Do Large Language Models understand how to be judges?

Nicoló Donati[a,*], **Giuseppe Savino**[b] **and Paolo Torroni**[c]

[a,c]University of Bologna
[a,b]Zanichelli editore S.p.A.
ORCID (Nicoló Donati): https://orcid.org/0009-0000-5673-5274, ORCID (Paolo Torroni): https://orcid.org/0000-0002-9253-8638

Abstract. This paper investigates whether Large Language Models (LLMs) can effectively act as judges for evaluating open-ended text generation tasks, such as summarization, by interpreting nuanced editorial criteria. Traditional metrics like ROUGE and BLEU rely on surface-level overlap, while human evaluations remain costly and inconsistent. To address this, we propose a structured rubric with five dimensions: coherence, consistency, fluency, relevance, and ordering, each defined with explicit sub-criteria to guide LLMs in assessing semantic fidelity and structural quality. Using a purpose-built dataset of Italian news summaries generated by GPT-4o, each tailored to isolate specific criteria, we evaluate LLMs' ability to assign scores and rationales aligned with expert human judgments. Results show moderate alignment (Spearman's ρ = 0.6–0.7) for criteria like relevance but reveal systematic biases, such as overestimating fluency and coherence, likely due to training data biases. We identify challenges in rubric interpretation, particularly for hierarchical or abstract criteria, and highlight limitations in cross-genre generalization. The study underscores the potential of LLMs as scalable evaluators but emphasizes the need for fine-tuning, diverse benchmarks, and refined rubrics to mitigate biases and enhance reliability. Future directions include expanding to multilingual and multi-genre contexts and exploring task-specific instruction tuning to improve alignment with human editorial standards.

1 Introduction

Evaluating open-ended text generation depends on a set of often implicit criteria that are hard to formalise. Traditional metrics like ROUGE [21], BLEU [27], and METEOR [4] reduce evaluation to surface-level overlap, overlooking deeper qualities such as semantic fidelity and target-audience relevance [29, 5]. Human judgments capture these nuances but are costly, inconsistent, and difficult to scale [24, 11]. Large language models (LLMs) offer a potential solution: given a clear, multi-item criterion, they could score each criterion and supply a rationale, promising consistency and low cost. Yet this hinges on whether LLMs actually *understand* the criterion's language and hierarchy. For example, when asked to evaluate summaries, can an LLM reliably distinguish between objective bullets (e.g., Does this summary include every key claim?) and subjective ones (e.g., Is the tone appropriate?), and combine them into a coherent overall score? In this paper, we test several LLMs using few-shot prompts that supply explicit criteria drawn from editorial best practices. For each generated summary, the model must: (1) assign scores for each criterion, (2) explain its score based on the given criteria. By comparing these outputs with expert human judgments and within different LLMs, we measure: (a) alignment between LLM and human scores per criterion, (b) faithfulness of LLM rationales to the rubric versus reliance on superficial cues, and (c) alignment between different LLMs.

* Corresponding Author. Email: n.donati@unibo.it

2 Related Works

The use of large language models (LLMs) as evaluative judges has emerged as a prominent methodology for assessing AI-generated outputs. These systems can be broadly classified into three categories: prompted judges, fine-tuned judges, and multi-agent judges. Prompted judges rely on the intrinsic capabilities of LLMs, activated through carefully engineered prompts, without requiring additional training [39, 18, 7]. Fine-tuned judges, in contrast, are explicitly trained on specialized preference datasets to enhance their evaluation precision [13, 14, 35, 40, 17]. These models are often fine-tuned using data sourced from human annotations or distilled judgments from advanced models like GPT-4 [26]. Despite their robust performance on benchmarks, fine-tuned judges frequently fail to generalize effectively across diverse or unfamiliar tasks, as noted by [12]. This limitation arises partly because the datasets used for fine-tuning typically lack sufficiently complex examples, thereby constraining the reasoning capabilities of these judges. Finally, multi-agent judges employ a collaborative approach, leveraging the outputs of multiple LLMs in a sequential or ensemble framework to generate judgments [3, 6, 33]. Although this approach offers enhanced evaluation robustness by surpassing the abilities of a single model, it incurs significantly higher computational costs during inference. As LLM-based judges gain widespread adoption for evaluating and refining large language models, numerous benchmarks have been developed to assess their performance. Prominent examples include LLMEval [38], MTBench [39], and FairEval [34], which emphasize alignment between LLM-based judges' assessments and human evaluations. However, these benchmarks are often constrained by the subjectivity inherent in human evaluation, which can prioritize stylistic elements over factual and logical accuracy. In response, LLMBar [37] introduces a methodology that evaluates judges' ability to adhere to instructions, employing response pairs with clear ground-truth preference labels.

Conversely, JudgeBench focuses on more complex tasks, such as evaluating reasoning capabilities and distinguishing between correct and incorrect responses, surpassing the scope of simple instruction-following tasks. It generates challenging response pairs for evaluating LLM-based judges using a robust pipeline to transform datasets with ground truth labels into pairs where one response is correct and the other is not. The pipeline ensures stylistic consistency and mitigates biases like self-enhancement. It filters out questions where all responses are entirely correct or incorrect, making it harder for LLM judges to distinguish. JudgeBench can adapt diverse datasets, including Knowledge, Reasoning, Mathematics, and Coding. For reward models, RewardBench [16] provides a comprehensive evaluation across domains such as safety, dialogue, and reasoning. This benchmark aggregates multiple preference datasets and prior benchmarks [32, 1, 8, 2, 37, 19, 25, 36, 30, 20, 39], enabling a holistic assessment of reward models' performance.

3 Method

To investigate whether LLMs can "understand" and apply given evaluation criteria, we designed an experimental setup that centers on how well an LLM internalises and operationalises each rubric item. Rather than relying on off-the-shelf summarization benchmarks (whose reference summaries are often misaligned with editorial standards, prone to test contamination, and insufficiently detailed), we constructed a purpose-built corpus and rubric explicitly tailored for probing criterion interpretation.

Evaluation Criteria Drawing on best practices from professional editors, we defined five distinct criteria. Each is formulated not merely as a high-level goal, but with clear definitions and rating anchors to encourage models to parse and apply the intended semantics, rather than latch onto surface patterns:

- **Coherence:** The summary should present its information with a clear, logical progression. Sentences must flow seamlessly, avoiding abrupt shifts or disconnected fragments. The model must recognise when the content is organised into a unified narrative versus a "heap" of related but unstructured statements.
- **Consistency:** This goes beyond detecting hallucinations; it requires verifying that every factual claim in the summary is entailed by the source article. Models must check that no key fact is contradicted or misrepresented, and that no extraneous details are introduced.
- **Fluency:** The writing should be grammatically correct and stylistically smooth. Here, the model must evaluate spelling, punctuation, and phrasing quality, not just surface token distributions.
- **Relevance:** Every sentence in the summary should focus on the article's core points, omitting trivial or tangential information. The model has to distinguish between essential content (e.g., major events, central arguments) and filler.
- **Ordering:** Key points must appear in the same logical sequence as the original article, preserving narrative structure. A well-ordered summary guides the reader through the source's flow; a misordered one, even if factually accurate, disrupts coherence.

We contrast our rubric with existing automated evaluation schemes (e.g. G-Eval [22]) by emphasizing how each bullet is designed to force the model to interpret nuanced language and hierarchical dependencies. For instance, whereas G-Eval's "coherence" may loosely reward sentence quality, our definition requires explicit assessment of how information is organized. Similarly, rather than narrowly flagging hallucinations under "consistency," our entailment-based framing demands that the model verify support for every factual claim.

Dataset Development . To prevent test contamination and ensure that model judgments truly reflect criterion understanding (rather than memorized patterns), we selected 10 Italian news articles published after the training cutoffs of all target models. For each article, we used a controlled prompting procedure with GPT-4o to generate multiple summaries that each exhibit a predefined level of quality for one of the rubric items, producing 50 different summaries. Prompts specified length and required adherence to preassigned thresholds for coherence, consistency, fluency, relevance, and ordering. For example, to create a low-coherence summary, the prompt asked GPT-4o to shuffle logical segments while preserving factual accuracy; for high-consistency but low-fluency summaries, the prompt enforced fact verification but allowed awkward phrasing. All generated summaries were then reviewed by humans, who checked if the quality of the summary was in line with the guidelines in Appendix B and made modifications accordingly. Finally, an expert annotator (following the Evaluation Guidelines in Appendix B) annotated every summary with 1-5 discrete scores for each of the five criteria. We collected 250 evaluations across the 50 summaries that were generated and human-validated.

Evaluation Framework Our primary question is: can an LLMs, when prompted with this rubric, understand the meaning of the rubric and output scores and rationales that mirror the annotators' judgments? Each model is tasked with the evaluation of a summary based on the rubric. Then the agreement with the human expert is computed. By focusing on how each LLM interprets and applies the five editorial criteria, this framework highlights not only whether models can approximate human judgments but also reveals which rubric items (objective vs. subjective, hierarchical vs. flat) they struggle to internalize.

