OpenNLGauge: An Explainable Metric for NLG Evaluation with Open-Weights LLMs

Ivan Kartáč and **Mateusz Lango** and **Ondřej Dušek**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czechia
{kartac,lango,odusek}@ufal.mff.cuni.cz

Abstract

Large Language Models (LLMs) have demonstrated great potential as evaluators of NLG systems, allowing for high-quality, reference-free, and multi-aspect assessments. However, existing LLM-based metrics suffer from two major drawbacks: reliance on proprietary models to generate training data or perform evaluations, and a lack of fine-grained, explanatory feedback. We introduce OpenNLGauge, a fully open-source, reference-free NLG evaluation metric that provides accurate explanations based on individual error spans. OpenNLGauge is available as a two-stage ensemble of larger open-weight LLMs, or as a small fine-tuned evaluation model, with confirmed generalizability to unseen tasks, domains and aspects. Our extensive meta-evaluation shows that OpenNLGauge achieves competitive correlation with human judgments, outperforming state-of-the-art models on certain tasks while maintaining full reproducibility and providing explanations more than twice as accurate.

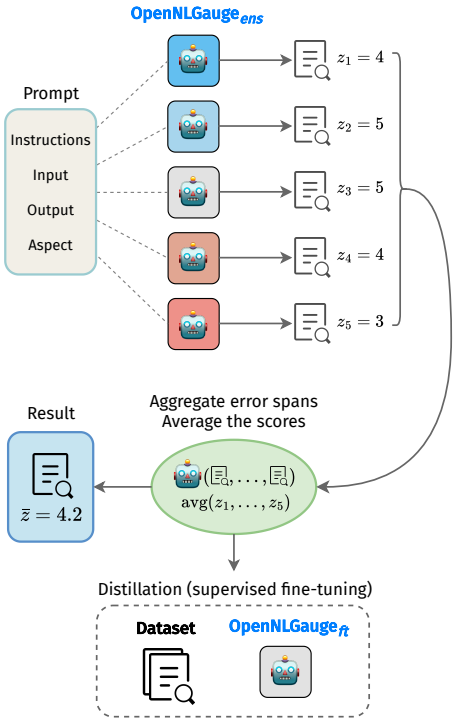

Figure 1: The ensemble metric OpenNLGauge$_{ens}$ and its distilled version OpenNLGauge$_{ft}$.

1 Introduction

Evaluating Natural Language Generation (NLG) systems remains a challenging research problem. Traditional overlap-based metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), are still widely used but exhibit limited correlation with human judgments, particularly when assessing modern NLG systems (Novikova et al., 2017a). With the rise of pre-trained language models, the research community began to shift toward model-based metrics that better capture semantic similarity, yet their performance was still unsatisfactory (Yan et al., 2023; Glushkova et al., 2023).

Recently, Large Language Models (LLMs) have demonstrated remarkable potential in imitating human evaluation of generated text (Jiang et al., 2024; Xu et al., 2023; Hu et al., 2024b). LLM-based metrics are often general enough to evaluate diverse NLG tasks and can provide high evaluation performance without the need for reference texts (Liu et al., 2023). They also offer better adaptability to evaluate specific aspects of generated text and are able to evaluate beyond semantic correctness, taking into account aspects such as factual consistency or relevance (Gu et al., 2024). However, existing LLM-based NLG metrics lack two important features: (1) they are not fully open-source, with most relying on data generated by proprietary LLMs such as GPT-4 (OpenAI, 2023), and (2) they do not provide precise explanations for their evaluations, with most of them producing only overall scores, or explanations limited to a single short comment (see Table 1 for details).

In this paper, we introduce OpenNLGauge –

Proceedings of the 18th International Natural Language Generation Conference, pages 292–337
October 29–November 2, 2025.

a versatile, reference-free metric for NLG tasks that provides precise, error-span-based explanations (Figure 1). Unlike existing LLM-based metrics, OPENLGAUGE is built entirely on open-weight models and does not rely on human-annotated datasets. It supports a wide range of NLG tasks, including data-to-text, summarization or story generation. Moreover, it allows for fine-grained evaluation of specific, customizable aspects of the generated text such as faithfulness, fluency or coherence. It can also evaluate other user-defined aspects appropriate for a given application.

Our contributions are as follows:

- We introduce an effective prompting strategy and a two-stage LLM ensemble to obtain high-quality evaluations of NLG outputs. The proposed OPENLGAUGE$_{ens}$ demonstrates higher correlations with human judgments on some NLG tasks than previously proposed metrics based on proprietary LLMs.
- We collect outputs from over 100 NLG systems on 15 NLG tasks and use OPENLGAUGE$_{ens}$ to annotate them to construct a comprehensive dataset for training open NLG metrics.
- The collected dataset is used to construct OPENLGAUGE$_{ft}$ – a fine-tuned version of the smallest model from the Llama 3.1 family (Grattafiori et al., 2024), which is able to provide accurate explanations and reliable quality assessments of NLG outputs in a more cost-effective way.
- We perform an extensive meta-evaluation of the proposed metrics on seven datasets covering five NLG tasks. The experimental analysis includes human evaluation of explanation quality, comparison of different score aggregation methods, ablation experiments, and evaluation on tasks, domains, and aspects not seen during training.

Our experiments show that OPENLGAUGE achieves higher explanation quality than other metrics trained on data generated by proprietary LLMs, while also delivering strong evaluation performance. On tasks such as summarization, OPENLGAUGE achieves higher correlations with human judgments than metrics based on the strong proprietary GPT-4 model. While the state-of-the-art metric, Themis – which leverages both human-annotated data and data from proprietary LLMs – remains superior in terms of correlations with human judgments, OPENLGAUGE proves highly competitive and even outperforms it on certain tasks, while also providing more detailed error explanations.

Metric	No Ref.	Aspects	Open	Err. Span
G-Eval	✓	✓	✗	✗
Prometheus	✗	✓	✗	✗
Auto-J	✓	✗	✗	✗
InstructScore	✗	✗	✗	✓
TIGERScore	✓	✗	✗	✓
Themis	✓	✓	✗	✗
OPENLGAUGE	✓	✓	✓	✓

Table 1: Comparison of properties in different LLM-based metrics for NLG: reference-free (No Ref.), customizable aspects (Aspects), built exclusively using open-weight LLMs (Open), and producing precise explanations with error-span annotation (Err. Span). The metrics compared are: G-Eval (Liu et al., 2023), Prometheus (Kim et al., 2024b), Auto-J (Li et al., 2024a), InstructScore (Xu et al., 2023), TIGERScore (Jiang et al., 2024), Themis (Hu et al., 2024b) and OPENLGAUGE (this work).

