LLMs instead of Human Judges?
A Large Scale Empirical Study across 20 NLP Evaluation Tasks

Anna Bavaresco[1], Raffaella Bernardi[2], Leonardo Bertolazzi[2], Desmond Elliott[3], Raquel Fernández[1], Albert Gatt[4], Esam Ghaleb[5], Mario Giulianelli[6], Michael Hanna[1], Alexander Koller[7], André F. T. Martins[8], Philipp Mondorf[9], Vera Neplenbroek[1], Sandro Pezzelle[1], Barbara Plank[9], David Schlangen[10], Alessandro Suglia[11], Aditya K Surikuchi[1], Ece Takmaz[4], Alberto Testoni[12]

[1]University of Amsterdam, [2]University of Trento, [3]University of Copenhagen, [4]Utrecht University, [5]Max Planck Institute for Psycholinguistics, [6]ETH Zürich, [7]Saarland University, [8]Universidade de Lisboa & Unbabel, [9]LMU Munich & MCML, [10]University of Potsdam, [11]Heriot-Watt University [12]Amsterdam UMC

judgebench@googlegroups.com[*]

Abstract

There is an increasing trend towards evaluating NLP models with LLMs instead of human judgments, raising questions about the validity of these evaluations, as well as their reproducibility in the case of proprietary models. We provide JUDGE-BENCH, an extensible collection of 20 NLP datasets with human annotations covering a broad range of evaluated properties and types of data, and comprehensively evaluate 11 current LLMs, covering both open-weight and proprietary models, for their ability to replicate the annotations. Our evaluations show substantial variance across models and datasets. Models are reliable evaluators on some tasks, but overall display substantial variability depending on the property being evaluated, the expertise level of the human judges, and whether the language is human or model-generated. We conclude that LLMs should be carefully validated against human judgments before being used as evaluators.

https://github.com/dmg-illc/JUDGE-BENCH

1 Introduction

For many natural language processing (NLP) tasks, the most informative evaluation is to ask humans to judge the model output. Such judgments are traditionally collected in lab experiments or through crowdsourcing, with either expert or non-expert annotators, as illustrated in Fig. 1. Recently, there has been a trend towards replacing human judgments with automatic assessments obtained via large language models (LLMs; Chiang and Lee, 2023; Wang et al., 2023; Liu et al., 2023; Li et al., 2024; Zheng et al., 2024, *inter alia*). For example, the LLM

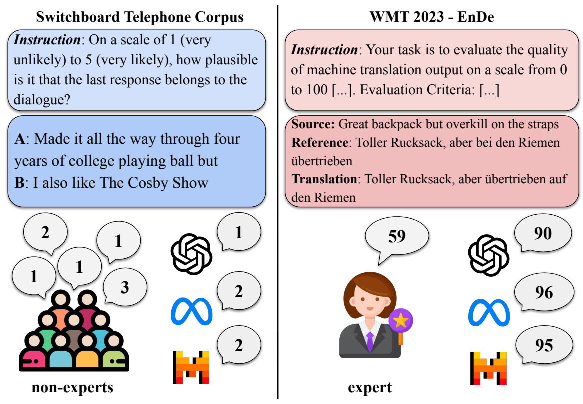

Figure 1: Evaluation by expert and non-expert human annotators and by LLMs for two tasks involving human-generated (left) and machine-generated text (right).

could be instructed to rate a response generated by a dialogue system for its perceived plausibility, on a scale from 1 to 5. This drastically reduces the evaluation effort and is claimed to yield more reliable results across multiple evaluation rounds (Landwehr et al., 2023; Jiang et al., 2023b; Reiter, 2024; Dubois et al., 2024).

At the same time, the use of LLMs as judges of linguistic output raises new concerns: LLMs may be prone to errors or systematic biases that differ from those of humans, especially on subtle tasks such as evaluating toxicity, or reasoning. This may distort evaluation results and lead to incorrect conclusions. The problem is aggravated by explicit or implicit data leakage (Balloccu et al., 2024), which undermines the ability to make broad, generalisable claims beyond the single specific dataset under analysis. Specifically for closed models such as OpenAI's GPT series, there are serious reproducibility concerns, as LLMs may be retrained or retired at any time, making subsequent

[*]Authors listed in alphabetical order.

comparisons invalid or impossible.

Previous studies offer mixed evidence regarding the reliability of LLM evaluators. Some research concludes that they are effective, correlating well with human judgments (Liu et al., 2023; Zheng et al., 2024; Chen et al., 2023; Verga et al., 2024; Törnberg, 2023; Huang et al., 2024; Naismith et al., 2023; Gilardi et al., 2023; Kocmi and Federmann, 2023b), albeit with some caveats (Wang et al., 2023; Wu and Aji, 2025; Hada et al., 2024; Pavlovic and Poesio, 2024). In some cases, LLM evaluators can also provide pairwise preference judgments (Kim et al., 2024; Liusie et al., 2024; Liu et al., 2024a; Park et al., 2024; Tan et al., 2025), or fine-grained evaluation beyond a single score, such as error spans (Fernandes et al., 2023; Kocmi and Federmann, 2023a). In contrast, some studies highlight substantial biases in LLMs' behaviour as evaluators, both as compared against human judgments (Koo et al., 2024; Zeng et al., 2024; Baris Schlicht et al., 2024) and through intrinsic analyses (Wang et al., 2024; Liu et al., 2024b; Stureborg et al., 2024). These discrepancies likely stem from the limitations of this previous work, which typically relies on a few datasets and models, often restricted to closed-source proprietary models. The observation of such limitations has motivated recent work to develop finetuning methods for LLM judges designed to overcome certain biases (Zhu et al., 2025).

In this paper, we examine how well current LLMs can approximate human evaluators on a large scale. We prompt 11 among the most recent open-weight and proprietary LLMs to generate judgments on 20 datasets with human annotations on a wide range of quality dimensions, prompt styles, and tasks. Our evaluation goes beyond existing work by including a wide variety of *datasets* that differ in the *type of task* (e.g., translation, dialogue generation, etc.), the *property* being judged (e.g., coherence, fluency, etc.), the *type of judgments* (categorical or graded), and the *expertise of human annotators* (experts or non-experts). We provide JUDGE-BENCH, a benchmark which includes, upon release, a total of over 70,000 test instances with associated human judgments with an extensible codebase.

Our results indicate that LLMs align well with human judgments on certain tasks, like instruction following. However, their performance is *inconsistent* across and within annotation tasks. Elicitation methods like Chain-of-Thought prompting (Wei et al., 2022) do not reliably improve agreement, in line with recent findings (Sprague et al., 2025). Some proprietary models—in particular, GPT-4o—align better to humans, but there is a rather small gap with large open-source models, holding promise for the reproducibility of future evaluation efforts. Altogether, at the current stage of LLM development, we recommend validating LLM judges against task-specific human annotations before deploying them for any particular task.

2 Construction of JUDGE-BENCH

One key feature that differs across the datasets included in JUDGE-BENCH is the source of the data being evaluated, i.e., whether the items to be judged are generated by a model or produced by humans, as illustrated in Figure 1.

For model-generated items, the goal is to evaluate an NLP system. This includes both classic tasks such as machine translation or dialogue response generation, as well as less standard tasks for which automation has recently become an option thanks to LLMs, such as the generation of plans or logical arguments. For human-generated items, the goal is to assess properties of interest such as grammaticality or toxicity. This distinction allows us to understand whether LLMs have a positive bias towards machine-generated outputs—a tendency reported in prior work (Xu et al., 2024).

The datasets we consider cover a wide span of properties of interest, ranging from grammaticality and toxicity to coherence, factual consistency, and verbosity, *inter alia*. Many properties are relevant across multiple tasks (e.g., fluency and coherence), while others are more task-specific (e.g., the success of a generated plan or the correctness of a multi-step mathematical reasoning trace).

Our study focuses on English datasets or language pairs which include English as one of the languages. We keep track of whether the original annotation guidelines are available and whether the annotations are provided by experts or non-experts. We retain all available individual annotations. Dataset information is summarised in Table 2, Appendix A. All 20 datasets are formatted following a precise data schema to facilitate the integration of additional datasets. This makes JUDGE-BENCH easily extensible. We provide more details about the data schema in Appendix B.