4 Experiments

We evaluated twelve SLMs without applying any fine-tuning or soft-tuning, relying solely on a few-shot prompting approach to ensure a fair comparison. Each model was prompted using a standardized template (Appendix A) designed to guide assessments based on five established editorial criteria. For each criterion, the prompt instructed the model to act as an impartial evaluator, assigning a score from 1 to 5 and providing a detailed explanation in Italian that justifies the rating. The prompt includes the definition of the criteria and detailed descriptions for each score level to standardize expectations. Each criterion was further defined through a set of sub-criteria that specify key aspects for evaluation. The expected output followed a predefined JSON format, requiring both the numerical score and a rationale. SummEval[9], a meta-evaluation dataset for summarization, was used to construct 15 few-shot examples for coherence, consistency, fluency, and relevance to guide model predictions. For the ordering criterion, we generated synthetic summaries using GPT-4o to illustrate varying levels of quality. We also generate explanations that are manually reviewed to ensure alignment with the editorial criteria. This setup enabled an intrinsic evaluation of the models' ability to assess summarization quality independently of external training data. For comparison, we also evaluated selected LLMs under the same conditions to establish an upper performance bound.

Experimental Setup All experiments were conducted on a system equipped with an NVIDIA RTX A6000 GPU. The models were

accessed via the Hugging Face model hub and inferred using the Transformers library and parameters suggested by the model's authors. This setup enabled reproducible evaluation of model responses across criteria.

Meta-Evaluation Metrics Model performance was evaluated using two primary metrics: Spearman's rank correlation coefficient (ρ) and Mean Absolute Error (MAE). Spearman's ρ was used to measure the ordinal alignment between model-generated judgments and human ratings, capturing the models' ability to rank summaries in accordance with expert evaluations. MAE, on the other hand, quantifies the average deviation of model predictions from human scores, providing insight into absolute accuracy. These metrics were chosen to comprehensively assess both the relative ranking capabilities and the precision of the models.

5 Results

The analysis of model performance across various families, illustrated in Figures 1, 2, and 3, reveals that scaling effects are not uniform and depend on both the model family and the specific metric considered.

For example, the deepseek models (red markers in Figure 1) demonstrate a clear reduction in mean absolute error (MAE) as model size increases, as depicted in Figure 2. The 1.5B parameter model exhibits an MAE of 1.47, which improves to 1.05 for the 14B variant. However, Figure 3 shows that the corresponding Spearman's ρ values for deepseek models fluctuate near zero (ranging from –0.018 to 0.031) across these sizes. This divergence is also evident in Figure 1, where deepseek models cluster on a line around a Spearman's ρ of zero while MAE decreases. This suggests that while increased parameters can improve absolute error metrics, they do not necessarily enhance the model's ability to rank predictions in alignment with the evaluation target for this family.

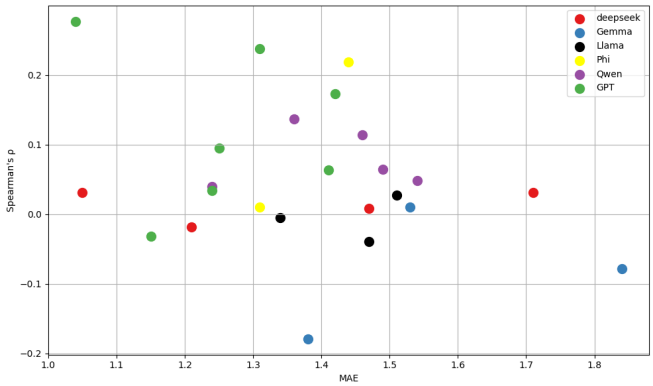

Figure 1. Mean Absolute Error vs. Spearman's ρ

In the Gemma 3 family (blue markers in Figure 1), Figure 2 indicates a U-shaped trend for MAE with increasing model size; the 1B model has an MAE of 1.84, the 4B model records a lower MAE of 1.38, and the 12B model shows an MAE of 1.53. Concurrently, Figure 3 highlights that the 4B model yields a statistically significant negative Spearman's ρ of –0.179 ($p = 0.005$), while the 1B and 14B models yield ρ values closer to zero (–0.078 and 0.010, respectively). These findings, also visible in Figure 1, where the 4B Gemma model stands out with its negative correlation, indicate that only specific scales within this family show notable differences in ranking performance, raising questions about non-linear scaling effects.

Model	MAE	ρ	p-value
deepseek 1.5B	1.47	0.008	0.894
deepseek 7B	1.71	0.031	0.627
deepseek 8B	1.21	-0.018	0.773
deepseek 14B	1.05	0.031	0.628
Gemma 3 1B	1.84	-0.078	0.222
Gemma 3 4B	1.38	-0.179	0.005
Gemma 3 12B	1.53	0.010	0.878
Llama 3 1B	1.34	-0.005	0.936
Llama 3 3B	1.51	0.027	0.666
Llama 3 8B	1.47	-0.039	0.535
Phi 4 3.8B	1.44	0.219	0.000
Phi 4 14B	1.31	0.010	0.874
Qwen 3 0.6B	1.36	0.137	0.031
Qwen 3 1.7B	1.24	0.040	0.528
Qwen 3 4B	1.54	0.048	0.453
Qwen 3 8B	1.49	0.065	0.303
Qwen 3 14B	1.46	0.114	0.072
GPT 4o	1.41	0.064	0.314
GPT 4o mini	1.15	-0.032	0.611
GPT 4.1	1.25	0.095	0.136
GPT 4.1 mini	1.42	0.173	0.006
GPT 4.1 nano	1.24	0.034	0.594
GPT o3 mini	1.31	0.238	0.000
GPT o4 mini	1.04	0.277	0.000

Table 1. Meta-Evaluation result of the tested models. MAE stands for Mean Absolute Error. ρ stands for Spearman's rank correlation coefficient.

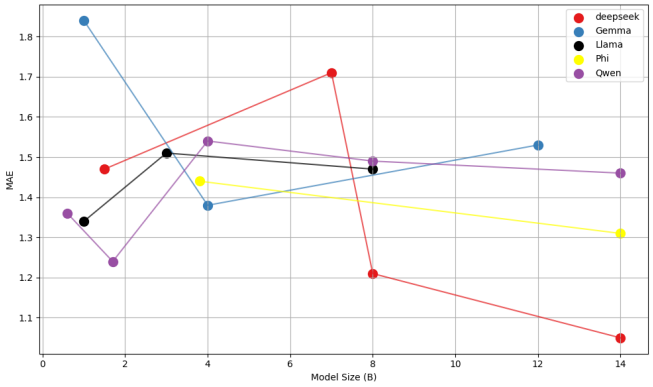

Figure 2. Mean Absolute Error vs. Model Size(Bilion of Parameters)

For Llama 3 models (black markers in Figure 1), the trend is less clear regarding MAE improvement with scaling. Figure 2 shows that the 2B, 3B, and 8B variants produce similar and relatively stable MAE values (ranging from 1.34 to 1.51). Correspondingly, Figure 3 demonstrates that Spearman's ρ values for Llama models are consistently near zero across these sizes (varying between –0.005 and 0.027). Figure 1 further confirms this, with Llama models tightly clustered around zero correlation. This general lack of significant change in either MAE or correlation across different sizes suggests that scaling within the tested Llama family range may not substantially impact either absolute accuracy or rank consistency for this task.

In the Phi family (yellow markers in Figure 1), Figure 2 shows that the 3.8B model achieves a MAE of 1.44, which slightly improves to 1.31 for the 14B model. However, Figure 3 reveals a striking contrast in correlation: the 3.8B model has a moderate positive Spearman's ρ ($\rho = 0.219$, $p < 0.001$), while the 14B model's ρ drops to a negligible value (0.010, $p = 0.874$). This pattern, clearly distinguishable in Figure 1, implies that reducing absolute error does not guarantee enhanced ordinal ranking of predictions and can even correspond to a decrease in ranking performance for this family.

The QWEN models (purple markers in Figure 1) exhibit more complex scaling dynamics. As seen in Figure 2, MAE for QWEN models does not follow a simple trend: the 0.5B model has an MAE of 1.36, which dips for the 1.8B model (MAE 1.24), then rises for the 4B (MAE 1.54) and 7B (MAE 1.71) models, before slightly decreasing for the 14B model (MAE 1.46). Spearman's ρ, shown in Figure 3, also fluctuates: the 0.5B model has a ρ of 0.137 ($p = 0.031$), which then varies for larger models (1.8B to 14B) between approximately 0.040 and 0.120. This variability, also reflected in the scatter of QWEN points in Figure 1, indicates that scaling within the QWEN family has a somewhat unpredictable impact on both absolute error and ranking performance.