All our experiments were conducted using only quantized open-weight models and two GPUs with 48GB of VRAM each, ensuring that the results can be reproduced in many AI research labs. Our code and data are available on Github.[1]

2 Related Work

Although NLG has traditionally been evaluated using simple word-overlap-based metrics such as BLEU, these are known to have low correlations with human judgments (Novikova et al., 2017a; Reiter, 2018). This improved somewhat with the use of trained models for metrics in the past few years (Yuan et al., 2021; Zhong et al., 2022; Mehri and Eskénazi, 2020b), but the correlations remained moderate. Recently, numerous studies explored the application of LLMs in NLG evaluation. A prominent line of research focuses on leveraging proprietary LLMs such as GPT-4 (OpenAI, 2023), with direct prompting for an overall score (Fu et al., 2024; Kocmi and Federmann, 2023b) or even annotating error spans with categories, which has been explored in machine translation (Kocmi and Federmann, 2023a; Fernandes et al., 2023; Lu et al., 2024). However, using proprietary models is costly and comes with a reproducibility penalty, as some LLM versions become unavailable or are modified in a non-transparent way (Chen et al., 2024). An-

[1] https://github.com/ivankartac/OpeNLGauge

Context: *[SUMMARY OF THE STORY]*

Question: What is Dr. Heywood Floyd's mission on the Clavius Base?

Answer: According to the summary provided[1], Dr. Heywood Floyd's mission on the Clavius Base is to investigate a recently found artifact buried four million years ago. Specifically, his mission is to ride in a Moonbus to the artifact, which is a monolith identical to the one encountered by the man-apes in the summary.[2]

Evaluation aspect: *Conciseness* – Extent to which the answer is concise and to the point

[1] Explanation: This introductory phrase is an error affecting conciseness as it adds unnecessary words to the answer. The response could directly state Floyd's mission without referencing the source (summary), given the context of the question.
Severity: 2

[2] Explanation: This phrase is considered an error affecting conciseness because it provides additional, unnecessary details about the mission. The initial sentence already clearly states Floyd's mission, making the subsequent elaboration redundant and wordy.
Severity: 3

Overall score: Fair (3/5)

Explanation of the score: While the generated answer generally addresses the question, its conciseness is compromised by the inclusion of redundant information and an unnecessary introductory phrase. Removing these elements would enhance the answer's directness and efficiency, potentially elevating the score to 'Good'.

Figure 2: Example error span annotation provided by OPENLGAUGE for the narrative question answering task. The answer to the question, grounded in the story summary, is evaluated for *conciseness*.

other concern is data leakage, which affects results, but cannot be controlled in the case of proprietary models (Balloccu et al., 2024).

Many recent LLM-based metrics, such as Themis (Hu et al., 2024b), TIGERScore (Jiang et al., 2024), InstructScore (Xu et al., 2023) or Prometheus (Kim et al., 2024b), are built on open LLMs, but in fact they all rely on proprietary LLMs such as GPT-4 (OpenAI, 2023) to generate, filter, or annotate their training data. Therefore, some of these metrics can be viewed as sophisticated knowledge distillation methods from proprietary to open-weight LLMs. This retains the reproducibility disadvantage, as reconstructing the metric from scratch or adapting it to a new task requires access to closed-source LLMs.

While several metrics based on open-weight LLMs provide some level of interpretability, this is often limited to a short, free-text review of the evaluated output (Hu et al., 2024b; Kim et al., 2024b). Fine-grained error-span annotation offers several advantages over singular scores and comments: (1) it can be easily processed automatically, allowing its use in post-processing steps or to provide feedback to the model or training algorithms, (2) it offers greater precision and clarity, as errors are associated with a particular part of the output, making it easier to find and correct issues, (3) it is more human-like, resembling the output of human annotation schemes such as MQM (Freitag et al., 2021) or ESA (Kocmi et al., 2024). The alignment with human evaluation allows easier comparison to humans or even use as pre-annotation, helping to accelerate human annotation process. While some metrics do provide annotation on the error level, these tend to use closed LLMs and their scope is generally limited to single NLG tasks, such as the machine translation approaches mentioned above or Kasner and Dusek (2024)'s work on data-to-text generation.

Scoring outputs with a single LLM may introduce bias (Zheng et al., 2023; Panickssery et al., 2024), which can be alleviated by combining multiple LLMs as evaluators. Verga et al. (2024) apply an ensemble of mostly proprietary LLMs, aggregating their scores either by majority voting over binary ratings, or by averaging for ordinal scores. For pairwise evaluation, Li et al. (2024b) propose two methods which combine preferences of multiple LLMs, including an iterative multi-agent discussion.

3 Problem Statement

We formulate the problem of evaluating the output of an NLG system while providing error-based explanations as follows. Given an input x, an output y, and an evaluation aspect a, the task is to return a tuple $\langle z, \{e_1, ..., e_n\}\rangle$, where z is a numeric score assigned to y, and $\{e_1, ..., e_n\}$ represents a set of error annotations. Each error annotation $e_i = \langle s_i, t_i, l_i\rangle$ includes a span of text $s_i \in y$ corresponding to the problematic segment, a textual explanation t_i, and a severity level l_i. Although the term *error* is used throughout this work, it should

be understood in a broader sense as any issue in the text related to the evaluated aspect a. Examples of outputs provided by OPENLGAUGE are presented in Figure 2 and in Appendix L.

4 Open LLM Ensemble as Evaluator

To achieve a high-quality evaluation of NLG outputs, we propose OPENLGAUGE$_{ens}$, a two-stage ensemble of open-weight LLMs. The ensemble consists of n *annotator models*, which perform independent analyses of the provided NLG output, and a *consolidator model*, which is responsible for merging their results and filtering inaccuracies. A high-level overview of the approach is presented in Figure 1.

Although our approach requires multiple LLMs, using the ensemble with a handful of models ($n = 5$ in our experiments) is still less computationally demanding than sampling multiple outputs to obtain statistical estimates, as required by some metrics, e.g. G-Eval ($n = 20$, Liu et al. (2023)). Furthermore, we only use quantized LLMs to limit the computational requirements.

Annotator models The annotator models are open-weight LLMs prompted to identify error spans in the text and to provide detailed explanations and severity levels for each error. This approach facilitates the interpretability of the evaluation process, but also acts as a chain-of-thought mechanism (Wei et al., 2022), helping the model to ground its decisions in a structured reasoning path.

