.
3. **Criteria specific prompt**: We prompt each LLM-as-a-judge with an instruction to measure the quality for a specific criteria e.g., **coherence**. We only provide the name of the criteria, not a definition. We rely on the models' prior knowledge of the specific quality criteria only.
4. **Full rubric prompt**: We prompt each LLM-as-a-judge with an instruction to measure the quality for a specific quality criterion, together with a definition of the criterion and instructions when to assign each rubric score e.g., "*Score 1: Incoherent text with many logical flaws*."

We evaluate different LLMs-as-a-judge under the above settings on several different benchmarks (as described in the next subsection). We use the criteria as specified in the corresponding annotation guidelines from the benchmark datasets. For setting 4, we use all available annotation guidelines with information about the criteria and when to assign each score. We extract this information directly from benchmarks into a full rubric containing information about the criteria and the scores. For our experiments, we structure the settings from least instructive (Perplexity / no prompting) to most instructive (Full rubric information with instructions when to assign a score). In Fig. 1 we show the different setting of prompting for an example quality criterion.

2. Generic prompt

```
# Sample to evaluate
{example}

# Instructions
Evaluate the quality of the response from the sample and return a score between 1 (bad) and 5 (very good) as:
## Score: [Number]
```

3. Criteria specific prompt

```
# Sample to evaluate
{example}

# Rubrics
Logicality

# Instructions
Evaluate the quality of the response from the sample and return a score between 1 (bad) and 5 (very good) as:
## Score: [Number]
```

4. Full rubric prompt

```
# Sample to evaluate
{example}

# Rubrics
Logicality: Measure how much the story obeys your commonsense.
Score 1: The story is full of absurd things.
Score 2: The story has one or two things make sense, but generally very absurd.
Score 3: The story roughly makes sense.
Score 4: The story largely makes sense, except one or two things.
Score 5: The story totally complies with commonsense.

# Instructions
Evaluate the quality of the response from the sample and return a score between 1 and 5 as:
## Score: [Number]
```

Figure 1: Our prompting settings. We measure how much influence the information about the actual evaluation has for model performance as LLM-as-a-judge. For setting 1, perplexity, we don't prompt the models but calculate the models' perplexity for the task solution in the example instead. The example prompts shown above are used for the LLMs-as-a-judge to measure the quality for the criterion **logicality** as defined in the benchmark dataset **TheNextChapter**.

Datasets

We use 8 different open-source benchmark datasets commonly used for LLM-based evaluations with human annotations for several evaluation criteria per task. The datasets cover tasks which span several aspects from coarse-grained NLG-quality evaluations, to fine-grained very task specific evaluations with detailed information about how to score the example solutions.

Firstly, we leverage two of the most prominently used datasets for coarse-grained NLG-quality evaluations: The **SummEval** (Fabbri et al. 2021) dataset contains news article summaries generated by different models together with human annotations for 4 different quality criteria e.g., fluency; and the **TopicalChat** (Gopalakrishnan et al. 2019) dataset contains human conversations over 8 different topics annotated by humans for 5 different quality criteria e.g., engagement. Further, we use two more challenging benchmark datasets for coarse-grained NLG-evaluations: the **OpinSummEval** (Shen and Wan 2023) dataset is a opinion summarization dataset, which consists of review summaries annotated for aspects, opinions and sentiments; the **InstruSumm** (Liu et al. 2023a) dataset, consists of news article summaries following specific instructions with human annotations for content specific quality-criteria e.g., amount of missing information.

Second, we use two benchmark datasets for more fine-grained NLG evaluations: the **Hanna** (Chhun et al. 2022) dataset and the **TheNextChapter** (Xie, Cohn, and Lau 2023) dataset contain creative stories generated for a given initial user prompt. Each story is annotated by humans for NLG and style based criteria e.g., coherence, but also for more unconventional criteria like surprise. Finally, we use two task-specific evaluation benchmark datasets with quality-criteria depending in task solution quality: **Roscoe** (Golovneva et al. 2023) is a collection of datasets of reasoning tasks, together with GPT3 generated step-by-step solutions. The human annotations cover coarse-grained task specific evaluation criteria like "missing step"; the **Flask** (Ye et al. 2024) dataset contains several knowledge and problem solving tasks with LLM generated solutions. The human annotations cover more fine-grained task-specific criteria like completeness and factuality. To efficiently evaluate most criteria, the models need an understanding of the solution.