	Dataset (# properties judged)	GPT-4o	Llama-3.1-70B	Mixtral-8x22B	Gemini-1.5	Mixtral-8x7B	Comm-R+	σ	*UB*
Categorical Annotations	CoLa (1)	0.34	0.46	0.54	0.45	**0.55**	0.12	0.16	-
	CoLa-grammar (63)	**0.47** ±0.22	0.28 ±0.24	0.28 ±0.23	0.26 ±0.24	0.21 ±0.18	0.13 ±0.14	0.14	-
	ToxicChat (2)	**0.49** ±0.36	0.41 ±0.26	0.45 ±0.27	0.45 ±0.35	0.36 ±0.12	0.28 ±0.35	0.1	-
	LLMBar-natural (1)	**0.84**	0.8	0.72	0.79	0.54	0.56	0.13	-
	LLMBar-adversarial (1)	**0.58**	0.46	0.2	0.29	0.06	0.11	0.2	-
	Persona Chat (2)	0.24 ±0.34	0.24 ±0.33	**0.58** ±0.59	-0.03 ±0.04	0.54 ±0.65	0.48 ±0.74	0.2	0.88
	Topical Chat (2)	**0.05** ±0.07	-0.02 ±0.02	-0.03 ±0.04	-0.03 ±0.04	0.02 ±0.03	0.01 ±0.02	0.07	0.58
	ROSCOE-GSM8K (2)	0.59 ±0.35	**0.64** ±0.27	0.62 ±0.38	0.6 ±0.24	0.58 ±0.36	0.0	0.15	-
	ROSCOE-eSNLI (2)	0.29 ±0.06	**0.38** ±0.08	0.13 ±0.13	0.11 ±0.18	0.1 ±0.11	0.03 ±0.05	0.14	-
	ROSCOE-DROP (2)	**0.29** ±0.08	0.27 ±0.07	0.2 ±0.12	0.08 ±0.05	0.13 ±0.21	0.03 ±0.04	0.13	-
	ROSCOE-CosmosQA (2)	0.16 ±0.07	**0.25** ±0.02	0.09 ±0.17	0.14 ±0.17	0.19 ±0.05	-0.03 ±0.01	0.1	-
	QAGS (1)	**0.72**	0.7	0.66	0.65	0.68	0.13	0.23	0.74
	Medical-safety (2)	0.01 ±0.03	-0.03 ±0.06	-0.02 ±0.09	-0.03 ±0.08	0.0 ±0.06	0.01 ±0.02	0.03	-
	DICES-990 (1)	-0.24	-0.17	-0.16	-0.12	-0.2	-0.09	0.05	0.27
	DICES-350-expert (1)	-0.2	-0.13	-0.15	-0.03	-0.11	0.01	0.08	-
	DICES-350-crowdsourced (1)	-0.22	-0.18	-0.08	-0.02	-0.11	-0.08	0.07	0.32
	Inferential strategies (1)	**0.42**	0.4	0.02	0.22	0.06	-0.02	0.19	1.0
	Average Cohen's κ	0.28 ±0.32	0.28 ±0.30	0.24 ±0.30	0.22 ±0.28	0.21 ±0.28	0.10 ±0.18		
Graded Annotations	Dailydialog (1)	**0.69**	0.6	0.55	0.63	0.63	0.52	0.06	0.79
	Switchboard (1)	**0.66**	0.45	0.63	0.59	0.56	0.36	0.11	0.8
	Persona Chat (4)	**0.22** ±0.11	-0.02 ±0.2	0.16 ±0.1	0.1 ±0.09	0.02 ±0.15	0.07 ±0.13	0.2	0.61
	Topical Chat (4)	0.26 ±0.03	**0.28** ±0.1	0.13 ±0.04	0.17 ±0.12	0.21 ±0.18	0.14 ±0.05	0.07	0.56
	Recipe-generation (6)	**0.78** ±0.05	0.66 ±0.07	0.6 ±0.15	0.67 ±0.09	0.57 ±0.24	0.32 ±0.28	0.18	0.65
	ROSCOE-GSM8K (2)	0.82 ±0.12	**0.83** ±0.11	0.81 ±0.14	0.81 ±0.12	0.79 ±0.13	0.68 ±0.2	0.15	-
	ROSCOE-eSNLI (2)	**0.49** ±0.24	0.4 ±0.16	0.38 ±0.17	0.35 ±0.21	0.32 ±0.12	0.09 ±0.08	0.14	-
	ROSCOE-DROP (2)	0.57 ±0.22	**0.59** ±0.16	0.44 ±0.15	0.44 ±0.13	0.32 ±0.12	0.21 ±0.22	0.13	-
	ROSCOE-CosmosQA (2)	**0.57** ±0.18	0.55 ±0.18	0.51 ±0.16	**0.57** ±0.17	0.53 ±0.21	0.33 ±0.25	0.1	-
	NewsRoom (4)	**0.59** ±0.02	**0.59** ±0.03	0.44 ±0.05	0.55 ±0.03	0.5 ±0.07	0.36 ±0.06	0.1	0.62
	SummEval (4)	0.35 ±0.06	0.44 ±0.14	**0.54** ±0.08	0.38 ±0.02	0.48 ±0.02	0.19 ±0.06	0.13	-
	WMT 2020 En-De (1)	**0.63**	0.37	0.51	0.46	0.2	0.42	0.15	0.81
	WMT 2020 Zh-En (1)	**0.54**	0.39	0.48	0.41	0.25	0.42	0.1	0.62
	WMT 2023 En-De (1)	0.22	0.14	**0.23**	0.16	0.17	0.22	0.04	-
	WMT 2023 Zh-En (1)	0.17	0.14	**0.19**	0.14	0.15	0.15	0.02	-
	Average Spearman's ρ	0.50 ±0.21	0.43 ±0.22	0.44 ±0.19	0.43 ±0.21	0.38 ±0.22	0.30 ±0.17		

Table 1: Scores per dataset for the models with ≥98% valid response rates (results for all models in Tab. 6, App. H): Cohen's kappa for categorical annotations and Spearman's correlation for graded annotations. Boldface marks best model performance per dataset. Spearman's correlations are generally significant ($p < 0.05$), with the exception of the Persona Chat and Topical Chat datasets (see Tab. 6 in Appendix H for more details). Datasets with both categorical and graded annotations appear twice. Datasets in blue concern human-generated language, while those in red concern model-generated text. 'σ' denotes the standard deviation of the scores across models per dataset (averaged over properties if more than one is judged per dataset). Upper-bound estimates (*UB*) indicate the agreement between individual and aggregated human judgments.

3 Model Selection and Experiment Design

Models. We select representative proprietary and open-weight models of various sizes that show high performance across several tasks on the Open LLM and Chatbot Arena Leaderboards (Chiang et al., 2024): GPT-4o (OpenAI, 2024), LLaMA-3.1 (8B and 70B; AI@Meta 2024), Gemini-1.5 (Reid et al., 2024), Mixtral (8x7B and 8x22B; Jiang et al. 2024), Command R and Command R+ (Cohere and Cohere for AI, 2024a,b), OLMo (Groeneveld et al., 2024), Starling-7B (Zhu et al., 2024), and Mistral (Jiang et al., 2023a). See Appendix E for inference procedure details.

Prompts. Since most datasets include the original instructions used to gather human judgments, we use these instructions directly as prompts for the model, with additional guidelines to constrain the models' output and minimise verbosity: '*Answer with one of {}. Do not explain your answer.*' When the original instruction for collecting human judgments is unavailable, we create a prompt based on relevant information from the original paper, such as the task description and the definitions of the evaluation metrics. We also experiment with alternative prompting strategies, including Chain-of-Thought, few-shot prompts, and prompt paraphrases. However, none of these strategies leads to systematic improvements. See Appendix H for full details and results. All prompts are provided in the codebase.

Evaluation. Models do not always respond to the prompts as requested (e.g., they may refuse to

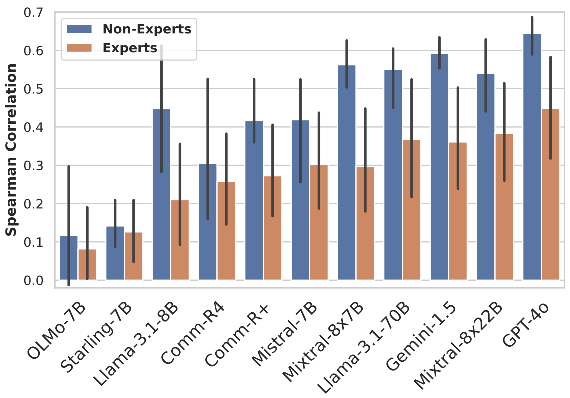

Figure 2: Average model correlation with human experts vs. non-experts in datasets with graded annotations.

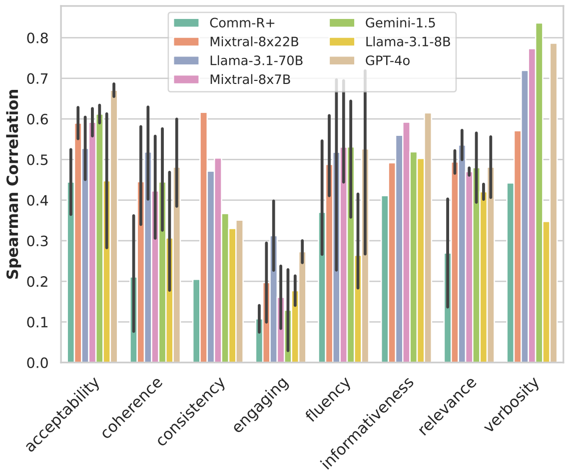

Figure 3: Correlation for properties with graded judgments. Averages and error bars when the property is present in more than one dataset.

answer if they perceive the prompt as sensitive). We therefore use the following evaluation protocol:

- To obtain the same number of judgments across models for a given dataset, we replace invalid LLM responses with judgments randomly sampled from the relevant set of categorical or graded annotations. Figure 5 in Appendix F shows the rate of valid responses per model.
- Graded annotations, such as in WMT 2020 (Freitag et al., 2021), assess language quality on a continuous scale (e.g., a score from 0 to 100, or Likert-scale ratings), capturing varying degrees of fluency, adequacy, or overall translation quality; whereas categorical annotations, like those in CoLa (Warstadt et al., 2019), involve binary judgments (e.g., grammatically acceptable or not). For the former, we compute Spearman's correlation (ρ) between model and human judgments; for the latter, we compute Cohen's κ.
- When multiple individual human judgments are available (typically three, see Table 2 in Appendix A), we estimate an upper bound by computing the average Spearman's ρ or Cohen's κ between bootstrapped single-rater responses and the aggregated responses across raters. Appendix C provides details on the upper bounds.