The full annotator model prompt is provided in Appendix F. It contains the description of the evaluated task (e.g., data-to-text), the definition of the evaluated aspect, and a template for the model's response. Since LLMs are known to confuse different evaluation aspects (Hu et al., 2024a), the prompt also contains several rules that instruct the model to remain focused on the specific evaluation aspect, not to make additional assumptions, and to justify any score lower than the maximum by at least one identified error. Inspired by Liu et al. (2023), we also include detailed steps for error identification.

Furthermore, we provide a description of the overall scoring scale, including an explanation of the lowest and highest scores. The scale is presented as categorical (*Unacceptable* < *Poor* < *Fair* < *Good* < *Excellent*), based on the intuition that adjectival categorical scales may be easier for language models to interpret than numerical scales. We use an integer severity scale (1-5) for scoring individual errors in order to avoid confusion with the overall scoring scale.[2]

Finally, the prompt contains the input that was originally used to generate the evaluated output (e.g., the source text for summarization or the input data in data-to-text). Some tasks may involve multiple inputs; for instance, evaluating knowledge-grounded question answering requires knowledge of both the question and the context. In such cases, the context is also presented to the model under separate headers. Although structured formats like JSON might make parsing of the output easier, prompting LLMs to reason within strict structured outputs has been shown to impair their performance (Tam et al., 2024; Beurer-Kellner et al., 2024). Therefore, the models are instructed to produce textual outputs.

Consolidator model The final score of the NLG output is computed as a simple average of the scores provided by the annotator models. However, to meet the requirement of explainable output, the error analyses of multiple annotators need to be unified. This is the task of the consolidator LLM.

This open-weight LLM is instructed to: (1) merge errors detected by multiple models, (2) unify their output format, and (3) clean up the annotations. Specifically, the model is instructed to merge all error annotations that refer to the same issue at approximately the same location in the text, while maintaining annotation granularity. It is also prompted to fix potential deviations from the expected output format. To simplify the cleaning process, error analyses produced by outlier annotator models[3] are filtered out from the input to the consolidator model. The full prompt for the consolidator model is provided in Appendix F.

5 Training a Distilled Model

5.1 Synthetic data generation

To distill knowledge from the ensemble, we collected the outputs of over a hundred NLG systems and applied OPENLGAUGE$_{ens}$ to produce synthetic evaluation data. We include NLG outputs produced on 15 datasets covering five task categories and almost 40 aspects. Table 2 shows the basic statistics of the constructed dataset.

[2] A categorical scale was also initially used for error severity. However, we found that LLMs sometimes confused the error severity scale with the overall score scale.

[3] The annotation is considered to be an outlier if the score provided by the annotator differs from the mean score by at least two standard deviations, and this difference is at least 1.

Task	Src.	Sys.	Asp.	Examples
Summarization	5	39	6	12,070
Data-to-text	4	29	7	7,894
Dialogue	3	36	9	10,074
Story Generation	1	9	6	3,200
Question Answering	2	15	11	4,849

Table 2: Training dataset statistics for OPENLGAUGE$_{ft}$: number of source datasets (Src.), systems (Sys.), evaluation aspects (Asp.) and training examples for each task.

The NLG outputs included cover the following tasks: data-to-text, summarization, question answering, dialogue response generation and story generation. These were chosen to represent a diverse range of tasks, each with unique objectives, input-output relationships and evaluation aspects. Note that the underlying datasets are used only as inputs for NLG systems and do not include human evaluations of any kind.

For each task, we select a set of relevant evaluation aspects with their definitions. We consider two aspects distinct if they are associated with different tasks. For instance, *coherence* in dialogue refers to coherence of the response with respect to the dialogue history, while in summarization it refers to the internal coherence of a summary. We list the datasets and aspects for each task in Appendix A.

To obtain output texts with varying quality and diverse types of errors, we sample outputs from a variety of systems, ranging from rule-based approaches to state-of-the-art LLMs. For older systems, we use pre-generated outputs from existing datasets, while outputs from more recent systems including LLMs are newly generated. An overview of the evaluated systems is shown in Appendix B.

To limit computational requirements for dataset generation, we apply a sampling procedure that ensures data diversity and broad coverage of different NLG systems and aspects while significantly reducing dataset size. For each input, we randomly sample the outputs of N systems, followed by sampling M aspects for each input-output pair. The sampling includes all aspects described in Appendix A and all systems listed in Appendix B. This results in $N \times M$ (input, output, aspect) triples for each input, which are then passed to OPENLGAUGE$_{ens}$ to obtain synthetic annotations. For most tasks, we set $N = 4$ and $M = 3$. This sampling strategy aims for a balanced distribution of inputs, NLG system outputs and evaluation aspects. It also ensures exposure to different outputs for the same input, i.e., it should not prime models to evaluate based solely on patterns in the inputs. Finally, by presenting multiple aspects for the same input-output pair, models are encouraged to learn differences in output quality between different evaluation aspects.

To keep the merged evaluation outputs internally consistent, we remove outliers before merging, as described in Section 4. Table 12 in Appendix D shows the proportions of outliers detected for all LLMs and task categories (3.4% on average).

5.2 Fine-tuning Procedure

We use the dataset described in Section 5.1 for supervised fine-tuning of a specialized LLM evaluator, with an instruction-tuned version of Llama 3.1 8B as the backbone. To avoid training the LLM to predict floating-point scores, we convert them to integers in the range 0-100 and then bin them to the nearest multiple of five. This extends the output space from five to twenty values (instead of 100), which is a trade-off between greater granularity in predictions and manageable task complexity.

As the model is expected to learn the task from training data, we used a simpler prompt template for fine-tuning, with a brief description of the task, the definition of the evaluated aspect, and the input and output to be evaluated (see Appendix F).

6 Experimental Setup

6.1 Ensemble

The ensemble consists of six open-weight LLMs: five annotators and one consolidator. Each selected LLM is distributed under a license that permits at least non-commercial use and allows the model's outputs to be used as training data. At the time of the experiments, these models ranked among the top-performing open-weight LLMs on the Chatbot Arena Leaderboard[4] (Chiang et al., 2024).