Criteria taxonomy

Previous approaches are mainly split between textual quality criteria and correctness judgments concerning LLMs-as-a-judge. These approaches miss a common classification of the evaluation criteria, making it difficult to study how the prompts with instructions influence the judgements generated by the models across them. We introduce a simple taxonomy based on current state-of-the-art benchmark datasets and quality criteria commonly used for automatic evaluations by LLMs. We define 4 groups of quality criteria, relevant for automatic evaluation:

1. **Content-based criteria**: Measure how well the solution is presented to the user, for example, whether a news article summary is fluent.
2. **Engagement-based criteria**: Measure how engaging the solution is, for example, whether a generated story contains an element of surprise.

Figure 2: Taxonomy of quality criteria summarizing current state-of-the-art benchmark datasets and criteria used for automatic evaluations with LLMs. We group all 34 quality criteria as defined in the 8 different benchmark datasets into 4 groups: **Content-based, Engagement-based, Integrity-based, Relevance-based criteria**.

3. **Integrity-based criteria**: Measure how consistent and logical coherent the solution is, for example, whether a math solution is correct.
4. **Relevance-based criteria**: Measure how relevant the solution is for the given task, for example, whether a legal advice answer contains irrelevant information.

We assign each of the criteria used in the benchmark datasets above to these 4 groups (Fig. 2) based on the descriptions of the metrics provided. For **content-based criteria**, we are interested in how to measure the quality of the content as it is presented to the user. This includes mainly criteria of the textual quality of the solution. For example, the criterion **fluency** is used for measuring the quality of the summaries in the SummEval dataset and hence is a content-based criteria. The **engagement-based criteria**, combine criteria of how the AI generated solution engages with the user. This includes for example the **empathy** criterion used to measure the quality of the generated stories in the Hanna dataset. The remaining two groups concentrate on more task specific evaluation criteria. **Integrity-based criteria** measure the coherence of task solution and whether it makes sense logically. For example the criterion **logical correctness**, used for measuring the quality of (mathematical) reasoning or coding task solutions in the Flask dataset, is a integrity-based criterion. Finally, **relevance-based criteria** measure the direct relevance of a task solution to the actual task. This includes for example the criterion **relevance**, used for measuring the connections of task solutions and initial task in several of the benchmark datasets e.g., TheNextChapter dataset. The separation of the quality criteria groups are not a 100% perfect and there are overlaps, for example content-based criteria like **readability**, can also be seen as engagement-based, since less legible solution are also less engaging. For a detailed list of the assigned criteria see technical report (Murugadoss et al. 2024).

Setting	1	2	3	4
GPT4-Turbo	-	0.414	0.468	**0.469**
GPT3.5-Turbo	-	0.269	0.310	**0.313**
Llama3 70b	0.295	0.299	0.349	**0.367**
Phi3-Medium	0.289	0.324	**0.367**	0.334
Llama3 8b	0.288	0.256	0.294	**0.352**
Mistral	**0.324**	0.261	0.259	0.311
Prometheus-2	**0.333**	-	-	0.266

Table 1: Pearson correlations of the scores generated by different LLMs-as-a-judge with the human annotations from the different datasets for the different settings (**1 - Perplexity / No prompt, 2 - Generic prompt, 3 - Specific prompt, 4 - Full rubric**). Bold numbers show highest agreement with human annotations under the setting for each model.