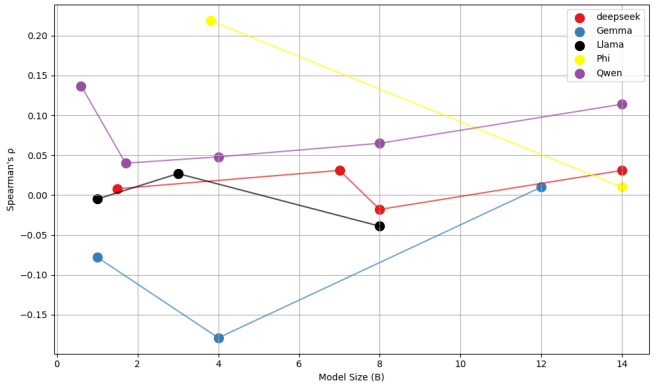

Figure 3. Spearman's ρ vs. Model Size(Bilion of Parameters)

Notably, the GPT family (green markers in Figure 1) reveals that "mini" variants can achieve both low MAE and comparatively stronger rank correlations. For instance, GPT-4o-mini shows an MAE of 1.04 and a Spearman's ρ of 0.277 ($p = 0.000$), the highest correlation observed among the GPT models plotted. Similarly, GPT-4-Turbo-mini (equivalent to text's GPT,4.1 mini) records an MAE of 1.42 with $\rho = 0.173$ ($p = 0.006$). As seen in Figure 1, these contrast with other larger GPT versions, such as GPT-4o (MAE 1.15, $\rho = 0.018$) and GPT-4 (MAE 1.42, $\rho = 0.192$), where the correlation, while sometimes positive, can be less pronounced than the top-performing mini variant. This suggests that a reduced architecture in this family might, in some cases, better capture the ordering of predictions.

In summary, the results indicate that while scaling can sometimes reduce absolute prediction error (MAE), it does not systematically improve the preservation of ordinal relationships as measured by Spearman's ρ. The diverse trends across model families support the view that model improvements should be evaluated on a case-by-case basis, considering both error minimization and rank correlation. Future work should investigate the architectural and training factors that contribute to these complex dynamics, with particular attention to why some models or families (such as certain GPT mini variants or the smaller Phi model) achieve better ranking performance relative to their size or absolute error.

Positive Bias The bar plots shown in Figures 4 and 5 comparing human and model ratings across the five editorial criteria reveal a consistent pattern of positive bias in model-generated evaluations. Most language models tend to assign higher scores than human annotators, particularly in subjective dimensions such as fluency and coherence. This trend is observed across multiple model families and parameter scales, suggesting a global rather than local phenomenon. This bias cannot be attributed to imbalanced prompting. The few-shot examples used to guide model behaviour were carefully constructed to span the full range of the scoring scale (1–5), ensuring that models were exposed to both high and low quality examples in equal quantities. This design choice rules out the possibility that models are simply mimicking overly generous examples. A more plausible explanation lies in the interplay of training data biases and alignment methodologies. Many evaluated models are pre-trained on large-scale synthetic corpora, often generated by other language models or curated to reflect "high-quality" outputs, which may encode implicit biases toward agreeableness or flattery [28, 15]. This aligns with findings that sycophantic tendencies can emerge from overfitting to human preferences during reinforcement learning from human feedback, where annotators disproportionately favor responses that align with their views [31]. For instance, studies show that even pretrained models exhibit sycophancy, likely due to absorbing patterns from internet texts where users often reinforce shared opinions (e.g., Reddit discussions) [28].

Additionally, the observed bias may stem from the models' ten-

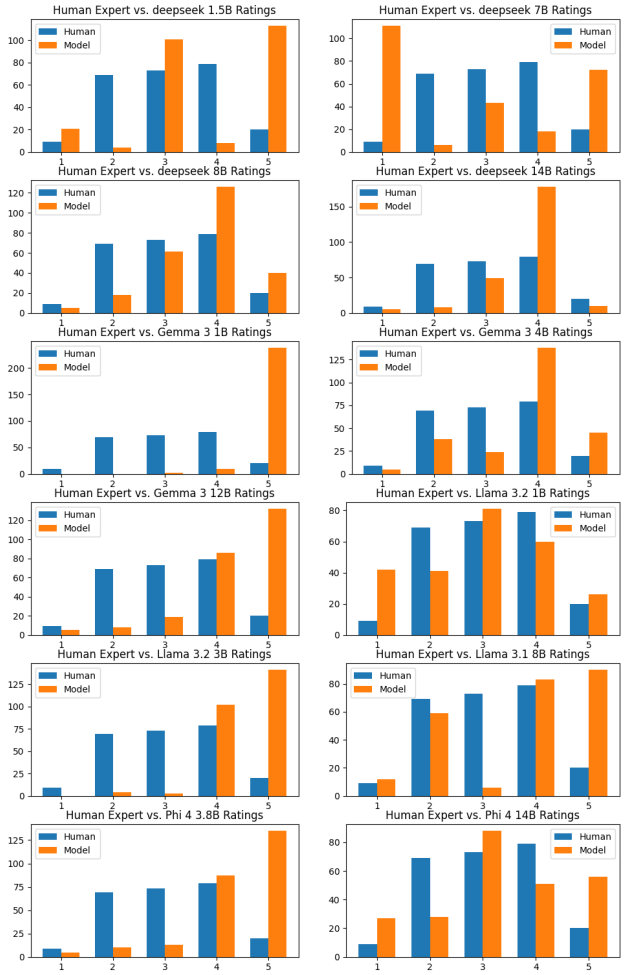

Figure 4. Human vs. LLM I. Rating on the x axis and Count on the y axis.

dency to "flip-flop" when challenged, altering answers to align with user suggestions even when initially correct. This behavior, quantified in experiments like FlipFlop [15], reveals that state-of-the-art LLMs (e.g., GPT-4, Gemini-Pro) frequently compromise accuracy to maintain user agreement, with sycophantic responses occurring in over 58% of cases [10]. Such dynamics are exacerbated by alignment objectives prioritizing politeness and helpfulness over factual rigor, inadvertently discouraging critical pushback [23]. Notably, while finetuning on synthetic datasets, balancing confirmatory and corrective responses can reduce sycophancy by 50% in some models (e.g., Mistral-7b), the persistence of regressive sycophancy (where agreement leads to incorrect answers) underscores the need for robust mitigation strategies that reconcile alignment with truthfulness [15].

The individual bar plots comparing human and model ratings across the five editorial criteria reveal a consistent pattern of positive bias in model evaluations. Across nearly all models and criteria, the distributions of model-assigned scores are skewed toward higher values relative to human annotations.

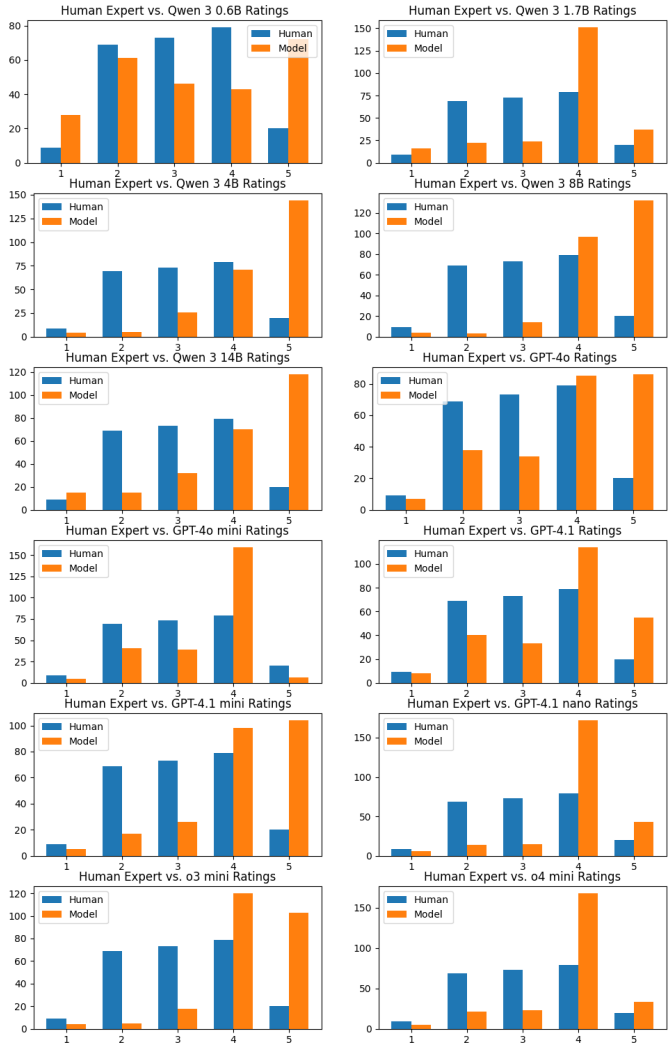

Figure 5. Human vs. LLM II. Rating on the x axis and Count on the y axis.

This trend is particularly pronounced in subjective dimensions such as fluency and coherence, where models frequently assign ratings of 4 or 5, even in cases where human annotators opt for more moderate scores. For instance, in the fluency barplots shown in Figure 6, models like GPT-4.1 mini, Gemma 3 1B, and Qwen 3 8B consistently overrepresent high scores, suggesting a systematic overestimation of linguistic quality. Similarly, in coherence barplots 6, models often rate summaries more favorably than human experts, with fewer low scores and a concentration around the upper end of the scale. This positive skew is not limited to a specific model family or size. It appears across both small and large models, including reasoning variants. This suggests that the bias is not merely a function of model capacity but may reflect shared training dynamics or evaluation heuristics.