The annotator models include the following: Llama 3.1 Nemotron 70B (Wang et al., 2025), Qwen 2.5 72B (Yang et al., 2024), Gemma 2 27B (Team et al., 2024), Command R+ 104B (Cohere For AI, 2024), and Mistral Large 2 123B[5]. We apply Llama 3.3 70B (Dubey et al., 2024) as the consolidator model. To address computational constraints, we use quantized versions available through the Ollama platform.[6] For synthetic data generation, we set the temperature to zero to obtain

[4] https://lmarena.ai/leaderboard
[5] https://mistral.ai/news/mistral-large-2407/
[6] https://ollama.com/

Metric	QAGS-CNN/DM			QAGS-XSUM			Average		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
ROUGE-1	0.338	0.318	0.248	-0.008	-0.049	-0.040	0.165	0.134	0.104
ROUGE-2	0.459	0.418	0.333	0.097	0.083	0.068	0.278	0.250	0.200
ROUGE-L	0.357	0.324	0.254	0.024	-0.011	-0.009	0.190	0.156	0.122
BERTScore	0.576	0.505	0.399	0.024	0.008	0.006	0.300	0.256	0.202
MoverScore	0.414	0.347	0.271	0.054	0.044	0.036	0.234	0.195	0.153
FactCC	0.416	0.484	0.376	0.297	0.259	0.212	0.356	0.371	0.294
QAGS	0.545	-	-	0.175	-	-	0.375	-	-
BARTScore	0.732	0.680	0.555	0.175	0.171	0.139	0.454	0.425	0.347
UniEval	0.682	0.662	0.532	0.461	0.488	0.399	0.572	0.575	0.466
G-Eval (GPT-3.5)	0.477	0.516	0.410	0.211	0.406	0.343	0.344	0.461	0.377
G-Eval (GPT-4)	0.631	0.685	0.591	0.558	0.537	0.472	0.595	0.611	0.532
LLM Evaluation (GPT-3.5)	0.454	0.514	0.417	0.279	0.348	0.295	0.366	0.431	0.356
LLM Evaluation (GPT-4)	0.735	0.746	0.626	0.541	0.528	0.439	0.638	0.637	0.532
Auto-J	0.291	0.238	0.214	0.225	0.214	0.203	0.258	0.226	0.209
TIGERScore	0.574	0.562	0.479	0.424	0.445	0.412	0.499	0.504	0.446
InstructScore	0.287	0.278	0.233	-0.096	-0.134	-0.119	0.095	0.072	0.057
Themis	**0.747**	**0.761**	**0.680**	0.599	0.607	0.546	0.673	0.684	**0.613**
OpenNLGauge$_{ens}$	0.738	0.753	0.627	**0.630**	**0.624**	0.531	**0.684**	**0.689**	0.579
• Command R+ 104B	0.676	0.675	0.617	0.540	0.541	0.515	0.608	0.608	0.566
• Gemma 2 27B	0.579	0.646	0.579	0.592	0.614	**0.563**	0.585	0.630	0.571
• Llama 3.1 Nemotron 70B	0.705	0.733	0.650	0.587	0.586	0.540	0.646	0.659	0.595
• Mistral Large 2 123B	0.658	0.704	0.635	0.577	0.570	0.541	0.617	0.637	0.588
• Qwen 2.5 72B	0.678	0.720	0.635	0.568	0.569	0.526	0.623	0.644	0.581
Llama 3.1 8B	0.275	0.242	0.219	0.218	0.230	0.218	0.247	0.236	0.219
OpenNLGauge$_{ft}$	0.668	0.695	0.584	0.607	0.607	0.524	0.638	0.651	0.554

Table 3: Segment-level Pearson (r), Spearman (ρ) and Kendall (τ) correlations of different metrics for factual consistency on QAGS. The best correlations are highlighted in bold, the second best are underlined.

consistent results. For details on the models, see Appendix C.

6.2 Distillation

To produce OpenNLGauge$_{ft}$, we apply LoRA (Hu et al., 2022) with rank 16 and alpha 32 to fine-tune the instruction-tuned version of Llama 3.1 8B. The model is trained for one epoch with a learning rate of 2e-4, using AdamW optimizer (Loshchilov and Hutter, 2017) with a weight decay of 0.01. We apply a linear learning rate schedule with a warm-up period corresponding to the first 5% of training steps. The batch size is set to 16. This setup enables resource-efficient training that requires only around 20GB of VRAM, completing one training epoch in just six hours on a single A40 GPU.

6.3 Evaluation Datasets

We used seven popular meta-evaluation datasets to assess how our metric correlates with human judgments (for detailed descriptions of these datasets, including details on aggregation of human scores, see Appendix E). The datasets cover the following tasks: summarization – SummEval (Fabbri et al., 2021b), QAGS (Wang et al., 2020a); story generation – HANNA (Chhun et al., 2022); data-to-text – SFRES and SFHOT (Wen et al., 2015), text simplification – Wiki-DA (Alva-Manchego et al., 2021) and dialogue generation – TopicalChat (Gopalakrishnan et al., 2019). For OpenNLGauge$_{ft}$, text simplification is a task unseen during training and TopicalChat contains an unseen aspect (groundedness). For human evaluation of error spans, we used the RotoWire dataset (Thomson and Reiter, 2020) with annotations from a data-to-text task in the basketball domain, a task unseen during training by the evaluated fine-tuned metrics.

Dataset	Llama 3.1	OpeNLGauge$_{ft}$	Δ
QAGS	0.236	0.651	+0.415
SummEval	0.186	0.502	+0.316
TopicalChat	0.309	0.578	+0.269
SFRES/SFHOT	0.108	0.315	+0.207
HANNA	0.150	0.425	+0.275
Wiki-DA	0.405	0.789	+0.384

Table 4: Comparison of Spearman (ρ) correlations of the backbone model (Llama 3.1 8B) and our metric OpenNLGauge$_{ft}$ fine-tuned from the backbone on the dataset described in Section 5.1. For each dataset, the correlations are averaged across all evaluated aspects.

6.4 Baselines

We compare our methods with a variety of commonly used evaluation metrics, including traditional metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) and METEOR (Agarwal and Lavie, 2008), distance-based metrics MoverScore (Zhao et al., 2019) and BERTScore (Zhang et al., 2020a), and trained metrics BARTScore (Yuan et al., 2021) and UniEval (Zhong et al., 2022). We also include the following LLM-based metrics: GPTScore (Fu et al., 2024), G-Eval (Liu et al., 2023), LLM Evaluation (Chiang and Lee, 2023a), Prometheus (Kim et al., 2024b), Auto-J (Li et al., 2024a), InstructScore (Xu et al., 2023), TIGERScore (Jiang et al., 2024) and Themis (Hu et al., 2024b). To measure the improvement of OpenNLGauge$_{ft}$ relative to the base model, we include the instruction-tuned version of Llama 3.1 8B as an additional baseline. The specific metrics reported differ depending on the dataset and the evaluated NLG task.

Additionally, our comparisons include some task-specific and aspect-specific metrics: QAGS (Wang et al., 2020a) and FactCC (Kryscinski et al., 2020) for evaluating factual consistency in summarization, SARI (Xu et al., 2016) and LENS (Maddela et al., 2023) for text simplification, and USR (Mehri and Eskénazi, 2020b) for dialogue response generation tasks. The latter metric has several variants; we use the variant with the best Pearson correlation for each aspect in our comparison.