Model Selection for LLM-as-a-judge

To understand how model size and finetuning affect performance across the different quality criteria and settings of prompting, we test several current LLMs: GPT4-Turbo-0125 (OpenAI 2023) as large closed-model baseline; Llama3 70b (Meta 2024) as a medium size open-model; Llama3 8b, Mistral-v0.3 (Mistral.AI 2023) as small open-models, Phi3-Medium-128k (Microsoft 2024) as fine-tuned model for reasoning, and Prometheus-2 (Kim et al. 2024) as fine-tuned models for evaluation tasks.

Results

In this section, we present the main results of the evaluations using the different LLMs-as-a-judge under the different settings of prompting. Analogous to previous work (Liu et al. 2023b), to measure the quality of the evaluations we calculate the Pearson correlation of the generated scores by the LLMs-as-a-judge, respectively the perplexity values, and the human annotations given for each quality criteria from the benchmark datasets. We split the results section into model level, dataset level and criteria level results. In the model level results subsection, we present the results comparing the different LLMs under the settings of prompting, averaging over all criteria; the dataset level results subsection presents the results when we compare the different datasets under the settings of prompting, averaged over all criteria; the criteria level results subsection presents the results when we compare models and settings of prompting under the different groups of criteria. Finally, we present the results of a detailed analysis for each group of criteria from the introduced taxonomy.

Model level results

There is only small effect adding full rubric information. Providing the LLMs-as-a-judge with more detailed rubric information of the quality criteria, generally has only small influence on evaluation performance for the large and mid-size models, and might even be disadvantageous in certain

Setting	1	2	3	4
Flask	**0.448**	0.365	0.409	0.408
Hanna	0.237	0.232	0.262	**0.318**
TheNextChapter	0.275	0.193	0.273	**0.340**
Summeval	**0.408**	0.345	0.369	0.376
TopicalChat	0.189	0.421	**0.426**	**0.426**
InstruSum	0.160	0.130	**0.163**	0.153
OpinSummEval	0.179	0.316	**0.342**	0.328
Roscoe	0.159	0.294	0.349	**0.392**

Table 2: Pearson correlations of the scores generated by different LLMs-as-a-judge with human annotations from the **InstruSum** dataset for each setting.

situations (see Tab.1). For instance, Phi3's performance decreases when complete rubric details are provided compared to simple prompts which only mention the criterion name in the prompt. Here, Phi3's prior knowledge about evaluating the criteria has higher agreement with human annotators compared to when using full rubric information. Only the smaller Llama3 8b and Mistral models see improvements when given comprehensive rubric information for assessment. Among the open models, Llama3, both the 70b and 8b versions, perform best. Meanwhile, Mistral and Prometheus-2 do not show improvements when the LLM is prompted, with models' perplexity having higher correlation then the generated scores. For Prometheus-2, we only report perplexity and full rubric information in the prompts since this aligns with the fine-tuning data for this model and both setting 2 and 3 did return very poor results.

GPT4 performs best among all models. As may be expected, prompting GPT4-as-a-judge, even for a generic quality judgement, results in the highest performance in terms of agreement with human annotations compared to all other models tested. Further, GPT4's judgements do only improve marginally from prompting setting 3 to 4, indicating that GPT4's prior knowledge about evaluating does already agree with the human judgements to a high degree without the need to add more detailed rubric information about the evaluation.

Dataset level results

Perplexity correlates with text quality criteria. We observe (see Tab. 2) that the quality criteria from datasets with simple textual content creation tasks e.g., summarization in the SummEval dataset or story generation in the Hanna dataset, show high agreement with models' perplexity compared to simple prompting (setting 2 and 3). For more complex NLG tasks, which depend on several aspects and multiple possible steps, the human annotation correlate less strongly with perplexity compared prompting the LLMs-as-a-judge with more information. For example the opinion summary evaluations from the OpinSummEval dataset uses criteria which depend on sentiment identification and extractions of the key aspects, in these cases prompting the LLM seems necessary.