4 Results

Scores vary substantially across models. For any given model, they vary both across datasets and properties being judged. Table 1 presents detailed results for the 6 models that exhibit the largest rate of valid responses ($\geq$98%). GPT-4o ranks first across several evaluation scenarios, but the Llama-3.1-70B and Mixtral-8x22B open models are relatively close and outperform GPT-4o on some assessment types, such as categorical sentence acceptability (CoLa) and graded summary quality (SummEval). Overall, the high degree of variability is not fully accounted for by the inherent difficulty of the annotation tasks, as reflected in the human upper bound. Moreover, except for a few datasets (e.g., QAGS, Recipe-generation, and NewsRoom), model scores remain notably below the upper bound.

Among the property types with the lowest human-model alignment are toxicity and safety (in particular on DICES and Medical-safety), where model scores can be even negative and valid response rates particularly low (see Fig. 6 in Appendix F). This is due in part to the guardrails associated with these tasks (Weidinger et al., 2023). We find that, especially in the medical domain, many models tend to provide explanations instead of producing a judgment (see Appendix G).

Despite the high variability across models and datasets, we observe several notable trends. For graded annotations (Fig. 2), all models achieve higher correlations with annotations by non-expert human judges compared to expert annotators, echoing recent findings by Aguda et al. (2024). One possible explanation is that non-experts might rely on surface-level features, which could align more closely with the patterns LLMs are most attuned to, while experts apply stricter, domain-specific criteria. This remains speculative and calls for further investigation.

Figure 3 shows correlation results across different datasets for the subset of properties that exclusively have graded judgments. When applicable,

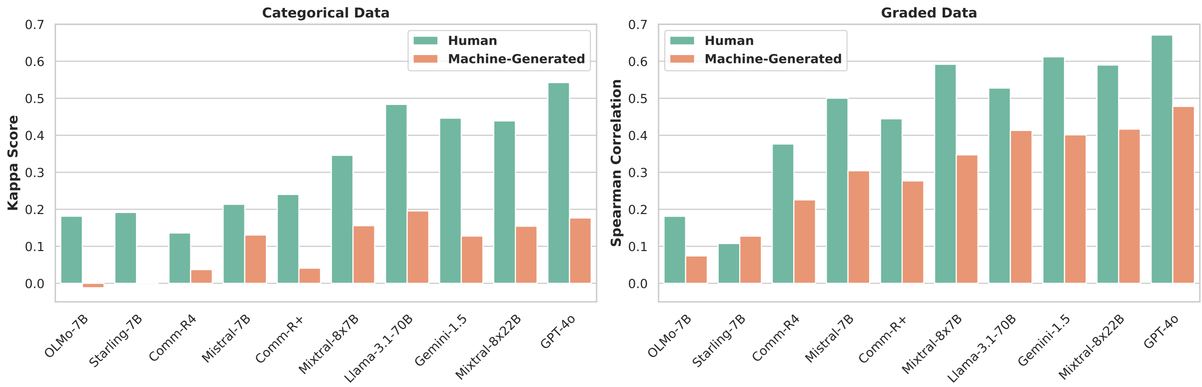

Figure 4: Scores (Cohen's κ for categorical annotations and Spearman's correlation for graded annotations) on test items involving human language vs. machine-generated outputs.

we average results across datasets including annotations for the same property. We provide more details about these properties in Appendix D. The proprietary models GPT-4o and Gemini-1.5 exhibit the highest scores when evaluating acceptability and verbosity, while the two Mixtral open models show the strongest correlations for coherence and consistency. Correlation with the engagingness property remains consistently low across all models. Overall, *no single model demonstrates a clear superiority* over others across all properties; instead, different quality dimensions are better assessed by different models. This calls into question the widespread practice of using a single model – typically a proprietary one like those from the GPT family—to evaluate a diverse range of linguistic properties.

Finally, as shown in Figure 4, all models achieve better alignment with human judgments when evaluating human language than when assessing machine-generated text, both for categorical and graded annotations. This result aligns with the findings by Xu et al. (2024), suggesting that LLMs display a bias towards their own generation. More broadly, this trend calls for caution when using LLMs to automatically evaluate the output of NLP systems.

5 Conclusions

In response to current trends in evaluation, in this paper we conducted a large-scale study of the correlation between human and LLM judgments across 20 datasets, considering factors such as the properties being assessed, the expertise level of the human judges, and whether the data is model- or human-generated. On some tasks, such as instruction following and the generation of mathematical reasoning traces, models can be reliably used as evaluators. Overall, however, models' agreement with human judgments varies widely across datasets, evaluated properties, and data sources; and depends on the level of expertise of human judges. Furthermore, elicitation strategies such as Chain-of-Thought prompting do not consistently improve agreement levels, in line with recent findings (Sprague et al., 2025). We recommend validation and calibration of LLMs against task-specific human judgments prior to their deployment as evaluators. To facilitate this process, we release JUDGE-BENCH, a benchmark that enables systematic evaluation across a diverse range of tasks and is easily extensible to include any new task of interest.

Limitations

One limitation of the experimental design of our work is that correlation with human judges may not be the most appropriate way to validate LLM evaluators. Indeed, there may be domains where human annotators and LLM evaluators appear aligned simply because they are affected by similar biases. Therefore, depending on the task at hand, it may be necessary to validate the reliability of human annotators as well.

Another limitation concerns the use of existing tasks and datasets without reassessing their quality or representativeness of actual downstream tasks. While we did our best to select a wide set of tasks meaningful to the NLP community, we acknowledge that these tasks could not be equally meaningful for end-users, and that employing existing datasets could arguably lead to potential risks and shortcomings, such as data leakage.

In contrast to approaches that use LLMs for pairwise preference evaluation, e.g., PairEval (Park et al., 2024) or JudgeBench (Tan et al., 2025),[1] this paper focuses on evaluating the performance of LLMs on generating judgements for categorical and graded responses. We leave the extension of JUDGE-BENCH to include pairwise preference evaluation and other recent evaluation methods, such as Prometheus 2 (Kim et al., 2024), for future work.

Finally, our work mostly focuses on English-language datasets—with the exception of datasets focussing specifically on machine-translation outputs. It remains to be seen whether LLMs' meta-evaluation abilities vary across different languages.

Acknowledgements

This work emerged from discussions at a workshop organised by Raquel Fernández and Sandro Pezzelle at MFO, the Oberwolfach Research Institute for Mathematics in the German Black Forest, on behalf of the ELLIS NLP programme. The event was funded by the state of Baden-Württemberg (Germany) and organised in collaboration with the ELLIS Institute Tübingen and the Max Planck Institute for Intelligent Systems. We furthermore acknowledge our funders. In particular, AB, RF, and AT were supported by the European Research Council (ERC Consolidator Grant DREAM 819455 to RF), as well as BP (ERC Consolidator Grant DIALECT 101043235 to BP). DE was supported by a research grant (VIL53122) from VILLUM FONDEN. EG was supported by the Dutch Research Council (Gravitation grant 024.001.006 to the Language in Interaction consortium). MG was supported by an ETH Zurich Postdoctoral Fellowship. MH was supported in part by an OpenAI Superalignment Fellowship. AM was supported by the European Research Council (DECOLLAGE, ERC-2022-CoG 101088763) and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020. We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

[1]Some time after an early version of this paper became available as a pre-print, accompanied by our JudgeBench code, the independent work by Tan et al. (2025) appeared, which describes a benchmark that the authors named JudgeBench. This name clash is unfortunate, but since in the meantime our paper has seen some uptake, we have decided against trying to resolve it.

References

Gavin Abercrombie and Verena Rieser. 2022. Risk-graded safety for handling medical queries in conversational AI. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 234–243, Online only. Association for Computational Linguistics.

Toyin D. Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, and Charese Smiley. 2024. Large language models as financial data annotators: A study on effectiveness and efficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10124–10145, Torino, Italia. ELRA and ICCL.

AI@Meta. 2024. Llama 3.1 model card.

Lora Aroyo, Alex Taylor, Mark Díaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. In *Advances in Neural Information Processing Systems*, volume 36, pages 53330–53342. Curran Associates, Inc.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Ipek Baris Schlicht, Defne Altiok, Maryanne Taouk, and Lucie Flek. 2024. Pitfalls of conversational LLMs on news debiasing. In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE) @ LREC-COLING 2024*, pages 33–38, Torino, Italia. ELRA and ICCL.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Cohere and Cohere for AI. 2024a. Command R Model Card.

Cohere and Cohere for AI. 2024b. Command R+ Model Card.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig, Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT rates natural language explanation quality like humans: But on which scales? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3111–3132, Torino, Italia. ELRA and ICCL.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023b. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.