7 Results

7.1 Score Correlation with Humans

The results for factual consistency on QAGS are presented in Table 3 and the results for other datasets are available in Tables 14–20 in Appendix G. On QAGS, OpenNLGauge$_{ens}$ achieves the highest average performance on both Pearson's r and Spearman's ρ. On Kendall's τ, our method is outperformed by Themis, which can be attributed to different score granularities used by these two methods (see Section G.1 for a discussion). The distilled version of our metric was consistently the third best measure. Notably, it outperformed the metrics based on the proprietary GPT-4 on this task.

On SummEval (Table 14), the best performing metric was Themis, closely followed by OpenNLGauge$_{ens}$. However, note that training data for Themis include almost 62,000 *human-annotated* examples for the summarization task.

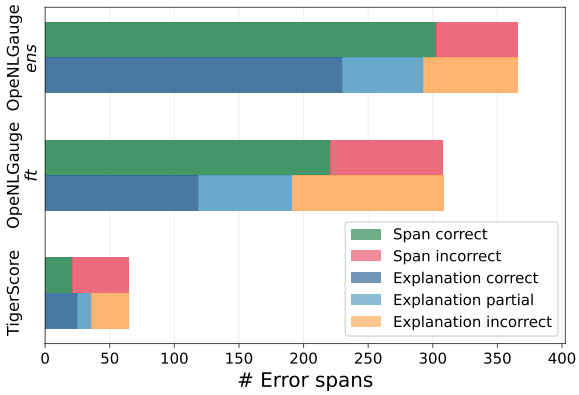

Figure 3: Results of human evaluation of error spans and explanations. Top half of each bar: Error spans marked as correct or incorrect (hallucinated spans, no span provided, or spans without errors). Bottom half: Explanations marked as correct, partial (partially correct or incomplete) or incorrect (not addressing actual errors, vague or incorrect). The differences between TigerScore and OpenNLGauge$_{ens}$ are statistically significant (t-test, $p < 0.05$). See Table 13 for more details.

Comparing our approach to TigerScore, another method that provides a similar level of explainability (error span annotations), we observe 15 p.p. improvements on Spearman's ρ averaged over all aspects. The smaller OpenNLGauge$_{ft}$ outperformed all other fine-tuned LLM-based metrics except Themis, and also surpassed metrics based on prompting GPT-3.5 by a large margin.

On TopicalChat (Table 15), the *LLM Evaluation* metric based on GPT-4 emerged as the strongest evaluator, while OpenNLGauge$_{ens}$ ranked between this method and its GPT-3.5-based version. OpenNLGauge$_{ft}$ achieved 27 p.p. improvement in Spearman's ρ compared to its Llama 3.1 backbone.

In data-to-text, OpenNLGauge$_{ens}$ excelled in evaluating naturalness, while OpenNLGauge$_{ft}$ stands out as the strongest evaluator of informativeness on the SFRES dataset (Table 16). Interestingly, on data-to-text problems, the distilled version of our metric achieved slightly better results on average than our ensemble. A similar situation is observed for story generation (Tables 17–19), where the distilled version obtained better average scores on Spearman's ρ and Kendall's τ, and closely followed OpenNLGauge$_{ens}$ on Pearson's r.

In text simplification (Table 20), our ensemble achieved superior performance across the board, outperforming LENS, a strong baseline metric specialized for this task.

Additionally, Table 4 presents a comparison of

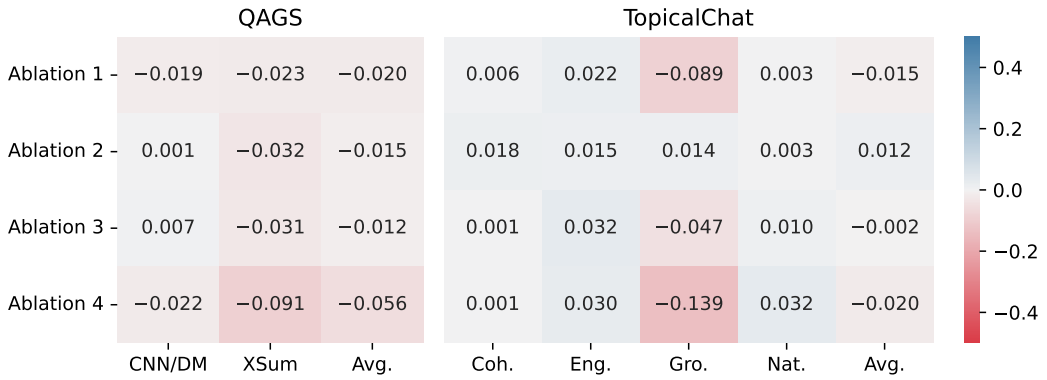

Figure 4: Ablation results on QAGS and TopicalChat for OpenNLGauge$_{ens}$. Plotted values represent differences in Spearman's ρ correlations with human scores between the ensemble with the original prompt and the corresponding ablation. For TopicalChat, **Coh.** = coherence, **Eng.** = engagingness, **Gro.** = groundedness, **Nat.** = naturalness, **Avg.** = average for all aspects.

OpenNLGauge$_{ft}$ with Llama 3.1 8B (instruct), indicating a considerable improvement over the prompted model.

Generalization to unseen tasks Although not originally trained on text simplification, OpenNLGauge$_{ft}$ outperforms all baseline metrics and individual LLMs in averaged Pearson correlation on the Wiki-DA dataset (Table 20) and achieves a particularly high score on meaning preservation.

Generalization to unseen aspects On TopicalChat (Table 15), a large improvement for OpenNLGauge$_{ft}$ over Llama 3.1 8B is observed on groundedness, which is an aspect unseen during training. OpenNLGauge$_{ft}$ also surpassed most fine-tuned metrics and most individual LLMs on this aspect. Moreover, Wiki-DA contains additional unseen aspects (meaning preservation and simplicity) on which OpenNLGauge$_{ft}$ shows considerable improvement over Llama 3.1 8B and outperforms most of the individual larger LLMs.

7.2 Human Evaluation of Error Spans

We performed a small in-house human evaluation study to compare the quality of explanations obtained by OpenNLGauge and TigerScore, another LLM-based metric that also provides error-span annotations (see Table 1). For this purpose, we used a data-to-text task in the basketball domain (Thomson and Reiter, 2020). Five expert annotators evaluated the output of all three systems on 40 instances (a total of 120 outputs and 950 error spans). We asked the annotators to: (1) evaluate provided error spans, marking them as *correctly identified*, *not containing an error*, *hallucinated*, and situations where *no span was provided*; (2) evaluate generated explanations, marking them as *correct*, *partially correct*, *incomplete*, *vague*, *incorrect*, or texts that *do not describe an error*.