Full rubric information helps for non-default textual quality evaluations. Unusual textual quality evaluation tasks which measure the quality beyond simple textual

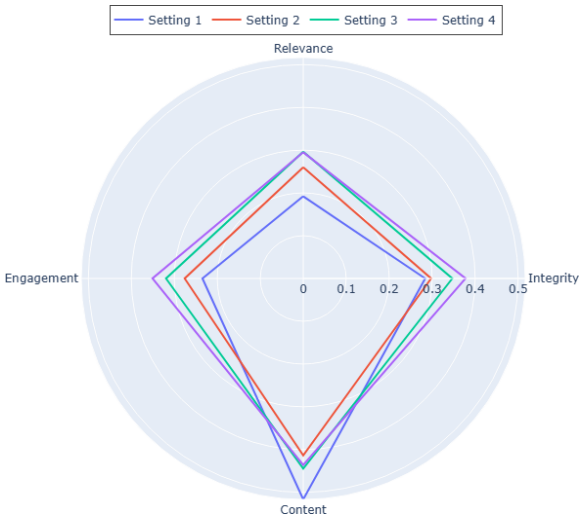

Figure 3: Radar chart of average Pearson correlations for the quality criteria groups for each of different settings (**1 - Perplexity / No Prompting, 2 - Generic prompt, 3 - Specific prompt, 4 - Full rubric**) over all models.

criteria like fluency, can benefit from more full rubric information about the the evaluation task. For example, we observe that for the TheNextChapter dataset, full rubric information in the prompts to the LLMs-as-a-judge leads to judgements with the highest correlations with human annotations. Compared to other datasets for textual quality evaluation, this datasets contains much more complex texts e.g., creative stories with non-default quality criteria like relatedness which is difficult to estimate without additional information by an LLM. Furthermore, evaluating more complex tasks which include more than text quality, like the logical reasoning tasks, benefit also from more detailed rubric information in the prompts. The logical and mathematical reasoning tasks in the Roscoe datasets for example do benefit from more information to effectively judge as shown by the higher correlations with the human judgements compared to prompting with less information or using perplexity.

Dataset level analysis can be misleading. Models' perplexity on both the Flask dataset and the SummEval dataset, outperforms simple prompting in aligning to human judgements. While the quality criteria in the SummEval dataset primarily focuses on textual quality where we expect perplexity to perform well for example, the Flask dataset consist a variety of different quality criteria which make it difficult to generalise and the average correlations values might be biased to the high values on the text related criteria. In the next subsection, we investigate this issue by using the previously introduced taxonomy to analyze results on a per-criteria group basis rather than average results per dataset.

Criteria level analysis

Content-based quality criteria correlate the most with perplexity. When evaluating quality with a focus on textual context, perplexity seems a viable alternative to prompting LLMs-as-a-judge. We observe that on average the agree-

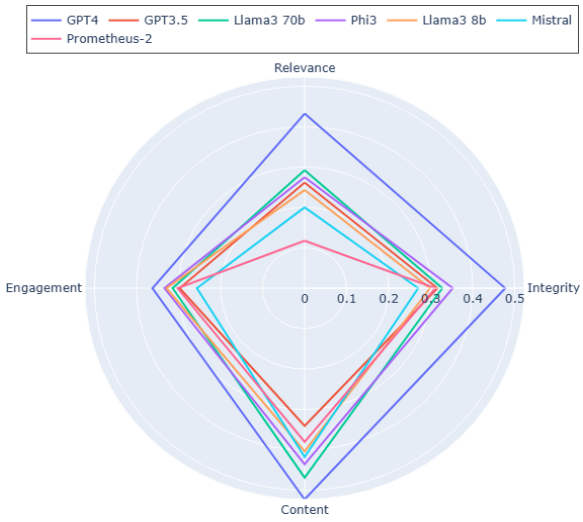

Figure 4: Radar chart of average Pearson correlations for the quality criteria groups for each of the different LLMs-as-a-judge over all setting.

ment with human annotations is more then 20% higher when using models' perplexity to judge the quality compared to prompting (Fig. 3). Further, there are only small differences between using a simple generic quality prompt for evaluation compared to all other settings of prompting, showing that models' prior knowledge generates judgements with high agreement with human judgements on textual quality.