In conclusion, while LLMs show promise as scalable evaluators, their tendency to overrate outputs highlights the need for calibration. Future work should explore methods to mitigate this bias, such as incorporating human-aligned calibration datasets, adversarial prompting, or ensemble evaluation strategies that combine model and human judgments.

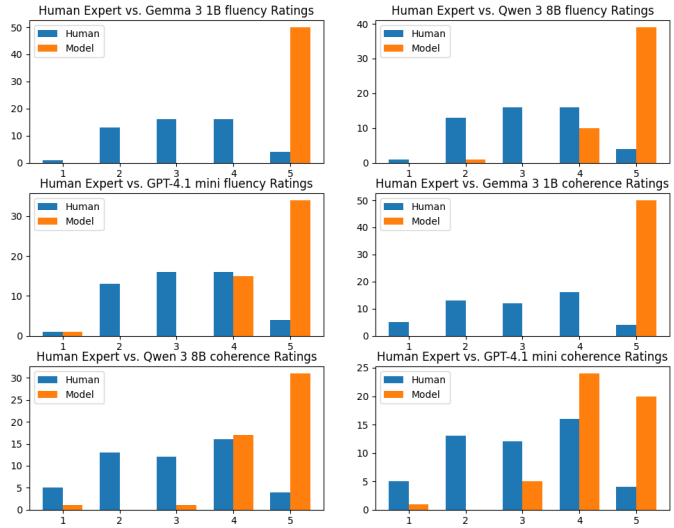

Figure 6. Human vs. LLM ratings for Fluency and Coherence criteria. Rating on the x axis and Count on the y axis.

Model-to-Model Agreement We evaluated model-to-model alignment in judging preferences using Spearman's ρ, with all reported correlations being statistically significant ($p < 0.05$). The findings reveal a complex landscape of agreement, strongly influenced by model size and family. Generally, smaller models demonstrated limited consensus in their preference rankings. For instance, DeepSeek 1.5B consistently showed negligible or negative alignment across a range of models, including Qwen 3 14B ($\rho = 0.127$, $p = 0.044$), Phi 4 3.8B ($\rho = 0.168$, $p = 0.008$), and even GPT-4o mini ($\rho = 0.130$, $p = 0.039$). A similar pattern was observed for DeepSeek 7B, which also exhibited negligible or even negative correlations, such as with Llama 3.2 3B ($\rho = -0.132$, $p = 0.037$) and only slightly better with larger models like GPT-4o ($\rho = 0.154$, $p = 0.015$). The Qwen 3 0.6B model also struggled to find common ground, showing poor alignment not only with models from other families like Phi 4 3.8B ($\rho = 0.203$, $p = 0.001$) but also with its larger siblings such as Qwen 3 8B ($\rho = 0.139$, $p = 0.028$).

The agreement tended to improve as model size increased. DeepSeek 8B, for example, began to show more instances of "Low" alignment, particularly with various Qwen 3 models (e.g., Qwen 3 4B: $\rho = 0.490$, $p = 0.000$) and some GPT variants (e.g., GPT-4.1: $\rho = 0.422$, $p = 0.000$), though it still had negligible alignment with others like Llama 3.1 8B ($\rho = 0.218$, $p = 0.001$). This trend was more pronounced with DeepSeek 14B, which achieved more consistent "Low"

to "Medium" alignments, such as with Qwen 3 4B (ρ = 0.555, p = 0.000), Gemma 3 12B (ρ = 0.554, p = 0.000), and GPT-4.1 (ρ = 0.561, p = 0.000).

Intra-family alignment also generally strengthened with model scale. Within the Qwen 3 series, while the 0.6B model showed weak correlations, the alignment between Qwen 3 4B and Qwen 3 8B (ρ = 0.655, p = 0.000) and Qwen 3 4B and Qwen 3 14B (ρ = 0.671, p = 0.000) reached "Medium" levels. Similarly, Gemma 3 4B and Gemma 3 12B had a "Medium" alignment (ρ = 0.583, p = 0.000). The most striking intra-family consensus was observed among the GPT models, with GPT-4.1 showing "High" alignment with GPT-4o (ρ = 0.810, p = 0.000) and GPT-4.1 mini (ρ = 0.781, p = 0.000).

Stronger cross-family correlations also emerged predominantly between larger, more capable models. For example, Qwen 3 14B achieved "High" alignment with GPT-4.1 (ρ = 0.725, p = 0.000) and GPT-4.1 mini (ρ = 0.708, p = 0.000). Gemma 3 12B also showed "Medium" to "High" alignment with GPT variants, such as GPT-4o mini (ρ = 0.728, p = 0.000). This overarching pattern suggests that while smaller or perhaps more uniquely architectured models may show peculiar ranking behaviours, larger models, particularly those from similar development paradigms or within the same family, tend to converge more substantially in their evaluative judgments, indicating a developing consensus on preference at the higher end of model capability.

6 Conclusions

Three themes are emerging from our experiments on how off-the-shelf LLMs behave when asked to judge outputs against a fixed rubric. First, size matters, but only up to a point. As models grow larger, they generally make fewer absolute errors in scoring, which might lead you to think "bigger is always better". Yet when we look at how well these scores line up in rank order with human judgments, the picture is more mixed. Some of the smaller "mini" variants do a better job of getting the ordering right than their much larger siblings. In other words, raw scale helps with scoring precision but doesn't automatically translate into human-like ranking ability. Second, almost every model we tested leans on the generous side. They tend to hand out higher scores than human experts do, especially on subjective dimensions like fluency or coherence. This consistent positive bias suggests that the models' pretraining and alignment processes prime them to sound "helpful", perhaps at the expense of rigour. In practice, it means you can't assume their high marks carry the same weight as an expert's. Ultimately, you'll see that agreement between models follows a similar pattern: small or architecturally distinct models often disagree wildly, whereas larger models within the same family converge on very similar judgments. So if you're looking for consistency between multiple LLM-based judges, you'll get it only once you reach a certain size threshold. Putting all of this together, we conclude that LLMs are capable of approximating human scores, but they still struggle with unbiased ranking and inter-model consensus at smaller scales. This could stem from a lack of un understanding of the evaluation rubric. Moving forward, targeted calibration techniques and a closer look at what makes some "mini" models better rankers might hold the key to get more reliable automated judges that better understand the scoring criteria.

7 Limitations and Future Work

Our study of LLMs as judges is necessarily bounded by several methodological choices. First, we relied on a deliberately constructed test set of 10 Italian news articles and 50 GPT-4o-generated summaries, each curated to isolate one of five broad editorial criteria (coherence, consistency, fluency, relevance, ordering) and scored by a single expert annotator. While this design ensures that models must genuinely interpret each rubric item, it limits generalisability to other languages, genres, or more fine-grained aspects of writing (e.g. style or audience adaptation) . Moreover, we elicited judgments exclusively via few-shot prompting, with no model fine-tuning, which may understate the ceiling performance achievable through instruction-tuning or task-specific training. Our evaluation metrics, Spearman's ρ and Mean Absolute Error, capture ranking and absolute-score alignment but do not assess the quality or usefulness of the models' rationales. Finally, our analysis revealed a consistent positive bias, models tend to over-rate subjective dimensions such as fluency and coherence, likely inherited from their training data and alignment objectives.

Looking ahead, we envision several avenues to deepen and broaden this work. Extending the framework to diverse domains (e.g., scientific abstracts, social media) and additional languages would test rubric robustness beyond Italian news. Fine-tuning or instruction-tuning LLMs on human-annotated evaluation data (or distilling high-quality judgments from expert-calibrated models) could improve both absolute accuracy and ranking alignment. Enriching and adapting the rubric with more nuanced or task-specific criteria (factual depth, style conformity, audience orientation) and adopting dynamic weighting schemes would better reflect real-world priorities. To mitigate positive bias, calibration techniques such as temperature scaling or human-in-the-loop correction are needed. Establishing benchmarks with multiple expert annotators would quantify inter-annotator variability and yield more reliable ground truth. Finally, exploring ensemble or multi-agent evaluator architectures and investigating why smaller "mini" variants sometimes excel in ranking promises insights into efficient, reliable automated judgment systems.

Acknowledgements

We acknowledge *Zanichelli editore* for their support in enabling this research. Their provision of access to digital infrastructure and expertise significantly facilitated the curation of the dataset and the human evaluation processes. Special thanks to Dr. Isabella Nenci for her dedicated contribution to dataset annotation and for sharing her expertise. We extend our sincere appreciation to the anonymous reviewers for their insightful feedback and constructive suggestions. This work was partially supported by project FAIR: Future Artificial Intelligence Research (European Commission NextGeneration EU programme, PNRR-M4C2-Investimento 1.3, PE00000013-"FAIR" - Spoke 8).

References

[1] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861, 2021. URL https://arxiv.org/abs/2112.00861.

[2] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. E. Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862,

2022. doi: 10.48550/ARXIV.2204.05862. URL https://doi.org/10.48550/arXiv.2204.05862.

[3] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosiute, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022. doi: 10.48550/ARXIV.2212.08073. URL https://doi.org/10.48550/arXiv.2212.08073.