To assess reliability of our human annotation, we computed Cohen's κ coefficient (Cohen, 1960) of inter-annotator agreement on 50 error spans with double annotations. We obtained $\kappa = 0.82$ for the evaluation of error spans, and $\kappa = 0.46$ for error explanations.

The results are shown in Figure 3, with more details provided in Table 13 in Appendix G. Both OpenNLGauge$_{ft}$ and OpenNLGauge$_{ens}$ are over twice more accurate than TigerScore at annotating error spans, while finding over ten times more correct error spans. The task of providing accurate error explanations was more difficult for all the approaches evaluated. OpenNLGauge$_{ens}$ achieved the highest performance and was almost twice as accurate as TigerScore and OpenNLGauge$_{ft}$, which achieved similar accuracies.

7.3 Ablation Experiments

Prompt ablations We explore the effect of using different scales for the overall score and error severity: integer scales for both (Ablation 1), integer scale for overall score and categorical scale for severity (Ablation 2), and categorical scale for both (Ablation 3). Recall that OpenNLGauge uses a categorical scale for overall score and integer scale for error severity – see Section 4.

Correlation differences between the full prompt and the ablations are summarized in Figure 4, for detailed results see Appendix H.1. Overall, the

change of scale has little effect on the average correlation of the whole ensemble, but it has a dramatic effect on some individual LLMs and aspects. This illustrates the ability of the ensemble to compensate for weaknesses in individual annotator models.

Finally, we examine the effect of removing the evaluation rules from the prompt (Ablation 4), which has an inconsistent effect on different models/aspects, but on average degrades the ensemble evaluation quality – up to 5.6 p.p.

Ensemble structure We also analyze the effect of ensemble size and the influence of its particular components on the correlations with human scores by recomputing the results for all ensemble combinations. The results are presented in Appendix H.2. For Wiki-DA, performance increases with ensemble size, with the full ensemble being the best combination. For other datasets, there are a few smaller combinations that actually rank higher, but none is consistently better than the full ensemble.

Inter-annotator agreement between LLMs To obtain additional insights into how the individual models of the ensemble diverge in their overall score predictions, we compute several measures of inter-annotator agreement. The results presented in Appendix I indicate only low to moderate agreement for most of the datasets, especially on exact overall scores, which suggests a sufficient diversity for combining outputs of these models into an ensemble.

Error analysis aggregation The consolidator model aggregates error span annotations from multiple LLMs, which could potentially lead to an overall larger number of detected errors. To estimate the extent of over-annotation by both the ensemble and its components, we analyze the number of detected errors for output-aspect pairs rated with maximum score by human annotators. The results presented in Appendix K indeed show some tendency of the ensemble to over-annotate due to error accumulation from the individual models. However, most spans annotated by the ensemble were marked as correct in the experiment in Section 7.2. This could indicate that OpeNLGauge$_{ens}$ finds subtle errors which human annotators overlook, but further analysis of this discrepancy is needed.

Score aggregation Additionally, we compare different methods of aggregating overall scores of individual LLMs to a final score in Appendix J. These results indicate that despite its simplicity, simple averaging is the most effective approach, generally providing the highest correlations with human scores.

8 Summary

In this work, we present OpeNLGauge – a versatile method for evaluating NLG that uses only open-weight models and provides fine-grained explainability. The method provides a much better explanation quality than previous methods and achieves competitive correlations with human judgments.

Limitations

OpeNLGauge is a method for evaluating a variety of NLG tasks. While this paper presents the evaluation on several NLG tasks and the method achieves good performance on unseen aspects, domains and tasks, the actual performance on new NLG tasks is unknown. In particular, the metric has not been tested in a multilingual setting. Moreover, previous research has shown that some LLM-based metrics have a bias towards texts generated by LLMs (Liu et al., 2023).

Acknowledgments

This work was supported by the European Research Council (Grant agreement No. 101039303, NG-NLG) and the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO. It used resources of the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

References

Abhaya Agarwal and Alon Lavie. 2008. Meteor, M-BLEU and M-TER: Evaluation Metrics for High-correlation with Human Rankings of Machine Translation Output. In *Proceedings of the Third Workshop on Statistical Machine Translation, WMT at ACL 2008*, pages 115–118, Columbus, Ohio, USA.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Comput. Linguistics*, 47(4):861–889.

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, EACL 2024 -*

Volume 1: Long Papers, pages 67–93, St. Julian's, Malta.

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the Similarity Function of TextRank for Automated Summarization. *CoRR*, abs/1602.03606.

Luca Beurer-Kellner, Marc Fischer, and Martin T. Vechev. 2024. Guiding LLMs The Right Way: Fast, Non-invasive Constrained Generation. In *Forty-first International Conference on Machine Learning, ICML 2024*, Vienna, Austria.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review*, 6(2).

Mingje Chen, Gerasimos Lampouras, and Andreas Vlachos. 2018. Sheffield at E2e: Structured Prediction Approaches to End-to-end Language Generation. *E2E NLG Challenge System Descriptions*, 85.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020. Logical Natural Language Generation from Open-domain Tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 7929–7942, Online.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022*, pages 5794–5836, Gyeongju, Republic of Korea.

Cheng-Han Chiang and Hung-yi Lee. 2023a. A Closer Look Into Using Large Language Models for Automatic Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore.

Cheng-Han Chiang and Hung-Yi Lee. 2023b. 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.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas 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 *Forty-first International Conference on Machine Learning, ICML 2024*, Vienna, Austria.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling Instruction-finetuned Language Models. *J. Mach. Learn. Res.*, 25:70:1–70:53.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46.

Cohere For AI. 2024. C4ai-Command-R-Plus-08-2024 (Revision Dfda5ab).

DeepSeek-AI, Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y. Wu, Yukun Li, Huazuo Gao, Shirong Ma, Wangding Zeng, Xiao Bi, Zihui Gu, Hanwei Xu, Damai Dai, Kai Dong, Liyue Zhang, Yishi Piao, Zhibin Gou, Zhenda Xie, Zhewen Hao, Bingxuan Wang, Junxiao Song, Deli Chen, Xin Xie, Kang Guan, Yuxiang You, Aixin Liu, Qiushi Du, Wenjun Gao, Xuan Lu, Qinyu Chen, Yaohui Wang, Chengqi Deng, Jiashi Li, Chenggang Zhao, Chong Ruan, Fuli Luo, and Wenfeng Liang. 2024. DeepSeek-Coder-V2: Breaking the Barrier of Closed-source Models in Code Intelligence. *CoRR*, abs/2406.11931.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered Conversational Agents. In *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC*, pages 13042–13054, Canada.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie

Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Comput. Speech Lang.*, 59:123–156.