Engagement-based quality criteria benefit the most from full rubric information. These criteria are unconventional as they assess the likelihood of a user feeling personally engaged, as opposed to merely evaluating straightforward text quality. Access to full rubric information can help judging with directives, particularly when the evaluation is more unusual and different from the text quality alone.

Often, there is only little improvement of adding full rubric information for most criteria groups. Except for the engagement-based criteria, there is only limited effect on adding full rubrics information to the prompts. Further, simple prompts for generic quality judgements results in similar correlation values with the human annotations than detailed information for content and relevance based evaluation criteria. This confirms again that advanced models' prior knowledge of text quality or relevance already has a high agreement with human judgements.

GPT4 clearly outperforms all other models. GPT-4 particularly outperforms in relevance and integrity quality criteria, surpassing other models on these criteria (Fig. 4). Although Phi3 matches GPT-4's performance in engagement based criteria, it generally performs less consistent with human annotations across other evaluative measures; Mitral's performance falls short in criteria associated with relevance and integrity; LLama3-70b exhibits a marginally improved performance over Phi3 when it comes to content and relevance-based criteria.

Setting	1	2	3	4
harmlessness	**1.000**	0.486	0.749	0.928
completeness	**0.764**	0.708	0.707	0.707
readability	**0.731**	0.256	0.329	0.341
fluency	**0.430**	0.310	0.314	0.312
metacognition	0.374	**0.454**	0.436	0.330
insightfulness	0.266	0.327	**0.444**	0.339
naturalness	0.274	0.454	0.458	**0.466**

Table 3: Pearson correlations of scores generated by different LLMs-as-a-judge with human annotations for content-based evaluation criteria, from the benchmark dataset for respective settings (**1 - Perplexity / No prompt, 2 - Generic prompt, 3 - Specific prompt, 4 - Full rubric**)

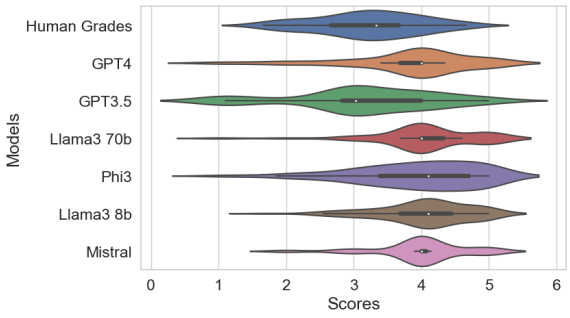

Figure 5: Violin plot of the generated scores by the LLMs-as-a-judge for the engagement-based criterion **empathy**, together with the corresponding human annotations.

Details on content-based criteria results

Drilling down the content-based evaluations criteria, we observe that perplexity outperforms prompting mainly on structural text quality criteria. For example, human annotations for **fluency** from the SummEval dataset have much higher agreement with perplexity compared to all prompting methods. This quality criterion judges grammar, spelling, and sentence structure for example. On the other hand, evaluating for more complex, specific and subjective content-based criteria like **naturalness**, which measures how natural the task response sounds, benefits from more instructions in the prompts for the LLMs-as-a-judge. Here, instructions to judge how much the task solution resembles a human answer improves agreement with human judgements.

Notably, we observe that model perplexity has 100% agreement with the human annotations for **harmlessness** on Flask datasets. This might reflect the strong influence fine-tuning has in inhibiting the generation of harmful content. Conversely, prompting for measuring harmfulness performs much lower until we provide full rubric information in the prompts.