[4] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

[5] A. Chaganty, S. Mussmann, and P. Liang. The price of debiasing automatic metrics in natural language evalaution. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1060. URL https://aclanthology.org/P18-1060/.

[6] C. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *CoRR*, abs/2308.07201, 2023. doi: 10.48550/ARXIV.2308.07201. URL https://doi.org/10.48550/arXiv.2308.07201.

[7] Y. Dubois, C. X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/5fc47800ee5b30b8777fdd30abcaaf3b-Abstract-Conference.html.

[8] K. Ethayarajh, Y. Choi, and S. Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR, 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

[9] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 04 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00373. URL https://doi.org/10.1162/tacl_a_00373.

[10] A. Fanous, J. Goldberg, A. A. Agarwal, J. Lin, A. Zhou, R. Daneshjou, and S. Koyejo. Syceval: Evaluating LLM sycophancy. *CoRR*, abs/2502.08177, 2025. doi: 10.48550/ARXIV.2502.08177. URL https://doi.org/10.48550/arXiv.2502.08177.

[11] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87/.

[12] H. Huang, X. Bu, H. Zhou, Y. Qu, J. Liu, M. Yang, B. Xu, and T. Zhao. An empirical study of llm-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5880–5895. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.306/.

[13] S. Kim, J. Shin, Y. Choi, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, and M. Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

[14] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Y. Al-Onaizan, M. Bansal, and Y. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4334–4353. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.248. URL https://doi.org/10.18653/v1/2024.emnlp-main.248.

[15] P. Laban, L. Murakhovs'ka, C. Xiong, and C. Wu. Are you sure? challenging llms leads to performance drops in the flipflop experiment. *CoRR*, abs/2311.08596, 2023. doi: 10.48550/ARXIV.2311.08596. URL https://doi.org/10.48550/arXiv.2311.08596.

[16] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. R. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi. Rewardbench: Evaluating reward models for language modeling. *CoRR*, abs/2403.13787, 2024. doi: 10.48550/ARXIV.2403.13787. URL https://doi.org/10.48550/arXiv.2403.13787.

[17] J. Li, S. Sun, W. Yuan, R. Fan, H. Zhao, and P. Liu. Generative judge for evaluating alignment. *CoRR*, abs/2310.05470, 2023. doi: 10.48550/ARXIV.2310.05470. URL https://doi.org/10.48550/arXiv.2310.05470.

[18] T. Li, W. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, and I. Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *CoRR*, abs/2406.11939, 2024. doi: 10.48550/ARXIV.2406.11939. URL https://doi.org/10.48550/arXiv.2406.11939.

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

[20] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

[21] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

[22] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.153. URL https://doi.org/10.18653/v1/2023.emnlp-main.153.

[23] L. Malmqvist. Sycophancy in large language models: Causes and mitigations. *CoRR*, abs/2411.15287, 2024. doi: 10.48550/ARXIV.2411.15287. URL https://doi.org/10.48550/arXiv.2411.15287.

[24] N. Mathur, T. Baldwin, and T. Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448. URL https://aclanthology.org/2020.acl-main.448/.

[25] N. Muennighoff, Q. Liu, A. Zebaze, Q. Zheng, B. Hui, T. Y. Zhuo, S. Singh, X. Tang, L. von Werra, and S. Longpre. Octopack: Instruction tuning code large language models. *CoRR*, abs/2308.07124, 2023. doi: 10.48550/ARXIV.2308.07124. URL https://doi.org/10.48550/arXiv.2308.07124.

[26] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

[28] E. Perez, S. Ringer, K. Lukosiute, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish, S. Johnston, S. Kravec, S. E. Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan,

T. Hume, Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. B. Grosse, D. Hernandez, D. Ganguli, E. Hubinger, N. Schiefer, and J. Kaplan. Discovering language model behaviors with model-written evaluations. In A. Rogers, J. L. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13387–13434. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.847. URL https://doi.org/10.18653/v1/2023.findings-acl.847.

[29] E. Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, Sept. 2018. doi: 10.1162/coli_a_00322. URL https://aclanthology.org/J18-3002/.

[30] P. Röttger, H. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In K. Duh, H. Gómez-Adorno, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5377–5400. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.301. URL https://doi.org/10.18653/v1/2024.naacl-long.301.

[31] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

[32] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize from human feedback. *CoRR*, abs/2009.01325, 2020. URL https://arxiv.org/abs/2009.01325.

[33] P. Verga, S. Hofstätter, S. Althammer, Y. Su, A. Piktus, A. Arkhangorodsky, M. Xu, N. White, and P. Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *CoRR*, abs/2404.18796, 2024. doi: 10.48550/ARXIV.2404.18796. URL https://doi.org/10.48550/arXiv.2404.18796.

[34] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, L. Kong, Q. Liu, T. Liu, and Z. Sui. Large language models are not fair evaluators. In L. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9440–9450. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.511. URL https://doi.org/10.18653/v1/2024.acl-long.511.

[35] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, W. Ye, S. Zhang, and Y. Zhang. Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization. *CoRR*, abs/2306.05087, 2023. doi: 10.48550/ARXIV.2306.05087. URL https://doi.org/10.48550/arXiv.2306.05087.

[36] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin. Do-not-answer: Evaluating safeguards in llms. In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 896–911. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-eacl.61.

[37] Z. Zeng, J. Yu, T. Gao, Y. Meng, T. Goyal, and D. Chen. Evaluating large language models at evaluating instruction following. *CoRR*, abs/2310.07641, 2023. doi: 10.48550/ARXIV.2310.07641. URL https://doi.org/10.48550/arXiv.2310.07641.

[38] X. Zhang, B. Yu, H. Yu, Y. Lv, T. Liu, F. Huang, H. Xu, and Y. Li. Wider and deeper LLM networks are fairer LLM evaluators. *CoRR*, abs/2308.01862, 2023. doi: 10.48550/ARXIV.2308.01862. URL https://doi.org/10.48550/arXiv.2308.01862.

[39] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

[40] L. Zhu, X. Wang, and X. Wang. Judgelm: Fine-tuned large language models are scalable judges. *CoRR*, abs/2310.17631, 2023. doi: 10.48550/ARXIV.2310.17631. URL https://doi.org/10.48550/arXiv.2310.17631.

A Prompts

Prompt Template for Coherence Prompt used for the Coherence criterion. Few-shot examples are provided in the GitHub repo [1].

```
As an impartial evaluator, your task is to assess
    the coherence of a given summary in relation
    to its source material by assigning a score
    from 1 to 5 and providing a detailed
    explanation in Italian that justifies your
    rating.
Focus on how well the summary is organized and
    whether it presents the source's information
    in a logical and structured way.
Coherence refers to how well the sentences in the
    summary flow together to form a unified whole.
    A coherent summary should present the main
    ideas in a clear, logical progression,
    avoiding any abrupt shifts or disjointed
    facts. The goal is for the reader to easily
    follow the line of reasoning or narrative
    without confusion.

Evaluation Criteria
To conduct a thorough assessment, consider the
    following sub-criteria:
    Logical Structure and Organization: Assess
        whether the summary follows a clear
        progression of ideas (introduction, body,
        conclusion) that mirrors the source
        material.
    Transitions: Evaluate if there are smooth
        transitions between sentences and
        paragraphs that facilitate the reader's
        understanding.
    Clarity and Conciseness: Determine if the
        language is precise and unambiguous,
        effectively conveying the core ideas
        without unnecessary complexity.

Evaluation Process
Review the Source Material: Thoroughly read the
    source document to understand its main facts,
    events, and details.
Analyze the summary: Compare the summary against
    the source material, evaluating it based on
    the sub-criteria outlined above.
Assign a Coherence Score and provide an
    Explanation:
Based on your analysis, assign a coherence score
    from 1 to 5, where the levels are defined as
    follows.
    Score 1 (Very Poor Coherence):
        The summary is highly disorganized with
            abrupt transitions. The summary
            exhibits little to no logical flow. It
            is difficult to understand the
            relationship between concepts.
    Score 2 (Poor Coherence):
        The summary shows some attempt at
            organization but remains fragmented
            with several abrupt shifts. Key points
            are only partially integrated in a
            fluent narrative. The sentences are
            fragmented with abrupt transitions.
            The lack of clear connections between
            ideas results in a choppy reading
            experience.
    Score 3 (Moderate Coherence):
```

[1] Code, Prompts, Data and Results: https://github.com/ZanichelliEditore/llm-summarization-evaluation

```
        The summary is reasonably organized with a
            generally logical progression.
            Transitions exist but may be uneven,
            they could be smoother.
    Score 4 (Good Coherence):
        The summary is well-structured with a
            clear and logical order of ideas. It
            features smooth transitions between
            sentences and paragraphs, making it
            easy to follow. The summary is
            coherent and flows well, with clear
            connections between ideas.
    Score 5 (Excellent Coherence):
        The summary exhibits exceptional
            coherence. The transitions are
            flawless and the presentation of the
            source material is clear and unified.
Provide your score along with a detailed
    explanation in italian that justifies your
    rating, referencing specific examples and
    observations from your evaluation.

Output in the following json template: {% raw
    %}```{'score': '<score between 1 and 5 from
    very poor to excellent>', 'explanation':
    '<spigazione del voto dato al riassunto
    basandosi sullo specifico criterio di
    valutazione>'}```{% endraw %}
Update values enclosed in <> and remove the <>.
Your response must only be the updated json
    template beginning with { and ending with }
Ensure the following output keys are present in
    the json: score explanation
{{few_shot_examples_coherence}}
Now Evaluate:
<Input>
    <Source_Material>
        <Text>{{document}}</Text>
    </Source_Material>
    <Summary>
        <Text>{{summary}}</Text>
    </Summary>
</Input>
<Output>
```