Henry Elder, Sebastian Gehrmann, Alexander O'Connor, and Qun Liu. 2018. E2E NLG Challenge Submission: Towards Controllable Generation of Diverse Natural Language. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 457–462, Tilburg University, The Netherlands. Association for Computational Linguistics.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021a. ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021b. SummEval: Re-evaluating Summarization Evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, pages 889–898, Melbourne, Australia.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. 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, WMT 2023*, pages 1066–1083, Singapore.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 Bilingual, Bi-directional WebNLG+ Shared Task: Overview and Evaluation Results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76.

Markus Freitag, George F. 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. *Trans. Assoc. Comput. Linguistics*, 9:1460–1474.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as You Desire. In *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*, pages 6556–6576, Mexico City, Mexico.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *CoRR*, abs/2101.00027.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting Summarization Evaluation for Dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 5693–5709, Seattle, WA, United States.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, Sydney, NSW, Australia.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. *CoRR*, abs/1911.12237.

Taisiya Glushkova, Chrysoula Zerva, and André F. T. Martins. 2023. BLEU Meets COMET: Combining Lexical and Neural Metrics Towards Robust Machine Translation Evaluation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 47–58, Tampere, Finland.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph M. Weischedel, and Nanyun Peng. 2020. Content Planning for Neural Story Generation with Aristotelian Rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 4319–4338, Online.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-grounded Open-domain Conversations. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019*, pages 1891–1895, Graz, Austria.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis,

Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

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, NAACL-HLT 2018, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, USA.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A Survey on LLM-as-a-judge. *CoRR*, abs/2411.15594.

Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu. 2020. A Knowledge-enhanced Pretraining Model for Commonsense Story Generation. *Trans. Assoc. Comput. Linguistics*, 8:93–108.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A Benchmark for Evaluating Open-ended Story Generation Metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, pages 6394–6407, Virtual Event.

Çaglar Gülçehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the Unknown Words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Volume 1: Long Papers*, Berlin, Germany.

Qipeng Guo, Zhijing Jin, Ning Dai, Xipeng Qiu, Xiangyang Xue, David Wipf, and Zheng Zhang. 2020. $\surd^2$: A Plan-and-Pretrain Approach for Knowledge Graph-to-Text Generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 100–106, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskénazi, and Jeffrey P. Bigham. 2019. Investigating Evaluation of Open-domain Dialogue Systems With Human Generated Multiple References. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019*, pages 379–391, Stockholm, Sweden.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015*, pages 1693–1701, Montreal, Quebec, Canada.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and

Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024a. Are LLM-based Evaluators Confusing NLG Quality Criteria? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9530–9570, Bangkok, Thailand. Association for Computational Linguistics.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024b. Themis: A Reference-free NLG Evaluation Language Model with Flexibility and Interpretability. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL*, pages 15924–15951, USA.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic Graph-enhanced Coherence Metric for Evaluating Open-domain Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 9230–9240, Online.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2020. Narrative Text Generation with a Latent Discrete Plan. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3637–3650, Online Event.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. TIGERScore: Towards Building Explainable Metric for All Text Generation Tasks. *Trans. Mach. Learn. Res.*, 2024.

Zhengbao Jiang, Yi Mao, Pengcheng He, Graham Neubig, and Weizhu Chen. 2022. OmniTab: Pretraining with Natural and Synthetic Data for Few-shot Table-based Question Answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 932–942, Seattle, United States. Association for Computational Linguistics.

Juraj Juraska, Panagiotis Karagiannis, Kevin Bowden, and Marilyn Walker. 2018. A Deep Ensemble Model with Slot Alignment for Sequence-to-Sequence Natural Language Generation. 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 152–162, New Orleans, Louisiana. Association for Computational Linguistics.

Zdenek Kasner and Ondrej Dusek. 2024. Beyond Traditional Benchmarks: Analyzing Behaviors of Open LLMs on Data-to-text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 12045–12072, Bangkok, Thailand.

Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024a. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 23–35, Mexico City, Mexico. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024b. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Trans. Assoc. Comput. Linguistics*, 6:317–328.

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, WMT 2023*, pages 768–775, Singapore.

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, EAMT 2023*, pages 193–203, Tampere, Finland.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popovic, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL*, pages 1440–1453, USA.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri,

Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. *Preprint*, arXiv:2411.15124.

Guy Lapalme. 2020. RDFjsRealB: A Symbolic Approach for Generating Text from RDF Triples. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 144–153, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 7871–7880, Online.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024a. Generative Judge for Evaluating Alignment. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023. Collaborative Evaluation: Exploring the Synergy of Large Language Models and Humans for Open-ended Generation Evaluation. *CoRR*, abs/2310.19740.

Ruosen Li, Teerth Patel, and Xinya Du. 2024b. PRD: Peer Rank and Discussion Improve Large Language Model based Evaluations. *Trans. Mach. Learn. Res.*, 2024.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017Volume 1: Long Papers*, pages 986–995, Taipei, Taiwan.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ao Liu, Haoyu Dong, Naoaki Okazaki, Shi Han, and Dongmei Zhang. 2022a. PLOG: Table-to-Logic Pretraining for Logical Table-to-Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5531–5546, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022b. TAPEX: Table Pre-training via Learning a Neural SQL Executor. In *The Tenth International Conference on Learning Representations, ICLR 2022*, Virtual Event.

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.

Zhengyuan Liu and Nancy Chen. 2021. Controllable Neural Dialogue Summarization with Personal Named Entity Planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. *CoRR*, abs/1711.05101.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-turn Dialogue Systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015*, pages 285–294, Prague, Czech Republic.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error Analysis Prompting Enables Human-like Translation Evaluation in Large Language Models. In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 8801–8816, Bangkok, Thailand.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A Learnable Evaluation Metric for Text Simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 16383–16408, Toronto, Canada.

Shikib Mehri and Maxine Eskénazi. 2020a. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting*, pages 225–235.

Shikib Mehri and Maxine Eskénazi. 2020b. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 681–707, Online.

Pablo N. Mendes, Max Jakob, and Christian Bizer. 2012. DBpedia: A Multilingual Cross-domain Knowledge Base. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 1813–1817, Istanbul, Turkey.

Simon Mille and Stamatia Dasiopoulou. 2018. FORGe at E2E 2017. In *Proceedings of the E2E NLG Challenge System Descriptions (2018)*.