Details on engagement-based criteria results

Engagement-based criteria are more challenging to judge since they are often subjective. We observe that there are fewer performance differences between the models compared to the results on the other criteria e.g., GPT4 judgements have an agreement (by Pearson correlation) of 0.32,

Setting	1	2	3	4
GPT4-Turbo	-	0.363	**0.365**	0.318
GPT3.5-Turbo	-	0.227	0.242	**0.274**
Llama3 70b	0.084	**0.301**	0.288	**0.301**
Phi3	0.011	0.260	**0.263**	0.249
Llama3 8b	0.007	0.180	**0.226**	0.180
Mistral	0.108	**0.154**	0.126	0.102

Table 4: Pearson correlations of the scores generated by different LLMs-as-a-judge with the human annotations for **groundedness** for the different setting.

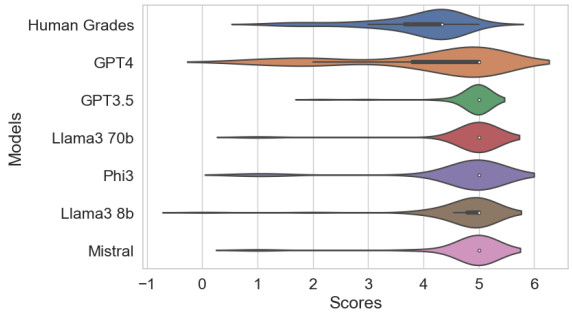

Figure 6: Violin plot of the generated scores by the LLMs-as-a-judge for integrity-based criterion **logical correctness**, together with human annotations.

Phi3 of 0.31 and Llama3 8b of 0.3. The overall performance of all models on these criteria is lower and the scores generated by the models have higher variance compared to other criteria. Further, human annotations for these criteria have on average lower scores with higher variance. For example, for the criterion **empathy** from the Hanna dataset (Fig. 5), human annotations show a significant degree of variation, and Phi3 notably generates scores with higher variability, unlike most other models that tend to cluster around a singular score. Hence, the distribution in engagement-based criteria may explain why Phi3 outperforms other models.

Details on relevance-based criteria results

For relevance-based quality criteria, we assume that the models need robust instructions about the problem to measure whether the information in the task solution is relevant, and estimate to what degree. Models' perplexity, but also generic prompts seem not sufficient for evaluation since relevance is more specific to certain aspects of the task solution. Still, we also observe that including a full rubric doesn't always appear necessary; instead the size of the models seem more important. For the criterion **groundedness** from the TopicalChat dataset for example (Tab. 4), we identify a clear trend of increasing agreement with the human judgements with larger models as LLM-as-a-judge.

Details on integrity-based criteria results

Similar to the relevance-based quality criteria, we assume that to measure the quality for integrity-based criteria the LLMs-as-a-judge need to have an understanding of the task, but also the ability to solve the task itself. We hypothesize that, for task-specific evaluations, the underlying LLM-as-a-judge actually needs to be able to solve the task itself to apply the correct score. As reported in (Lin et al. 2024), LLMs' ability to critic a task solution correlates with its ability to solve the task.

We exemplify this by the evaluations for the integrity-based evaluation criterion on the criterion **logical correctness** from the Flask dataset. This criterion reflects the correctness elements which are reflected in human annotations, which are more clearly scored as high (for logically correct) or low (for logically incorrect) scores. To select the appropriate scores, the LLM-as-a-judges need to know what is correct and what is wrong. Here, GPT-4 significantly outperforms all other models by a wide margin with a Pearson correlation of 0.68 with the human judgements in contrast to 0.34 for Llama3 70b and 0.33 for Phi3, for example.

To illustrate this, we plot the generated scores of the LLMs-as-a-judge (Fig. 6) and the human annotations. We observe that only GPT4 is able to generate lower scores to judge a task solution as "bad." All other models predominantly give high scores, consistently grading bad logically incorrect responses as "very good."