Fabian Landwehr, Erika Varis Doggett, and Romann M. Weber. 2023. Memories for virtual AI characters. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 237–252, Prague, Czechia. Association for Computational Linguistics.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA. Association for Computational Linguistics.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4694–4702, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

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

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024b. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701, Bangkok, Thailand. Association for Computational Linguistics.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Philipp Mondorf and Barbara Plank. 2024. Comparing inferential strategies of humans and large language models in deductive reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9402, Bangkok, Thailand. Association for Computational Linguistics.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. 2014. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553.

OpenAI. 2024. Gpt-4o model card.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. PairEval: Open-domain dialogue evaluation with pairwise comparison. In *First Conference on Language Modeling*.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Ehud Reiter. 2024. Can LLM-based eval replace human evaluation? Blog post.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

Katharina Stein, Lucia Donatelli, and Alexander Koller. 2023. From sentence to action: Splitting AMR graphs for recipe instructions. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 52–67, Nancy, France. Association for Computational Linguistics.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. JudgeBench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations*.

Petter Törnberg. 2023. ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. *arXiv preprint arXiv:2404.18796*.

Sarenne Carrol Wallbridge, Catherine Lai, and Peter Bell. 2022. Investigating perception of spoken dialogue acceptability through surprisal. In *Proc. Interspeech 2022*, pages 4506–4510.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Alex Warstadt and Samuel R. Bowman. 2020. Linguistic analysis of pretrained sentence encoders with acceptability judgments. *Preprint*, arXiv:1901.03438.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312, Abu Dhabi, UAE. Association for Computational Linguistics.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2024. Starling-7B: Improving LLM helpfulness & harmlessness with RLAIF. In *First Conference on Language Modeling*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned Large Language Models are Scalable Judges. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. arXiv. ArXiv:2310.17631 [cs].

Appendix

A Datasets

This section provides brief descriptions of the datasets employed in our study. Table 2 summarises relevant dataset information. Note that dataset sizes as reported in Table 2 refer to the number of annotated samples (not to the total number of collected annotations) and might therefore differ from the figures reported in the original papers. Table 3 reports Krippendorf's α for those datasets with multiple public human annotations.

CoLa (Warstadt et al., 2019). The Corpus of Linguistic Acceptability (CoLA) consists of 10657 sentences from 23 linguistics publications, expertly annotated for acceptability (grammaticality) by their original authors.

CoLa-grammar (Warstadt and Bowman, 2020). The dataset consists of a grammatically annotated version of the CoLA development set. Each sentence in the CoLA development set is labelled with boolean features indicating the presence or absence of a particular grammatical construction (usually syntactic in nature). Two related sets of features are considered: 63 minor features correspond to fine-grained phenomena, and 15 major features correspond to broad classes of phenomena.

ToxicChat (Lin et al., 2023). collect binary judgments on the toxicity and 'jailbreaking' nature (prompt hacks deliberately intended to bypass safety policies and induce models to generate unsafe content) of human prompts to LLMs. While the original dataset contains a mix of human- and automatically-annotated instances, here we only consider the human-annotated prompts.

LLMBar (Zeng et al., 2024). LLMBar is a dataset targeted at evaluating the instruction-following abilities of LLMs. Each entry of this dataset consists of an instruction paired with two different outputs, one correctly following the instruction and the other deviating from it. LLMBar has an `adversarial` split where deviating outputs are carefully constructed to 'fool' LLM-based evaluators and a `natural` split where deviating outputs are more naturalistic.

Dataset	Task	Size	# Annot.	Type	Guidelines	Expert	Leaked
CoLA (Warstadt et al., 2019)	Acceptability	1,043	-	Categorical	✗	✓	✓
CoLA-grammar (Warstadt and Bowman, 2020)	Acceptability	1,043	-	Categorical	✗	✓	✓
Switchboard (Wallbridge et al., 2022)	Acceptability	100	3-6	Graded	✓	✗	
Dailydialog (Wallbridge et al., 2022)	Acceptability	100	3-6	Graded	✓	✗	
Inferential strategies (Mondorf and Plank, 2024)	Reasoning	300	-	Categorical	✓	✓	✗
ROSCOE (Golovneva et al., 2023)	Reasoning	756	-	Categorical + Graded	✓	✓	
Recipe-generation (Stein et al., 2023)	Planning	52	-	Graded	✓		
Medical-safety (Abercrombie and Rieser, 2022)	Toxicity & Safety	3,701	-	Preference	✓	✓	
DICES (Aroyo et al., 2023)	Toxicity & Safety	1,340	~70 + ~120	Categorical	✗	Mixed	
ToxicChat (Lin et al., 2023)	Toxicity & Safety	5,654	-	Categorical	✗	✓	
Topical Chat (Mehri and Eskenazi, 2020)	Dialogue	60	3	Graded + Categorical	✗	✓	
Persona Chat (Mehri and Eskenazi, 2020)	Dialogue	60	3	Graded + Categorical	✗	✓	
WMT 2020 En-De (Freitag et al., 2021)	Machine Translation	14,122	3	Graded	✗	✓	
WMT 2020 Zh-En (Freitag et al., 2021)	Machine Translation	19,974	3	Graded	✗	✓	
WMT 2023 En-De (Kocmi et al., 2023)	Machine Translation	6,588	-	Graded	✗	✓	
WMT 2023 Zh-En (Kocmi et al., 2023)	Machine Translation	13,245	-	Graded	✗	✓	
G-Eval / SummEval (Liu et al., 2023)	Summarisation	1,600	-	Graded	✓		✓
QAGS (Wang et al., 2020)	Summarisation	953	3	Categorical	✓	✗	
NewsRoom (Grusky et al., 2018)	Summarisation	420	3	Graded	✓	✗	✓
LLMBar (Zeng et al., 2024)	Instruction Following	419	-	Categorical	✓	✓	✗

Table 2: Overview of the main features of the datasets considered in the study. Note that 'Size' refers to the number of annotated samples, not to the total number of human annotations. '# Annot.' refers to the number of available individual annotations, if any, which we use to estimate the human upper bound. Note that datasets with only a single annotation per sample, or which only report the average over multiple annotations are not included in '# Annot.'. Information on possible data leakage was retrieved from Balloccu et al. (2024).

Topical Chat and Persona Chat (Mehri and Eskenazi, 2020). These datasets contain human judgments on the quality of machine- and human-generated responses based on the provided dialogue context. The annotated dialogues were selected from Topical Chat (Gopalakrishnan et al., 2019)—a dataset collecting human-human conversations on provided facts—and Persona Chat (Zhang et al., 2018), which contains human-human persona-conditioned conversations. Each response is evaluated on 6 attributes: Understandable, Natural, Maintains Context, Interesting/Engaging, Uses Knowledge, and Overall Quality.

ROSCOE (Golovneva et al., 2023). collect human judgments assessing the quality of GPT-3's reasonings. The output reasonings are elicited by inputting GPT-3 with questions selected from 4 commonly used reasoning datasets, i.e., CosmosQA (Huang et al., 2019), DROP (Dua et al., 2019), e-SNLI (Camburu et al., 2018) and GSM8K (Cobbe et al., 2021). While ROSCOE provides annotations on each step of the reasoning trace, here we only consider the global judgments over the whole reasoning.

QAGS (Wang et al., 2020). QAGS consists of annotations judging the factual consistency of one-sentence model-generated summaries of news articles. The gold-standard summaries and articles are collected from CNN/DailyMail (Hermann et al., 2015) and XSUM (Narayan et al., 2018).

Medical-safety (Abercrombie and Rieser, 2022). This dataset consists of 3701 pairs of medical queries (collected from a subreddit on medical advice) and both machine-generated and human-generated answers. Queries were classified by human annotators according to their severity (from 'Not medical' to 'Serious', with 'Serious' indicating that emergency care would be required) and answers were categorised based on their risk level (from 'Non-medical' to 'Diagnosis/Treatment').

DICES (Aroyo et al., 2023). The DICES datasets consist of a series of machine-generated responses whose safety is judged based on the previous conversation turns (context). While the original dataset provides fine-grained annotations with answers to questions targeting specific aspects of safety, here we only consider the 'overall' categorisation comprehensive of all aspects. In `DICES 990` safety is judged by crowdsourced annotators, whereas in `DICES 350` both expert and crowdsourced annotations are provided.

Inferential strategies (Mondorf and Plank, 2024). This dataset contains annotations on the logical validity of reasoning steps that models—in this case, `Llama-2-chat-hf3` (Touvron et al., 2023), `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023a) and `Zephyr-7b-beta` (Tunstall et al., 2023)—generate when prompted to solve problems of propositional logic. Binary labels are assigned to each response, indicating whether the rationale provided by the model is sound (True) or not (False). Each model is assessed on 12 problems of propositional logic across 5 random seeds, resulting in a total of 60 responses per model.

Switchboard and Dailydialog (Wallbridge et al., 2022). Switchboard includes acceptability judgments collected using stimuli from the Switchboard Telephone Corpus (Godfrey et al., 1992). More specifically, the judgments refer to how plausible it is that a specific response belongs to a telephonic dialogue. The same kind of judgments are provided for Dailydialog, which collects written dialogues intended to mimic conversations that could happen in real life.

Recipe-generation (Stein et al., 2023). This dataset contains human annotations assessing the quality of machine-generated recipes based on 6 attributes: grammar, fluency, verbosity, structure, success, overall.