Prompt Template for Consistency Prompt used for the Consistency criterion. Few-shot examples are provided in the GitHub repo [2].

```
As an impartial evaluator, your task is to assess
    the consistency of a given summary in relation
    to its source material by assigning a score
    from 1 to 5 and providing a detailed
    explanation in Italian that justifies your
    rating.
Consistency refers to the degree to which the
    summary accurately and faithfully represents
    the factual content of the source without
    introducing contradictions, inaccuracies, or
    unsupported information.
A consistent summary should align closely with the
    source material, ensuring that all presented
    information is both accurate and verifiable.

Evaluation Criteria
To conduct a thorough assessment, consider the
    following sub-criteria:
    Factual Accuracy: Verify that the summary
        accurately represents explicit facts from
        the source, including names, dates,
        numbers, and locations. Cross-reference
        specific claims in the summary with the
        source to confirm their precision.
    Absence of Contradictions: Ensure that the
        summary does not contain information that
        directly contradicts the source material.
        Identify any opposing statements or
        conflicting details between the summary
        and the source.
    Absence of Hallucinations (Extrinsic
        Consistency): Check that the summary does
        not introduce information absent from the
        source. All details should be traceable to
        the original text, and any unsubstantiated
        additions should be noted.
    Logical Inferences (Intrinsic Consistency):
        Assess whether any inferences or
        conclusions drawn in the summary are
        logically supported by the information
        provided in the source. Ensure that
        deductions are valid and reasonable based
        on the source material.
    Terminology Alignment: Confirm that the
        summary uses the same key terms and refers
        to entities consistently with the source
        material. While paraphrasing is
        acceptable, maintaining consistency in
        terminology is important for clarity and
        accuracy.

Evaluation Process
Review the Source Material: Thoroughly read the
    source document to understand its main facts,
    events, and details.
Analyze the summary: Compare the summary against
    the source material, evaluating it based on
    the sub-criteria outlined above.
Assign a Consistency Score and provide an
    Explanation:
    Based on your analysis, assign a consistency
        score from 1 to 5, where the levels are
        defined as follows.
    Score 1 (Very Poor Consistency):
        The summary contains significant factual
            inaccuracies, contradictions,
            hallucinated details, or
            misrepresentations that severely
            distort the source material.
        The summary introduces entirely fabricated
            events or represents critical
            information such that it no longer
            reflects the source.
    Score 2 (Poor Consistency):
        The summary has multiple errors and
            inconsistencies; while some key facts
            may be correct, there are notable
            inaccuracies or added details that
            conflict with the source material.
        The summary includes several incorrect
            dates, names, or details that
            contradict the source, resulting in a
            misleading representation.
    Score 3 (Moderate Consistency):
        The summary is generally accurate but
            contains minor errors, omissions, or
            slight paraphrasing issues that affect
            the overall precision.
        Most details match the source, but a few
            minor discrepancies or vague terms
            slightly reduce the clarity of the
            summary.
    Score 4 (Good Consistency):
        The summary is largely consistent with the
            source, with only trivial
            discrepancies that do not impact the
            overall factual integrity.
        The summary accurately reflects the main
            facts and events, with only minor
            stylistic differences that do not
            alter the meaning.
    Score 5 (Excellent Consistency):
```

[2] Code, Prompts, Data and Results: https://github.com/ZanichelliEditore/llm-summarization-evaluation

```
        The summary is fully consistent with the
            source, accurately representing every
            key fact and detail without any added
            or contradictory information.
        The summary perfectly mirrors the source
            material, ensuring that every piece of
            information is correctly and
            completely conveyed.
Provide your score along with a detailed
    explanation in italian that justifies your
    rating, referencing specific examples and
    observations from your evaluation.

Output in the following json template: {% raw
    %}```{'score': '<score between 1 and 5 from
    very poor to excellent>', 'explanation':
    '<spigazione del voto dato al riassunto
    basandosi sullo specifico criterio di
    valutazione>'}```{% endraw %}
Update values enclosed in <> and remove the <>.
Your response must only be the updated json
    template beginning with { and ending with }
Ensure the following output keys are present in
    the json: score explanation
{{few_shot_examples_consistency}}
Now Evaluate:
<Input>
    <Source_Material>
        <Text>{{document}}</Text>
    </Source_Material>
    <Summary>
        <Text>{{summary}}</Text>
    </Summary>
</Input>
<Output>
```

Prompt Template for Fluency Prompt used for the Fluency criterion. Few-shot examples are provided in the GitHub repo [3].

```
As an impartial evaluator, your task is to assess
    the fluency of a given summary by assigning a
    score from 1 to 5 and providing a detailed
    explanation in Italian that justifies your
    rating.
Fluency refers to the readability and overall
    quality of the summary's writing. This
    includes assessing grammar, spelling,
    punctuation, word choice, and sentence
    structure. A fluent summary should be free
    from errors that make the text difficult to
    read or understand.

Evaluation Criteria
To conduct a thorough assessment, consider the
    following sub-criteria.
    Grammar: Check for accuracy, tense
        consistency, and overall syntax.
    Spelling: Identify any spelling mistakes or
        typographical errors.
    Punctuation: Assess proper punctuation usage
        and its contribution to clarity.
    Word Choice: Evaluate whether vocabulary and
        phrasing are appropriate for the context.
    Sentence Structure: Determine if sentences are
        well-constructed, varied, and natural.

Evaluation Process
Read the summary carefully.
Check for errors:
    Are there grammatical errors? Are there
        frequent or severe errors present?
    Is there any spelling or punctuation mistakes?
    Does the word choice suit the context without
        being overly complex or too simplistic?
Assign a Fluency Score and provide an Explanation:
    Based on your analysis, assign a coherence
        score from 1 to 5, where the levels are
        defined as follows.
    Score 1 (Very Poor Fluency):
        Numerous errors in grammar, spelling,
            punctuation, and word/sentence
            construction make the summary
            extremely difficult to read.
    Score 2 (Poor Fluency):
        Frequent errors are present that interfere
            with understanding. Sentence structure
            and vocabulary choices are suboptimal,
            leading to a choppy flow.
    Score 3 (Moderate Fluency):
        Errors exist, but they do not hinder
            summary understandability. Occasional
            awkward phrasing or punctuation
            mistakes are present.
    Score 4 (Good Fluency):
        The summary is well-written with only
            isolated, minor errors. Grammar,
            spelling, punctuation, and sentence
            structure are correct, ensuring smooth
            readability.
    Score 5 (Excellent Fluency):
        The summary is polished and flawless, with
            impeccable grammar, spelling,
            punctuation, word choice, and sentence
            structure that provide a natural flow.
Provide your score along with a detailed
    explanation in italian that justifies your
    rating, referencing specific examples and
    observations from your evaluation.

Output in the following json template: {% raw
    %}```{'score': '<score between 1 and 5 from
    very poor to excellent>', 'explanation':
    '<spigazione del voto dato al riassunto
    basandosi sullo specifico criterio di
    valutazione>'}```{% endraw %}
Update values enclosed in <> and remove the <>.
Your response must only be the updated json
    template beginning with { and ending with }
Ensure the following output keys are present in
    the json: score explanation
{{few_shot_examples_fluency}}
Now Evaluate:
<Input>
    <Source_Material>
        <Text>{{document}}</Text>
    </Source_Material>
    <Summary>
        <Text>{{summary}}</Text>
    </Summary>
</Input>
<Output>
```

Prompt Template for Relevance Prompt used for the Relevance criterion. Few-shot examples are provided in the GitHub repo [4].

```
As an impartial evaluator, your task is to assess
    the relevance of a given summary in relation
    to its source material by assigning a score
    from 1 to 5 and providing a detailed
    explanation in Italian that justifies your
    rating.
Relevance refers to how well the summary includes
    only the most important and necessary content
    from the source material, without introducing
    redundant or irrelevant details.
A relevant summary should focus on the key points
    of the source and avoid unnecessary or
    excessive information.

Evaluation Criteria
```

[3] Code, Prompts, Data and Results: https://github.com/ZanichelliEditore/llm-summarization-evaluation