Conclusion

In this paper, we investigate how increasing levels of prompting impact the automatic evaluations made by LLMs-as-a-judge in measuring the quality of AI-generated text. We introduce a new taxonomy of quality criteria, summarizing commonly used criteria in automatic evaluations with LLMs into four broad categories: Content, Relevance, Integrity, and Engagement. We systematically evaluated several LLMs, including GPT-4, Llama-3, and others, across all settings of prompting to determine if more detailed instructions enhance the LLMs' alignment with human judgements. Key findings include:

- Detailed quality criteria information might not be necessary in the most powerful models; for instance, GPT-4 shows a high level of agreement with human judgements even without detailed instruction.
- Simple perplexity values are very effective at estimating textual quality, often outperforming the results of prompting the LLMs-as-a-judge with basic instructions.
- Judging task-specific quality criteria like relevance or logical correctness requires more capable, larger models, aligning with previous research on the necessary model capabilities for critiquing (Lin et al. 2024).

Future Work

In this work, we concentrate on single example evaluations only with simplistic prompts to minimize the effects of the reported biases in related works e.g., positional biases. Future work could extend this to pair-wise evaluation and more complex instructions. Further, adversarial prompts which contain contradictory instructions with the model's prior judgements can pose a serious challenge for using LLMs-as-a-judge. While out-of-scope of this work, future work will further investigate how different models follow adversarial instructions and which evaluation criteria are most receptive.

References

Andrade, H. G. 2005. Teaching With Rubrics: The Good, the Bad, and the Ugly. *College Teaching*, 53(1): 27–31.

Ankner, Z.; Blakeney, C.; Sreenivasan, K.; Marion, M.; Leavitt, M. L.; and Paul, M. 2024. Perplexed by Perplexity: Perplexity-Based Data Pruning With Small Reference Models. arXiv:2405.20541.

Bavaresco, A.; Bernardi, R.; Bertolazzi, L.; Elliott, D.; Fernández, R.; Gatt, A.; Ghaleb, E.; Giulianelli, M.; Hanna, M.; Koller, A.; Martins, A. F. T.; Mondorf, P.; Neplenbroek, V.; Pezzelle, S.; Plank, B.; Schlangen, D.; Suglia, A.; Surikuchi, A. K.; Takmaz, E.; and Testoni, A. 2024. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. arXiv:2406.18403.

Chen, Y.; Wang, R.; Jiang, H.; Shi, S.; and Xu, R. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. arXiv:2304.00723.

Chhun, C.; Colombo, P.; Suchanek, F. M.; and Clavel, C. 2022. Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 5794–5836. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Chiang, C.-H.; and Lee, H.-y. 2023. Can Large Language Models Be an Alternative to Human Evaluations? In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15607–15631. Toronto, Canada: Association for Computational Linguistics.

Doddapaneni, S.; Khan, M. S. U. R.; Verma, S.; and Khapra, M. M. 2024. Finding Blind Spots in Evaluator LLMs with Interpretable Checklists. arXiv:2406.13439.

Fabbri, A. R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; and Radev, D. 2021. SummEval: Re-evaluating Summarization Evaluation. arXiv:2007.12626.

Gao, M.; Hu, X.; Ruan, J.; Pu, X.; and Wan, X. 2024. LLM-based NLG Evaluation: Current Status and Challenges. arXiv:2402.01383.

Golovneva, O.; Chen, M.; Poff, S.; Corredor, M.; Zettlemoyer, L.; Fazel-Zarandi, M.; and Celikyilmaz, A. 2023. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. arXiv:2212.07919.

Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; and Hakkani-Tür, D. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, 1891–1895.

Ji, Y.; Gong, Y.; Peng, Y.; Ni, C.; Sun, P.; Pan, D.; Ma, B.; and Li, X. 2023. Exploring ChatGPT's Ability to Rank Content: A Preliminary Study on Consistency with Human Preferences. arXiv:2303.07610.

Kim, S.; Shin, J.; Cho, Y.; Jang, J.; Longpre, S.; Lee, H.; Yun, S.; Shin, S.; Kim, S.; Thorne, J.; et al. 2023. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. *arXiv preprint arXiv:2310.08491*.

Kim, S.; Suk, J.; Longpre, S.; Lin, B. Y.; Shin, J.; Welleck, S.; Neubig, G.; Lee, M.; Lee, K.; and Seo, M. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. arXiv:2405.01535.