NewsRoom (Grusky et al., 2018). This dataset includes human judgments on the quality of system-generated summaries of news articles. More specifically, annotators evaluated summaries across two semantic dimensions (informativeness and relevancy) and two syntactic dimensions (fluency and coherence).

SummEval and G-Eval (Fabbri et al., 2021; Liu et al., 2023). These datasets include summaries generated by multiple recent summarisation models trained on the CNN/DailyMail dataset (Hermann et al., 2015). Summaries are annotated by both expert judges and crowdsourced workers on 4 dimensions: coherence, consistency, fluency, relevance.

WMT 2020 En-De and Zh-En (Freitag et al., 2021). These datasets are a re-annotated version of the English-to-German and Chinese-to-English test sets taken from the WMT 2020 news translation task. The annotation was carried out by raters who are professional translators and native speakers of the target language using a Scalar Quality Metric (SQM) evaluation on a 0–6 rating scale.

WMT 2023 En-De and Zh-En (Kocmi et al., 2023). These datasets are the English-to-German and Chinese-to-English test sets taken from the General Machine Translation Task organised as part of the 2023 Conference on Machine Translation (WMT). In contrast to previous editions, the evaluation of translation quality was conducted by a professional or semi-professional annotator pool rather than utilising annotations from MTurk. Annotators were asked to provide a score between 0 and 100 on a sliding scale.

	Dataset	Krippendorff's α
Categorical	Topical Chat	0.08
	QAGS	0.49
	DICES-990	0.14
	DICES-350-crowdsourced	0.16
	Persona Chat	0.33
	Inferential strategies	1.0
Graded	Dailydialog	0.59
	Switchboard	0.57
	Persona Chat	0.33
	Topical Chat	0.08
	Recipe-generation	0.41
	NewsRoom	0.11
	WMT 2020 En-De	0.5
	WMT 2020 Zh-En	0.09

Table 3: Inter-rater agreement for datasets with multiple human annotations. Datasets in blue concern human-generated language, while those in red concern model-generated text.

B The JUDGE-BENCH Data Schema

To facilitate extending our benchmark, we adopt a shared schema used to pre-process all datasets. Our publicly available code base includes an example[2] of this format as well as instructions on how to verify that newly added datasets comply with it.

The Json-based JUDGE-BENCH data schema ensures that the following fields are included for each dataset:

- `dataset`: the name of the dataset;
- `dataset_url`: the URL where the original dataset can be downloaded, as opensourced by their creators;
- `annotations`: an overview of the properties annotated for each dataset, along with information on how they are measured and prompt-like instructions similar to those originally given to the human annotators (when applicable);
- `instances`: dataset instances including the piece of text to be judged, aggregated human

[2] `https://github.com/dmg-illc/JUDGE-BENCH/blob/master/data/example.json`

judgments and, when available, individual human annotations.

We note that, while we do not systematically explore inter-annotator variations at the instance level, the data schema we adopt allows for conducting this type of analysis in future work.

C Upper Bound Estimation for Model Correlations

Whenever multiple human annotations were publicly available for a property (see Table 3 for inter-annotator agreement scores), we computed upper-bound estimates for the correlations achievable by models. The intuition behind these estimates, borrowed from neuroscience (Nili et al., 2014), is that the maximum correlation a model can achieve with aggregated human responses is bounded by the average correlation between single-participant responses and the aggregated responses across participants. We applied a similar logic to the human judgments used in the present study and combined it with a bootstrapping approach. For each annotated property, we bootstrapped single-participant responses by sampling 1000 times from the available human responses, excluding data points where a single annotation was available. Next, we computed the alignment between each of the bootstrapped-participant arrays and the array of aggregated responses. Alignment was computed as Spearman's correlation for graded judgments and Cohen's kappa for categorical judgments. Finally, we estimated the upper bound as the average of the 1000 alignment measures. In cases where alignment between bootstrapped and aggregated responses could not be computed—because the variance of the bootstrapped responses was null—values were replaced with an average of the 'non-nan' correlations.

We emphasise that these upper bounds are estimates and, as such, are subject to errors. Therefore, it may happen that model performance exceeds these upper bounds.

D Properties with Graded Judgments

In Figure 3, we display results for a set of graded properties annotated in one or more of the datasets we consider. The properties are defined as follows:

- *Acceptability* refers to whether it is plausible or not that a response belongs to a telephonic dialogue and was annotated in Swithboard and Dailydialog;
- *Coherence* was annotated for summaries and model-generated reasonings as part of the datasets NewsRoom, ROSCOE, and SummEval;
- *Consistency* refers to the alignment between facts described in a summary and in its source text, and was annotated in SummEval;
- *Engaging* indicates whether a response generated in the context of a dialogue is dull or interesting and was annotated in TopicalChat and PersonaChat;
- *Fluency* measures whether a piece of text is grammatically correct and well-formatted, and was annotated in NewsRoom, SummEval and Recipe-generation;
- *Informativeness* refers to the extent to which a summary captures the key points of the full text, and was annotated for summaries as part of the NewsRoom dataset;
- *Relevance* refers to whether a summary selects important information as opposed to including redundancies, and was annotated for NewsRoom and SummEval;
- *Verbosity* indicates whether a generated recipe is concise and avoids unnecessary repetitions, and was annotated in Recipe-generation.

E Inference Details

All open-model checkpoints were obtained using the HuggingFace pipeline and we access all proprietary models using their corresponding API libraries. The proprietary models were accessed from 06-06-2024 to 13-06-2024, for standard prompting and from 09-10-2024 to 13-12-2024, for CoT prompting. We obtain the model responses using greedy decoding, which we operationalise for the proprietary models by setting the temperature parameter to 0. We allow open models to generate a maximum of 25 new tokens and proprietary models to generate a maximum of 5 new tokens. For CoT prompting, we allow for a maximum of 1000 new tokens.

We leverage Nvidia A100 (80 GB) GPUs for a total of 321 compute hours. The cost of running experiments using Gemini-1.5-flash was €30.31, while the cost of experiments using GPT-4o was approximately $565.

F Valid Response Rates

Table 5 reports the rate of valid responses for each model and dataset. Valid response rates are summarised per model and dataset in Figures 5 and 6.

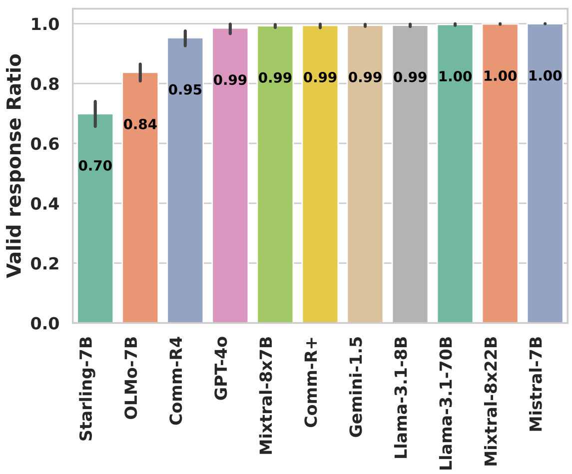

Figure 5: Valid response rate per model.

G More Details on Toxicity and Safety Evaluation

For the Medical-safety dataset, models often refused to answer. Instead they tended to generate explanations, copy what they had in the prompt, or tried to be generally helpful because they saw that it was a medical issue. Since we take a random answer when no answer could be detected, this contributes to lower the results obtained on this task. Scores for the DICES dataset were also low, even though the valid response rate was high, because in this case there is the 'Unsure' option, which (along with 'Unsafe') models preferred over calling anything 'Safe'. For ToxicChat, models performed reasonably well.

H Additional Results

In Table 6 we report human-model alignment scores per dataset for all models tested, thus complementing Table 1 in the paper.

Chain-of-Thought Prompts. For the results with CoT prompting, we use the same original instructions used to gather human judgments as prompts for the model but adapt the additional guidelines to emphasise multi-step reasoning rather than constrain the models' output. Specifically, we append the original instructions with the following additional guideline: '*Always end your answer with either {} regarding the entire context. Let's think step by step.*', in which *{}* is replaced with an enumeration of all possible answer labels formatted as '*Therefore, {label A} is correct, or therefore, {label B} is correct, or therefore [...].*'. This also allows for automatically extracting the final answers from model responses during evaluation. In this study, we evaluate nine models and exclude Mixtral-8x22B and Comm-R+ due to computational constraints. For the CoLa-grammar dataset, we obtain GPT-4o responses only for ten percent of its instances (that are randomly sampled) to address the slow processing times and rate limitations. While CoT prompting leads to improved agreement scores and correlations when used with some models for certain datasets (see Table 7), its overall effectiveness compared to the results obtained using standard prompts without CoT (see Table 6) is inconsistent.

Prompt Paraphrases. We experiment with paraphrased prompts for three datasets that models struggle with: DICES-350-expert, WMT 2023 En-De, and WMT 2023 Zh-En. The paraphrase for dices-350-expert elaborates on the concept of safety, compared to its short original prompt, whereas the paraphrases for the WMT datasets are more concise regarding what comprises a good translation compared to the original. We do not observe consistent improvements when using paraphrased prompts compared to the original prompts (Table 4).