[4] Code, Prompts, Data and Results: https://github.com/ZanichelliEditore/llm-summarization-evaluation

```
To conduct a thorough assessment, consider the
    following sub-criteria.
    Content Coverage and Accuracy: Does the
        summary capture all of the primary
        arguments, data points, or ideas presented
        in the source document? Is the information
        presented in the summary faithful to the
        original intent and details of the source?
    Conciseness and Clarity:Is the summary
        expressed in a concise manner that does
        not sacrifice the essential details? Are
        the ideas presented clearly and
        straightforwardly, ensuring that the
        summary Does not confuse the reader with
        verbose or circular language?
    Elimination of Redundancy and Irrelevance:
        Removal of Superfluous Information: Does
        the summary avoid including unnecessary
        background or repetitive details that do
        not contribute to understanding the
        source? Are only the important and
        relevant aspects of the source material
        captured, with a clear focus on the
        essential message?
    Omission of Critical Elements: Does the
        summary omit any critical elements or
        supporting details that are necessary for
        a complete and accurate understanding of
        the source document?


Evaluation Process
Review the Source Material: Thoroughly read the
    source document to understand its main facts,
    events, and details.
Analyze the summary: Compare the summary against
    the source material, evaluating it based on
    the sub-criteria outlined above.
Assign a Consistency Score and provide an
    Explanation:
    Score 1 (Very Poor Relevance):
        The summary includes little to none of the
            key points from the source.
        The summary is Overburdened with
            irrelevant, redundant, or incorrect
            details.
        Critical points are missing from the
            summary, leading to a distorted or
            incomplete picture.
    Score 2 (Poor Relevance):
        The summary captures some primary points,
            but many important aspects are either
            omitted or misrepresented.
        The summary includes redundant or
            extraneous information that dilutes
            the primary message.
        Key supporting details are missing,
            reducing the summary's overall
            reliability.
    Score 3 (Fair Relevance):
        The summary captures more than half of the
            key points, but some secondary details
            or nuanced information may be lacking.
        The summary is mostly concise with minor
            instances of unnecessary details or
            slight redundancy.
        Less-critical details may be omitted from
            the summary without drastically
            affecting the overall understanding.
    Score 4 (Good Relevance):
        Successfully includes nearly all important
            points and supporting details from the
            source.
        The summary is clear and succinct, with
            minimal, if any, redundant content.
        Rare omissions that do not significantly
            impair the overall understanding of
            the summary.
    Score 5 (Excellent Relevance):
        The summary completely captures all
            essential points and nuances of the
            source material.
        The summary is extremely concise and
            clear, with no unnecessary or
            redundant information.
        No significant information is omitted; the
            summary is a precise and complete
            representation of the source.
Provide your score along with a detailed
    explanation in italian that justifies your
    rating, referencing specific examples and
    observations from your evaluation.

Output in the following json template: {% raw
    %}```{'score': '<score between 1 and 5 from
    very poor to excellent>', 'explanation':
    '<spigazione del voto dato al riassunto
    basandosi sullo specifico criterio di
    valutazione>'}```{% endraw %}
Update values enclosed in <> and remove the <>.
Your response must only be the updated json
    template beginning with { and ending with }
Ensure the following output keys are present in
    the json: score explanation
{{few_shot_examples_relevance}}
Now Evaluate:
<Input>
    <Source_Material>
        <Text>{{document}}</Text>
    </Source_Material>
    <Summary>
        <Text>{{summary}}</Text>
    </Summary>
</Input>
<Output>
```

Prompt Template for Ordering Prompt used for the Ordering criterion. Few-shot examples are provided in the GitHub repo [5].

```
As an impartial evaluator, your task is to assess
    the ordering of a given summary in relation to
    the ordering of the source material by
    assigning a score from 1 to 5 and providing a
    detailed explanation in Italian that justifies
    your rating.
Focus on how well the sequence of information in
    the summary mirrors the order in which it is
    presented in the source material.
Ordering refers to how closely the summary adheres
    to the structure of the source material. A
    well-ordered summary should present the key
    points in the same sequence as they appear in
    the source, ensuring a logical and coherent
    flow of information.

Evaluation Criteria
To conduct a thorough assessment, consider the
    following sub-criteria.
    Chronological/Logical Order: Confirm that if
        the source is structured chronologically
        or by logical argument, the summary
        upholds that framework. Deviations should
        be penalized based on their impact on the
        intended progression.
    Segmentation and Grouping: Consider if the
        summary correctly groups related
        information as seen in the source.
        Grouping similar ideas ensures that the
        coherence of the original narrative is
        maintained.
    Cohesion and Comprehension Impact: Assess if
        any deviation
        (omission/insertion/reordering)
```

[5] Code, Prompts, Data and Results: https://github.com/ZanichelliEditore/llm-summarization-evaluation

```
        significantly affects the reader's ability
        to follow and understand the overall
        narrative.

Evaluation Process
Review the Source Material: Thoroughly read the
    source document focusing on the ordering of
    the main facts, events, and details.
Analyze the summary: Compare the ordering of the
    main facts, events, and details of the summary
    against the source material, evaluating it
    based on the sub-criteria outlined above.
Assign a Ordering Score and provide an Explanation:
    Score 1 (Very Poor Ordering):
        Key points are not only out of sequence
            but the entire summary structure is
            different than the source material.
        The summary introduces significant
            confusion, hindering comprehension of
            the source narrative.
        Major segments are reversed or
            intermingled.
    Score 2 (Poor Ordering):
        The majority of the summary's structure
            deviates from the source order.
        Several key transitional phrases and
            segments are misplaced or omitted.
        Although some basic structure might be
            discernible, it still leads to a
            disjointed narrative.
        Noticeable reordering with multiple
            inconsistencies.
    Score 3 (Fair Ordering):
        The summary preserves parts of the source
            order while containing noticeable
            reordering in other sections.
        Some transitions and sequencing are
            maintained correctly, though there are
            occasional inconsistencies.
        The overall narrative is understandable,
            but the flow is less coherent than the
            source.
        A mixed pattern of accurate segments and
            segments with minor shifts.
    Score 4 (Good Ordering):
        The summary largely follows the sequence
            of the source material.
        Most key points and transitional phrases
            maintain their original order.
        Minor deviations may exist but do not
            materially disrupt the overall
            coherence or logical flow.
        Nearly complete alignment with the
            source's narrative structure.
        Any reordering is minimal and does not
            limit comprehension.
    Score 5 (Excellent Ordering):
        The summary mirrors exactly the structure
            of the source material.
        All key segments, transitional cues, and
            the logical narrative flow are
            preserved.
        The reader can effortlessly follow the
            progression as intended in the
            original document.
        Consistent preservation of order, ensuring
            clarity and cohesion.
        The sequence of information is methodical
            and reflective of the source.
Provide your score along with a detailed
    explanation in italian that justifies your
    rating, referencing specific examples and
    observations from your evaluation.

Output in the following json template: {% raw
    %}```{'score': '<score between 1 and 5 from
    very poor to excellent>', 'explanation':
    '<spigazione del voto dato al riassunto
    basandosi sullo specifico criterio di
    valutazione>'}```{% endraw %}
Update values enclosed in <> and remove the <>.
Your response must only be the updated json
    template beginning with { and ending with }
Ensure the following output keys are present in
    the json: score explanation
{{few_shot_examples_ordering}}
Now Evaluate:
<Input>
    <Source_Material>
        <Text>{{document}}</Text>
    </Source_Material>
    <Summary>
        <Text>{{summary}}</Text>
    </Summary>
</Input>
<Output>
```

B Annotation Guidelines

We developed a dedicated set of annotator guidelines to support the evaluation of Italian summaries according to editorial standards. They aim to ensure consistency and inter-annotator agreement in the qualitative evaluation of summaries. You can find the guidelines in the pages below in both English and Italian.

Annotation Guidelines

These guidelines define the process and criteria for evaluating a text summary based on five dimensions: coherence, consistency, relevance, fluency, and ordering. For each dimension, the following are provided:

1. Description of the task
2. Definition of the evaluation criterion and subcriteria
3. Rating scale from 1 to 5 with description of each level

Each summary in the Google sheet should be rated according to the following 5 criteria by entering its rating from 1 to 5 in the column of the same name as the criterion.

1. Coherence

Task description: Assess how well the summary presents the information in the text in a logical and structured way.

Definition: Coherence measures the fluency and unity of the text, that is, how logically the sentences flow, avoiding abrupt or discontinuous transitions.

Subcriteria:

- **Logical progression of ideas:** Ideas are presented in an order that follows a logical and natural thread.
- **Clarity and conciseness:** Sentences are formulated clearly and concisely.
- **Presence of transitions:** Connectives and transitions are used to tie sentences and paragraphs together.

Rating scale:

- **1 (Very Poor):** Disorganised text, absent transitions, difficult to follow the thread of discourse.
- **2 (Poor):** Fragmented structure, abrupt transitions, narrative not very fluid.
- **3 (Moderate):** Generally logical progression, transitions present but irregular.
- **4 (Good):** Clear structure, smooth transitions, easy to follow.
- **5 (Excellent):** Impeccable coherence, perfectly connected passages.

2. Consistency

Task description: Check the factual accuracy of the summary against the original text.

Definition: Consistency measures the fidelity of facts: absence of contradictions, errors and information not present in the source text.

Subcriteria:

- **Factual Accuracy:** All statements in the summary correspond to the facts expressed in the source text.
- **Absence of contradictions:** No part of the summary contradicts what is stated in the source text.
- **Absence of hallucinations:** No invented or added information is present that does not appear in the source text.
- **Logical Inference:** The inferred information is consistent with and supported by the content of the original text.
- **Terminological alignment:** Terms used in the summary are consistent with those in the source text, especially for technical or specialised concepts.