Kocmi, T.; and Federmann, C. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. arXiv:2302.14520.

Li, J.; Sun, S.; Yuan, W.; Fan, R.-Z.; Zhao, H.; and Liu, P. 2023. Generative Judge for Evaluating Alignment. arXiv:2310.05470.

Li, Z.; Xu, X.; Shen, T.; Xu, C.; Gu, J.-C.; and Tao, C. 2024. Leveraging Large Language Models for NLG Evaluation: A Survey. arXiv:2401.07103.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Lin, Z.; Gou, Z.; Liang, T.; Luo, R.; Liu, H.; and Yang, Y. 2024. CriticBench: Benchmarking LLMs for Critique-Correct Reasoning. arXiv:2402.14809.

Liu, Y.; Fabbri, A. R.; Chen, J.; Zhao, Y.; Han, S.; Joty, S.; Liu, P.; Radev, D.; Wu, C.-S.; and Cohan, A. 2023a. Benchmarking Generation and Evaluation Capabilities of Large Language Models for Instruction Controllable Summarization. arXiv:2311.09184.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023b. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv:2303.16634.

Liu, Y.; Moosavi, N. S.; and Lin, C. 2024. LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores. arXiv:2311.09766.

Liu, Y.; Yang, T.; Huang, S.; Zhang, Z.; Huang, H.; Wei, F.; Deng, W.; Sun, F.; and Zhang, Q. 2023c. Calibrating LLM-Based Evaluator. arXiv:2309.13308.

Liu, Y.; Zhou, H.; Guo, Z.; Shareghi, E.; Vulić, I.; Korhonen, A.; and Collier, N. 2024. Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators. arXiv:2403.16950.

McAleese, N.; Pokorny, R. M.; Uribe, J. F. C.; Nitishinskaya, E.; Trebacz, M.; and Leike, J. 2024. LLM Critics Help Catch LLM Bugs. arXiv:2407.00215.

Meta. 2024. Llama 3. https://www.llama.com/.

Microsoft. 2024. Phi-3. https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms/.

Mistral.AI. 2023. Mistral 7B. https://mistral.ai/en.

Murugadoss, B.; Poelitz, C.; Drosos, I.; Le, V.; McKenna, N.; Negreanu, C. S.; Parnin, C.; and Sarkar, A. 2024. Evaluating the Evaluator: Measuring LLMs' Adherence to Task Evaluation Instructions. arXiv:2408.08781.

OpenAI. 2023. GPT-4 Turbo. https://platform.openai.com/docs/models/gpt-4-turbo.

Panickssery, A.; Bowman, S. R.; and Feng, S. 2024. LLM Evaluators Recognize and Favor Their Own Generations. arXiv:2404.13076.

Shen, Y.; and Wan, X. 2023. OpinSummEval: Revisiting Automated Evaluation for Opinion Summarization. arXiv:2310.18122.

Siledar, T.; Nath, S.; Muddu, S. S. R. R.; Rangaraju, R.; Nath, S.; Bhattacharyya, P.; Banerjee, S.; Patil, A.; Singh, S. S.; Chelliah, M.; and Garera, N. 2024. One Prompt To Rule Them All: LLMs for Opinion Summary Evaluation. arXiv:2402.11683.

Stureborg, R.; Alikaniotis, D.; and Suhara, Y. 2024. Large Language Models are Inconsistent and Biased Evaluators. arXiv:2405.01724.

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. 2023. Shepherd: A Critic for Language Model Generation. arXiv:2308.04592.

Xie, Z.; Cohn, T.; and Lau, J. H. 2023. The Next Chapter: A Study of Large Language Models in Storytelling. arXiv:2301.09790.

Ye, S.; Kim, D.; Kim, S.; Hwang, H.; Kim, S.; Jo, Y.; Thorne, J.; Kim, J.; and Seo, M. 2024. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. arXiv:2307.10928.

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. 2024. 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: Curran Associates Inc.