Few-shot Prompts. For the three datasets above—DICES-350-expert, WMT 2023 En-De, and WMT 2023 Zh-En—we also experiment with few-shot prompts (Table 4), where we provide the model with 6 examples for DICES-350-expert, 3 of safe conversations and 3 of unsafe conversations, and 4 examples for each WMT 2023 dataset, 2 of high-scoring translations and 2 of low-scoring translations. Using few-shot prompts does not improve correlations for dices-350-expert. On the WMT 2023 datasets, we observe higher correlations for Llama 3.1 8B but very moderate or no improvements on the other two models. Given that these improvements are inconsistent across datasets, we did not scale up the experiments to all 20 datasets and 11 models.

	Prompt	Llama 3.1 8B	Llama 3.1 70B	Mixtral-8x7B
DICES-350-expert	Original	0.01	-0.13	-0.11
	CoT	-0.07	-0.26	-0.02
	Few-shot	0.01	-0.22	-0.01
	Paraphrase	-0.13	-0.36	-0.09
WMT 2023 En-De	Original	0.08 (1)	0.14 (1)	0.17 (1)
	CoT	0.18 (1)	0.16 (1)	0.20 (1)
	Few-shot	0.19 (1)	0.20 (1)	0.20 (1)
	Paraphrase	0.01 ±0.09 (3)	0.08 ± 0.12 (3)	0.14 ±0.05 (3)
WMT 2023 Zh-En	Original	0.02 (1)	0.14 (1)	0.15 (1)
	CoT	0.13 (1)	0.13 (1)	0.16 (1)
	Few-shot	0.15 (1)	0.14 (1)	0.16 (1)
	Paraphrase	0.08 ±0.04 (3)	0.09 ±0.07 (2)	0.13 ±0.03 (3)

Table 4: Cohen's kappa for DICES-350-expert and Spearman's correlation for two WMT 2023 datasets, comparing the original prompt and CoT prompt to few-shot prompts and prompt paraphrases for a selection of models. For datasets with more than one paraphrased prompt, we report the average and standard deviation across paraphrases. For Spearman's correlations, we report the number of significant correlations ($p < 0.05$) for each model and dataset in brackets.

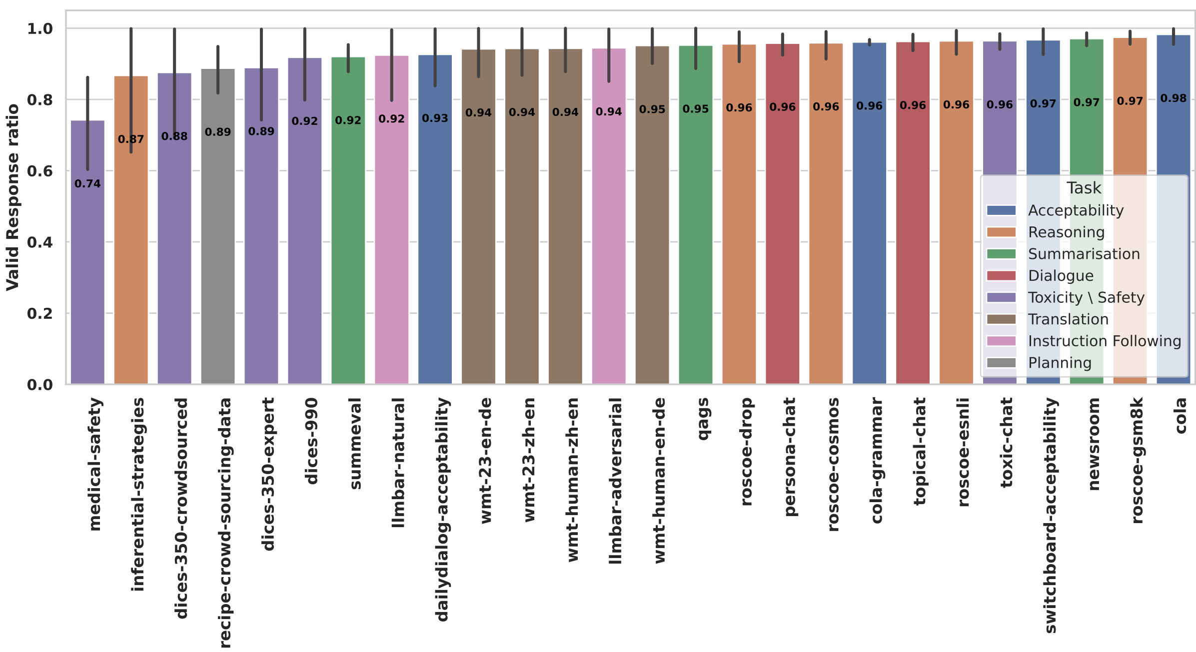

Figure 6: Average ratios of valid responses across datasets over the 11 models we tested.

Type	Dataset (#Subtasks)	GPT-4o	Llama-3.1-70B	Mixtral-8x22B	Gemini-1.5	Mixtral-8x7B	Comm-R+	Comm-R4	Llama-3.1-8B	Mistral-7B	Starling-7B	OLMo-7B
Categorical Annotations	CoLa (1)	1.0	1.0	1.0	1.0	0.98	1.0	1.0	1.0	1.0	0.85	0.98
	CoLa-grammar (63)	1.0	1.0	1.0	1.0	1.0	1.0	1.0±0.01	1.0	1.0	0.71±0.15	0.87±0.11
	LLMBar-natural (1)	1.0	1.0	1.0	1.0	0.95	1.0	0.95	1.0	1.0	0.33	0.94
	LLMBar-adversarial (1)	1.0	1.0	1.0	1.0	0.97	1.0	0.96	1.0	1.0	0.48	0.98
	ToxicChat (2)	1.0	1.0	1.0	0.99	0.96±0.06	0.99	0.91±0.11	0.98	1.0	0.86±0.02	0.92±0.08
	Persona Chat (2)	1.0	1.0	1.0	1.0	0.98±0.02	1.0	0.89±0.15	1.0	1.0	0.96±0.01	0.58±0.24
	Topical Chat (2)	1.0	1.0	1.0	1.0	1.0	1.0	0.99±0.01	1.0	1.0	0.7±0.12	0.77±0.24
	ROSCOE-GSM8K (2)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.92±0.01	0.8±0.23
	ROSCOE-eSNLI (2)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6±0.33	0.73±0.18
	DICES-990 (1)	1.0	1.0	1.0	0.99	0.98	1.0	1.0	1.0	1.0	0.77	0.35
	Inferential strategies (1)	1.0	1.0	1.0	1.0	0.99	0.97	1.0	1.0	1.0	0.05	0.53
	ROSCOE-CosmosQA (2)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.49±0.45	0.75±0.19
	QAGS (1)	1.0	1.0	1.0	0.97	1.0	1.0	1.0	1.0	1.0	0.73	0.78
	Medical-safety (2)	0.35±0.37	0.96±0.02	0.97±0.04	0.97±0.04	0.85±0.1	0.78±0.31	0.33±0.47	0.89±0.11	1.0	0.22±0.08	0.85±0.19
	DICES-350-expert (1)	1.0	1.0	1.0	0.99	1.0	0.98	0.99	1.0	1.0	0.55	0.27
	DICES-350-crowdsourced (1)	1.0	1.0	1.0	0.99	0.99	0.98	1.0	1.0	1.0	0.51	0.16
	ROSCOE-DROP (2)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.51±0.51	0.68±0.26
Graded Annotations	Dailydialog (1)	1.0	1.0	1.0	0.99	1.0	1.0	0.69	1.0	1.0	0.89	0.62
	Switchboard (1)	1.0	1.0	1.0	1.0	0.99	1.0	0.93	1.0	1.0	0.95	0.77
	Persona Chat (4)	1.0	1.0	1.0	1.0	1.0	1.0	0.97±0.03	1.0	1.0	0.71±0.27	0.92±0.15
	Topical Chat (4)	1.0	1.0	1.0	1.0	1.0	1.0	0.99±0.01	1.0	1.0	0.75±0.1	0.91±0.07
	Recipe-generation (6)	1.0	1.0	1.0	1.0	1.0±0.01	1.0	0.67±0.2	1.0	1.0	0.11±0.16	0.98±0.01
	ROSCOE-CosmosQA (2)	1.0	1.0	1.0	1.0	1.0	1.0	0.99±0.01	1.0	1.0	0.97±0.01	0.89
	ROSCOE-DROP (2)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99±0.01	0.85±0.11
	ROSCOE-eSNLI (2)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.89
	ROSCOE-GSM8K (2)	1.0	1.0	1.0	1.0	1.0	1.0	0.98	1.0	1.0	0.84±0.02	0.91±0.06
	NewsRoom (4)	1.0	1.0	0.98±0.01	0.99	1.0	1.0	1.0	1.0	1.0	0.89±0.1	0.83±0.04
	SummEval (4)	0.87±0.13	0.94±0.06	1.0	0.9±0.06	1.0	1.0	0.72±0.3	0.94±0.08	1.0	0.96±0.04	0.79±0.05
	WMT 2020 En-De (1)	1.0	1.0	1.0	0.99	0.87	1.0	1.0	1.0	1.0	0.85	0.75
	WMT 2020 Zh-En (1)	1.0	1.0	1.0	1.0	0.87	1.0	1.0	1.0	1.0	0.81	0.7
	WMT 2023 En-De (1)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.79	0.58
	WMT 2023 Zh-En (1)	1.0	1.0	1.0	1.0	0.99	1.0	1.0	1.0	1.0	0.78	0.61

Table 5: Ratios of valid responses per dataset for all models we evaluate.