Rating scale:

- **1 (Very poor):** Numerous inaccuracies and invented details.
- **2 (Poor):** Multiple errors and discrepancies.
- **3 (Moderate):** Generally accurate, but with minor inaccuracies.
- **4 (Good):** Very few negligible discrepancies.
- **5 (Excellent):** Total fidelity to the facts of the text.

3. Relevance

Task description: Assess whether the summary includes only the essential contents of the source text.

Definition: Relevance measures the inclusion of key points and avoids superfluous or irrelevant details.

Subcriteria:

- **Inclusion of main ideas:** The core concepts of the source text are present in the summary.
- **Conciseness:** The content is expressed briefly but completely.
- **Absence of redundancy:** There are no unnecessary repetitions.
- **Absence of critical omissions:** No essential concepts have been omitted.

Rating scale:

- **1 (Very poor):** Almost all key points are missing; it contains much irrelevant information.
- **2 (Poor):** Covers some main points but leaves out important aspects; presence of superfluous details.
- **3 (Fair):** Covers more than half of the key points; some redundancy or minor omissions.
- **4 (Good):** Includes almost all essential points; minimal redundancy.
- **5 (Excellent):** Fully covers key points; extremely concise.

4. Fluidity

Task description: Assess the linguistic quality of the summary: grammar, spelling, punctuation and style.

Definition: Fluency measures readability and the absence of linguistic errors.

Subcriteria:

- **Grammar:** Absence of grammatical errors.
- **Spelling:** Words are spelt correctly.
- **Punctuation:** Appropriate use of punctuation marks.
- **Lexical choice:** Vocabulary is appropriate and varied.
- **Sentence structure:** Sentences are well constructed and of appropriate length.

Rating scale:

- **1 (Very poor):** Numerous serious errors that hinder reading.
- **2 (Poor):** Frequent errors hinder comprehension.
- **3 (Moderate):** Minor errors do not hinder reading.
- **4 (Good):** Isolated minor errors; fluent reading.
- **5 (Excellent):** Text impeccable in every respect.

5. Ordering

Task description: Assess whether the order of the information in the summary reflects that of the source text.

Definition: Sorting measures the alignment of the information sequence with the original structure.

Subcriteria:

- **Chronological or logical order:** Events or ideas follow the temporal or logical sequence of the source text.
- **Grouping of related information:** Related information is presented together.
- **Impact on comprehensibility:** The chosen order facilitates comprehension of the content.

Rating scale:

- **1 (Very poor):** Completely different sequence; confusion.
- **2 (Poor):** Many shifts that disturb the flow.
- **3 (Fair):** Some correct sequences, but also deviations.
- **4 (Good):** Order generally respected; minimal deviations.
- **5 (Excellent):** Order identical to that of the source text.

Linee guida per l'annotazione

Queste linee guida definiscono il processo e i criteri di valutazione di una sintesi di un testo in base a cinque dimensioni: coerenza, consistenza, rilevanza, fluidità e ordinamento. Per ciascuna dimensione, sono forniti i seguenti elementi:

1. Descrizione del compito
2. Definizione del criterio e dei sottocriteri di valutazione
3. Scala di valutazione da 1 a 5 con descrizione dei singoli livelli

Ogni riassunto presente nel foglio google dovrà essere valutato secondo i seguenti 5 criteri inserendo nella colonna omonima al criterio la sua valutazione da 1 a 5.

1. Coerenza

Descrizione del compito: Valutare quanto la sintesi presenti le informazioni del testo in modo logico e strutturato.

Definizione: La coerenza misura la fluidità e l'unità del testo, ovvero quanto le frasi scorrono in modo logico evitando passaggi bruschi o discontinui.

Sottocriteri:

- **Progressione logica delle idee**: Le idee sono presentate in un ordine che segue un filo logico e naturale.
- **Chiarezza e concisione**: Le frasi sono formulate in modo chiaro e sintetico.
- **Presenza di transizioni**: Sono utilizzati connettivi e transizioni per legare le frasi e i paragrafi.

Scala di punteggio:

- **1 (Molto scarsa)**: Testo disorganizzato, transizioni assenti, difficile seguire il filo del discorso.
- **2 (Scarsa)**: Struttura frammentata, passaggi bruschi, narrazione poco fluida.
- **3 (Moderata)**: Progressione generalmente logica, transizioni presenti ma irregolari.
- **4 (Buona)**: Struttura chiara, transizioni fluide, facile da seguire.
- **5 (Eccellente)**: Coerenza impeccabile, passaggi perfettamente raccordati.

2. Consistenza

Descrizione del compito: Verificare l'accuratezza fattuale della sintesi rispetto al testo originale.

Definizione: La consistenza misura la fedeltà dei fatti: assenza di contraddizioni, errori e informazioni non presenti nel testo sorgente.

Sottocriteri:

- **Accuratezza fattuale**: Tutte le affermazioni presenti nella sintesi corrispondono ai fatti espressi nel testo sorgente.
- **Assenza di contraddizioni**: Nessuna parte della sintesi contraddice ciò che è riportato nel testo originale.
- **Assenza di allucinazioni**: Non sono presenti informazioni inventate o aggiunte che non compaiono nel testo sorgente.
- **Inferenza logica**: Le informazioni dedotte sono coerenti e supportate dal contenuto del testo originale.
- **Allineamento terminologico**: I termini utilizzati nella sintesi sono coerenti con quelli del testo sorgente, soprattutto per concetti tecnici o specialistici.

Scala di punteggio:

- **1 (Molto scarsa)**: Numerose imprecisioni e dettagli inventati.
- **2 (Scarsa)**: Errori e discrepanze multiple.
- **3 (Moderata)**: Generalmente accurata, ma con piccole imprecisioni.
- **4 (Buona)**: Pochissime discrepanze trascurabili.
- **5 (Eccellente)**: Fedeltà totale ai fatti del testo.

3. Rilevanza

Descrizione del compito: Valutare se la sintesi include solo i contenuti essenziali del testo sorgente.

Definizione: La rilevanza misura l'inclusione dei punti chiave ed evita dettagli superflui o irrilevanti.

Sottocriteri:

- **Inclusione delle idee principali**: I concetti fondamentali del testo sorgente sono presenti nella sintesi.
- **Concisione**: Il contenuto è espresso in modo breve ma completo.
- **Assenza di ridondanze**: Non vi sono ripetizioni inutili.
- **Assenza di omissioni critiche**: Nessun concetto essenziale è stato omesso.

Scala di punteggio:

- **1 (Molto scarsa)**: Mancano quasi tutti i punti chiave; contiene molte informazioni irrilevanti.
- **2 (Scarsa)**: Copre alcuni punti principali ma tralascia aspetti importanti; presenza di dettagli superflui.
- **3 (Discreta)**: Copertura di oltre metà dei punti chiave; qualche ridondanza o omissione minore.
- **4 (Buona)**: Include quasi tutti i punti essenziali; minima ridondanza.
- **5 (Eccellente)**: Copre completamente i punti chiave; estremamente concisa.

4. Fluidità

Descrizione del compito: Valutare la qualità linguistica della sintesi: grammatica, ortografia, punteggiatura e stile.

Definizione: La fluidità misura la leggibilità e l'assenza di errori linguistici.

Sottocriteri:

- **Grammatica**: Assenza di errori grammaticali.
- **Ortografia**: Le parole sono scritte correttamente.
- **Punteggiatura**: Uso appropriato dei segni di punteggiatura.
- **Scelta lessicale**: Il vocabolario è appropriato e vario.
- **Struttura della frase**: Le frasi sono ben costruite e di lunghezza adeguata.

Scala di punteggio:

- **1 (Molto scarsa)**: Numerosi errori gravi che impediscono la lettura.
- **2 (Scarsa)**: Errori frequenti che ostacolano la comprensione.
- **3 (Moderata)**: Errori lievi non ostacolano la lettura.
- **4 (Buona)**: Isolati errori minori; lettura scorrevole.
- **5 (Eccellente)**: Testo impeccabile sotto ogni profilo.

5. Ordinamento

Descrizione del compito: Valutare se l'ordine delle informazioni nella sintesi rispecchia quello del testo sorgente.

Definizione: L'ordinamento misura l'allineamento della sequenza informativa con la struttura originale.

Sottocriteri:

- **Ordine cronologico o logico**: Gli eventi o le idee seguono la sequenza temporale o logica del testo sorgente.
- **Raggruppamento di informazioni correlate**: Le informazioni connesse tra loro sono presentate insieme.
- **Impatto sulla comprensibilità**: L'ordine scelto facilita la comprensione del contenuto.

Scala di punteggio:

- **1 (Molto scarsa)**: Sequenza completamente diversa; confusione.
- **2 (Scarsa)**: Molti spostamenti che disturbano il flusso.
- **3 (Discreta)**: Alcune sequenze corrette, ma anche deviazioni.
- **4 (Buona)**: Ordine generalmente rispettato; deviazioni minime.
- **5 (Eccellente)**: Ordine identico a quello del testo sorgente.