Type	Dataset (# properties judged)	GPT-4o	Llama-3.1-70B	Mixtral-8x22B	Gemini-1.5	Mixtral-8x7B	Comm-R+	Comm-R4	Llama-3.1-8B	Mistral-7B	Starling-7B	OLMo-7B
Categorical Annotations	CoLa (1)	0.34	0.46	0.54	0.45	0.55	0.12	0.01	0.42	0.43	0.45	0.42
	CoLa-grammar (63)	0.47 ±0.22	0.28 ±0.24	0.28 ±0.23	0.26 ±0.24	0.21 ±0.18	0.13 ±0.14	0.08 ±0.1	0.1 ±0.14	0.09 ±0.13	0.07 ±0.08	0.04 ±0.06
	LLMBar-natural (1)	0.84	0.8	0.72	0.79	0.54	0.56	0.59	0.57	0.3	0.28	0.24
	LLMBar-adversarial (1)	0.58	0.46	0.2	0.29	0.06	0.11	-0.2	-0.18	-0.2	-0.12	-0.1
	ToxicChat (2)	0.49 ±0.36	0.41 ±0.26	0.45 ±0.27	0.45 ±0.35	0.36 ±0.12	0.28 ±0.35	0.2 ±0.21	0.34 ±0.29	0.45 ±0.18	0.27 ±0.26	0.3 ±0.13
	Persona Chat (2)	0.24 ±0.34	0.24 ±0.33	0.58 ±0.59	-0.03 ±0.04	0.54 ±0.65	0.48 ±0.74	0.01 ±0.01	0.5 ±0.7	0.47 ±0.75	-0.03 ±0.04	0.02 ±0.03
	Topical Chat (2)	0.05 ±0.07	-0.02 ±0.02	-0.03 ±0.04	-0.03 ±0.04	0.02 ±0.03	0.01 ±0.02	0.01 ±0.01	0.57 ±0.61	-0.03 ±0.05	0.04 ±0.06	0.03 ±0.04
	ROSCOE-GSM8K (2)	0.59 ±0.35	0.64 ±0.27	0.62 ±0.38	0.6 ±0.24	0.58 ±0.36	0.0	0.21 ±0.03	0.36 ±0.31	0.47 ±0.34	-0.03 ±0.01	-0.01 ±0.02
	ROSCOE-eSNLI (2)	0.29 ±0.06	0.38 ±0.08	0.13 ±0.13	0.11 ±0.18	0.1 ±0.11	0.03 ±0.05	-0.01 ±0.01	0.14 ±0.2	0.02 ±0.09	0.01 ±0.07	-0.04 ±0.09
	ROSCOE-DROP (2)	0.29 ±0.08	0.27 ±0.07	0.2 ±0.12	0.08 ±0.05	0.13 ±0.21	0.03 ±0.04	0.02 ±0.07	0.02 ±0.02	0.09 ±0.08	0.01 ±0.03	0.0 ±0.01
	ROSCOE-CosmosQA (2)	0.16 ±0.07	0.25 ±0.02	0.09 ±0.17	0.14 ±0.17	0.19 ±0.05	-0.03 ±0.01	-0.01 ±0.02	0.08 ±0.11	0.29 ±0.03	0.03	-0.18
	QAGS (1)	0.72	0.7	0.66	0.65	0.68	0.13	0.33	0.58	0.43	0.02	0.11
	Medical-safety (2)	0.01 ±0.03	-0.03 ±0.06	-0.02 ±0.09	-0.03 ±0.08	0.0 ±0.06	0.01 ±0.02	0.01 ±0.01	0.01	-0.03 ±0.12	0.0 ±0.02	-0.02 ±0.07
	DICES-990 (1)	-0.24	-0.17	-0.16	-0.12	-0.2	-0.09	-0.02	-0.11	-0.12	-0.05	0.0
	DICES-350-expert (1)	-0.2	-0.13	-0.15	-0.03	-0.11	0.01	0.01	0.01	0.01	0.01	-0.06
	DICES-350-crowdsourced (1)	-0.22	-0.18	-0.08	-0.02	-0.11	-0.08	0.01	-0.05	-0.04	0.01	-0.03
	Inferential strategies (1)	0.42	0.4	0.02	0.22	0.06	-0.02	-0.12	0.13	0.01	0.01	0.04
Graded Annotations	Dailydialog (1)	0.69 (1)	0.6 (1)	0.55 (1)	0.63 (1)	0.63 (1)	0.52 (1)	0.23 (1)	0.61 (1)	0.48 (1)	0.09 (0)	0.07 (0)
	Switchboard (1)	0.66 (1)	0.45 (1)	0.63 (1)	0.59 (1)	0.56 (1)	0.36 (1)	0.53 (1)	0.28 (1)	0.52 (1)	0.13 (0)	0.3 (1)
	Persona Chat (4)	0.22 ±0.11 (2)	-0.02 ±0.2 (0)	0.16 ±0.1 (1)	0.1 ±0.09 (0)	0.02 ±0.15 (0)	0.07 ±0.13 (0)	0.05 ±0.2 (0)	-0.02 ±0.14 (0)	-0.09 ±0.17 (1)	0.03 ±0.13 (0)	-0.06 ±0.14 (0)
	Topical Chat (4)	0.26 ±0.03 (2)	0.28 ±0.1 (2)	0.13 ±0.04 (0)	0.17 ±0.12 (1)	0.21 ±0.18 (1)	0.14 ±0.05 (0)	0.07 ±0.07 (0)	0.15 ±0.13 (0)	0.29 ±0.11 (3)	0.14 ±0.16 (1)	0.08 ±0.21 (1)
	Recipe-generation (6)	0.78 ±0.05 (6)	0.66 ±0.07 (6)	0.6 ±0.15 (6)	0.67 ±0.09 (5)	0.57 ±0.24 (5)	0.32 ±0.28 (5)	0.06 ±0.26 (3)	0.34 ±0.09 (5)	0.28 ±0.08 (4)	0.04 ±0.17 (1)	0.1 ±0.08 (0)
	ROSCOE-GSM8K (2)	0.82 ±0.12 (2)	0.83 ±0.11 (2)	0.81 ±0.14 (2)	0.81 ±0.12 (2)	0.79 ±0.13 (2)	0.68 ±0.2 (2)	0.7 ±0.08 (2)	0.76 ±0.15 (2)	0.63 ±0.18 (2)	0.46 ±0.13 (2)	0.1 ±0.07 (1)
	ROSCOE-eSNLI (2)	0.49 ±0.24 (2)	0.4 ±0.16 (2)	0.38 ±0.17 (2)	0.35 ±0.21 (2)	0.32 ±0.12 (2)	0.09 ±0.08 (0)	0.28 ±0.21 (1)	0.19 ±0.16 (1)	0.32 ±0.12 (2)	0.11 ±0.06 (0)	0.11 ±0.17 (1)
	ROSCOE-DROP (2)	0.57 ±0.22 (2)	0.59 ±0.16 (2)	0.44 ±0.15 (2)	0.44 ±0.13 (2)	0.32 ±0.12 (2)	0.21 ±0.22 (1)	0.37 ±0.18 (2)	0.23 ±0.1 (2)	0.22 ±0.22 (1)	0.16 ±0.17 (1)	0.15 ±0.21 (1)
	ROSCOE-CosmosQA (2)	0.57 ±0.18 (2)	0.55 ±0.18 (2)	0.51 ±0.16 (2)	0.57 ±0.17 (2)	0.53 ±0.21 (2)	0.33 ±0.25 (2)	0.48 ±0.17 (2)	0.44 ±0.26 (2)	0.57 ±0.2 (2)	0.13 ±0.04 (1)	0.49 ±0.24 (2)
	NewsRoom (4)	0.59 ±0.02 (4)	0.59 ±0.03 (4)	0.44 ±0.05 (4)	0.55 ±0.03 (4)	0.5 ±0.07 (4)	0.36 ±0.06 (4)	0.16 ±0.05 (4)	0.45 ±0.04 (4)	0.26 ±0.06 (4)	0.21 ±0.08 (4)	-0.01 ±0.04 (0)
	SummEval (4)	0.35 ±0.06 (4)	0.44 ±0.14 (4)	0.54 ±0.08 (4)	0.38 ±0.02 (4)	0.48 ±0.02 (4)	0.19 ±0.06 (4)	0.13 ±0.06 (4)	0.29 ±0.09 (4)	0.4 ±0.12 (4)	0.15 ±0.05 (4)	0.06 ±0.02 (2)
	WMT 2020 En-De (1)	0.63 (1)	0.37 (1)	0.51 (1)	0.46 (1)	0.2 (1)	0.42 (1)	0.15 (1)	0.11 (1)	0.36 (1)	0.15 (1)	-0.03 (1)
	WMT 2020 Zh-En (1)	0.54 (1)	0.39 (1)	0.48 (1)	0.41 (1)	0.25 (1)	0.42 (1)	0.15 (1)	0.14 (1)	0.39 (1)	0.15 (1)	0.01 (0)
	WMT 2023 En-De (1)	0.22 (1)	0.14 (1)	0.23 (1)	0.16 (1)	0.17 (1)	0.22 (1)	0.19 (1)	0.08 (1)	0.18 (1)	-0.09 (1)	-0.05 (1)
	WMT 2023 Zh-En (1)	0.17 (1)	0.14 (1)	0.19 (1)	0.14 (1)	0.15 (1)	0.15 (1)	0.14 (1)	0.02 (1)	0.15 (1)	0.01 (0)	0.01 (0)

Table 6: Scores per dataset for all models we evaluate: Cohen's kappa for categorical annotations and Spearman's correlation for graded annotations. For Spearman's correlations, we report the number of significant correlations ($p < 0.05$) for each model and dataset in brackets. Datasets in blue concern human-generated language while those in red concern model-generated text.

Type	Dataset (#properties judged)	GPT-4o	Llama-3.1-70B	Gemini-1.5	Mixtral-8x7B	Comm-R4	Llama-3.1-8B	Mistral-7B	Starling-7B	OLMo-7B
Categorical Annotations	CoLa (1)	0.35	0.41	0.45	0.47	0.3	0.35	0.51	0.39	0.26
	CoLa-grammar (63)	-0.04 ±0.06	0.35 ±0.25	0.33 ±0.23	0.21 ±0.16	0.05 ±0.09	0.24 ±0.21	0.19 ±0.19	0.16 ±0.16	0.04 ±0.06
	LLMBar-natural (1)	0.86	0.86	0.71	0.62	0.37	0.55	0.56	0.46	0.21
	LLMBar-adversarial (1)	0.67	0.92	0.32	-0.07	-0.25	-0.3	-0.29	-0.25	-0.05
	ToxicChat (2)	0.42 ±0.1	0.37 ±0.03	0.41 ±0.36	0.33 ±0.21	0.33 ±0.26	0.22 ±0.03	0.41 ±0.07	0.33 ±0.16	0.31 ±0.2
	Persona Chat (2)	0.83 ±0.25	0.13 ±0.19	0.57 ±0.6	0.0 ±0.01	0.47 ±0.75	-0.01 ±0.01	-0.01 ±0.02	-0.03 ±0.05	-0.03 ±0.05
	Topical Chat (2)	0.57 ±0.61	0.09 ±0.13	0.03 ±0.04	-0.02 ±0.03	0.48 ±0.74	-0.0	-0.03 ±0.05	-0.03 ±0.05	-0.03 ±0.05
	ROSCOE-GSM8K (2)	0.29 ±0.77	0.52 ±0.26	0.52 ±0.25	-0.29 ±0.02	-0.24 ±0.34	0.12 ±0.15	0.38 ±0.46	-0.06 ±0.18	-0.04 ±0.03
	ROSCOE-eSNLI (2)	0.05 ±0.16	0.1 ±0.09	-0.01 ±0.01	-0.03 ±0.05	-0.04 ±0.04	-0.04 ±0.04	-0.01 ±0.09	0.06 ±0.17	-0.03 ±0.06
	ROSCOE-DROP (2)	-0.05 ±0.01	0.13 ±0.15	-0.08 ±0.07	-0.11 ±0.15	-0.14 ±0.12	-0.05 ±0.05	-0.07 ±0.09	-0.03	-0.07 ±0.04
	ROSCOE-CosmosQA (2)	-0.29 ±0.06	0.03	-0.26 ±0.09	-0.29 ±0.12	-0.3 ±0.05	-0.11 ±0.16	-0.25 ±0.2	-0.09 ±0.16	-0.23 ±0.12
	QAGS (1)	0.69	0.7	0.66	0.66	0.34	0.58	0.46	0.49	0.07
	Medical-safety (2)	-0.01 ±0.09	-0.02 ±0.08	-0.01 ±0.09	0.03 ±0.07	-0.0 ±0.01	-0.02 ±0.01	-0.02 ±0.07	-0.01 ±0.01	-0.01 ±0.09
	DICES-990 (1)	-0.22	-0.15	-0.16	-0.14	-0.1	-0.16	-0.08	-0.08	-0.12
	DICES-350-expert (1)	-0.3	-0.26	-0.06	-0.02	-0.13	-0.07	0.06	0.01	-0.04
	DICES-350-crowdsourced (1)	-0.26	-0.21	-0.13	-0.02	-0.18	-0.19	0.0	-0.07	0.06
	Inferential strategies (1)	0.47	0.4	0.25	0.13	0.01	0.04	-0.01	0.12	0.02
Graded Annotations	Dailydialog (1)	0.69 (1)	0.62 (1)	0.56 (1)	0.25 (1)	0.42 (1)	0.51 (1)	0.4 (1)	0.34 (1)	0.27 (1)
	Switchboard (1)	0.6 (1)	0.48 (1)	0.53 (1)	0.07 (0)	0.38 (1)	0.17 (0)	0.36 (1)	0.35 (1)	0.08 (0)
	Persona Chat (4)	0.2 ±0.09 (2)	0.09 ±0.2 (1)	0.11 ±0.06 (0)	0.06 ±0.15 (0)	0.17 ±0.21 (1)	-0.04 ±0.22 (1)	0.04 ±0.13 (0)	-0.01 ±0.22 (1)	0.06 ±0.18 (0)
	Topical Chat (4)	0.22 ±0.02 (0)	0.14 ±0.13 (1)	0.11 ±0.1 (0)	0.06 ±0.18 (1)	0.1 ±0.14 (0)	0.16 ±0.17 (1)	0.25 ±0.05 (2)	0.17 ±0.09 (1)	0.08 ±0.13 (0)
	Recipe-generation (6)	0.67 ±0.12 (6)	0.64 ±0.14 (6)	0.65 ±0.09 (6)	0.42 ±0.18 (5)	0.14 ±0.15 (2)	0.31 ±0.15 (2)	0.41 ±0.07 (6)	0.34 ±0.2 (4)	0.09 ±0.1 (0)
	ROSCOE-GSM8K (2)	0.82 ±0.12 (2)	0.81 ±0.12 (2)	0.81 ±0.11 (2)	0.81 ±0.13 (2)	0.49 ±0.13 (2)	0.8 ±0.11 (2)	0.58 ±0.13 (2)	0.64 ±0.15 (2)	0.07 ±0.07 (0)
	ROSCOE-eSNLI (2)	0.49 ±0.29 (2)	0.39 ±0.38 (1)	0.31 ±0.32 (1)	0.31 ±0.09 (2)	0.33 ±0.01 (2)	0.23 ±0.17 (1)	0.17 ±0.03 (1)	0.19 ±0.13 (1)	-0.05 ±0.07 (0)
	ROSCOE-DROP (2)	0.54 ±0.17 (2)	0.55 ±0.19 (2)	0.44 ±0.12 (2)	0.29 ±0.14 (2)	0.29 ±0.24 (1)	0.44 ±0.07 (2)	0.28 ±0.03 (2)	0.17 ±0.15 (1)	-0.04 ±0.02 (0)
	ROSCOE-CosmosQA (2)	0.57 ±0.21 (2)	0.55 ±0.22 (2)	0.56 ±0.11 (2)	0.55 ±0.12 (2)	0.62 ±0.15 (2)	0.38 ±0.06 (2)	0.54 ±0.21 (2)	0.32 ±0.14 (2)	0.3 ±0.14 (1)
	NewsRoom (4)	0.57 ±0.05 (4)	0.53 ±0.03 (4)	0.53 ±0.05 (4)	0.46 ±0.02 (4)	0.19 ±0.06 (4)	0.49 ±0.04 (4)	0.19 ±0.04 (4)	0.22 ±0.12 (3)	0.09 ±0.06 (1)
	SummEval (4)	0.45 ±0.11 (4)	0.48 ±0.16 (4)	0.33 ±0.03 (4)	0.35 ±0.05 (4)	0.17 ±0.05 (4)	0.24 ±0.13 (4)	0.38 ±0.1 (4)	0.24 ±0.11 (4)	0.03 ±0.07 (1)
	WMT 2020 En-De (1)	0.57	0.44 (1)	0.38 (1)	0.39 (1)	0.13 (1)	0.34 (1)	0.33 (1)	0.3 (1)	0.0 (0)
	WMT 2020 Zh-En (1)	0.52 (1)	0.44 (1)	0.44 (1)	0.42 (1)	0.19 (1)	0.36 (1)	0.39 (1)	0.35 (1)	0.13 (1)
	WMT 2023 En-De (1)	0.2 (1)	0.16 (1)	0.18 (1)	0.2 (1)	0.2 (1)	0.18 (1)	0.21 (1)	0.19 (1)	-0.01 (0)
	WMT 2023 Zh-En (1)	0.15 (1)	0.13 (1)	0.15 (1)	0.16 (1)	0.16 (1)	0.13 (1)	0.11 (1)	0.13 (1)	0.06 (0)

Table 7: Scores per dataset for all models we evaluate using CoT prompts: Cohen's kappa for categorical annotations and Spearman's correlation for graded annotations. For Spearman's correlation, we report the number of significant correlations for each model and dataset in brackets. Datasets in blue concern human-generated language while those in red concern model-generated text.