Sébastien Montella, Betty Fabre, Tanguy Urvoy, Johannes Heinecke, and Lina Maria Rojas-Barahona. 2020. Denoising Pre-training and Data Augmentation Strategies for Enhanced RDF Verbalization with Transformers. *CoRR*, abs/2012.00571.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*, pages 280–290, Berlin, Germany.

Linyong Nan, Lorenzo Jaime Yu Flores, Yilun Zhao, Yixin Liu, Luke Benson, Weijin Zou, and Dragomir Radev. 2022a. R2D2: Robust Data-to-text with Replacement Detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 6903–6917, Abu Dhabi, United Arab Emirates.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryscinski, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir R. Radev. 2022b. FeTaQA: Free-form Table Question Answering. *Trans. Assoc. Comput. Linguistics*, 10:35–49.

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.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017a. Why We Need New Evaluation Metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2017b. The E2E Dataset: New Challenges For End-to-end Generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken*, pages 201–206, Germany.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC*, Canada.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002*, pages 311–318, Philadelphia.

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A Controlled Table-To-text Generation Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 1173–1186, Online.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 5370–5381, Florence, Italy.

Tanya Reinhart. 1980. Conditions for Text Coherence. *Poetics today*, 1(4):161–180.

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Comput. Linguistics*, 44(3).

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017 4, Volume 1: Long Papers*, pages 1073–1083, Vancouver, Canada.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What Makes a Good Conversation? How Controllable Attributes Affect Human Judgments. 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 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building End-To-end Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016*, pages 3776–3784, Phoenix, Arizona.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-decoder Model for Generating Dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017*, pages 3295–3301, San Francisco, California.

Charese Smiley, Elnaz Davoodi, Dezhao Song, and Frank Schilder. 2018. The E2E NLG Challenge: A Tale of Two Systems. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 472–477, Tilburg University, The Netherlands. Association for Computational Linguistics.

Marco Antonio Sobrevilla Cabezudo and Thiago A. S. Pardo. 2020. NILC at WebNLG+: Pretrained Sequence-to-Sequence Models on RDF-to-Text Generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 131–136, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to Summarize With Human Feedback. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014*, pages 3104–3112, Montreal, Quebec, Canada.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let Me Speak Freely? A Study On The Impact Of Format Restrictions On Large Language Model Performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *Preprint*, arXiv:2408.00118.

Craig Thomson and Ehud Reiter. 2020. A Gold Standard Methodology for Evaluating Accuracy in Data-To-text Systems. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, pages 158–168, Dublin, Ireland.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Trung Tran and Dang Tuan Nguyen. 2020. WebNLG 2020 Challenge: Semantic Template Mining for Generating References from RDF. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 177–185, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to Speak and Act in a Fantasy Text Adventure Game. 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 673–683, Hong Kong, China. Association for Computational Linguistics.

Pat Verga, Sebastian Hofstätter, 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. *CoRR*, abs/2404.18796.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015*, pages 2692–2700, Montreal, Quebec, Canada.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to Learn Automatic Summarization. In *Proceedings of the Workshop on New Frontiers in Summarization, NFiS at EMNLP 2017*, pages 59–63, Copenhagen, Denmark.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 5008–5020, Online.

Su Wang, Greg Durrett, and Katrin Erk. 2020b. Narrative Interpolation for Generating and Understanding Stories. *CoRR*, abs/2008.07466.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2025. HelpSteer2-Preference: Complementing Ratings with Preferences. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, Singapore.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. 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 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, New Orleans, LA, USA.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Peihao Su, David Vandyke, and Steve J. Young. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 1711–1721, Lisbon, Portugal.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *CoRR*, abs/1901.08149.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable Abstractive Dialogue Summarization with Sketch Supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5108–5122, Online. Association for Computational Linguistics.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The Next Chapter: A Study of Large Language Models in Storytelling. In *Proceedings of the 16th International Natural Language Generation Conference, INLG 2023*, pages 323–351, Prague, Czechia.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards Explainable Text Generation Evaluation with Automatic Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 5967–5994, Singapore.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT Has Universal Translations: An Analysis of Automatic Metrics by Minimum Risk Training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-Write: Towards Better Automatic Storytelling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 7378–7385, Honolulu, Hawaii, USA.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS*, pages 27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017 4, Volume 1: Long Papers*, pages 654–664, Vancouver, Canada.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing Precise and Robust Dialogue Response Evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 26–33, Online.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating With Contextualized Embeddings and Earth Mover Distance. 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 563–578, Hong Kong, China.

Yilun Zhao, Linyong Nan, Zhenting Qi, Rui Zhang, and Dragomir Radev. 2022. ReasTAP: Injecting Table Reasoning Skills During Pre-training via Synthetic Reasoning Examples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9006–9018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yilun Zhao, Zhenting Qi, Linyong Nan, Lorenzo Jaime Yu Flores, and Dragomir Radev. 2023a. LoFT: Enhancing Faithfulness and Diversity for Table-to-text Generation via Logic Form Control. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik*, pages 554–561, Croatia.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023b. Investigating Table-to-Text Generation Capabilities of Large Language Models in Real-World Information Seeking Scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a Unified Multi-dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 2023–2038, Abu Dhabi, United Arab Emirates.

A Aspects and datasets in the training data

This section provides a detailed overview of all source datasets and evaluation aspects used to generate our training data for OPENLGAUGE$_{ft}$.

A.1 Summarization

As source data for summarization, we utilize five datasets spanning three distinct tasks: news article summarization, forum post summarization and dialogue summarization. Our evaluation dataset includes six commonly used aspects: four related to meaning of the evaluated text, one to its form and one to both. Some of these aspects overlap significantly in their definitions. This is intentional, as our goal is to make the evaluator model robust to small variations in aspect definitions, and enable generalization to new aspects. Table 6 provides an overview of all summarization systems in our dataset.

A.1.1 Datasets

- **CNN/DailyMail** (Hermann et al., 2015) is a popular summarization dataset that consists of news articles from CNN and DailyMail, paired with corresponding bullet-point summaries. Originally developed for question answering, it was later adapted for summarization (Nallapati et al., 2016). For this dataset, we use pre-generated outputs from Stiennon et al. (2020), which include results from 11 different systems, including human references and an extractive baseline. Since our meta-evaluation datasets for summarization contain inputs from CNN/DailyMail, we ensure there is no overlap in the source texts between these datasets when sampling the inputs.

- **Newsroom** (Grusky et al., 2018) is a large-scale dataset of news articles and their summaries, collected from diverse sources, domains, authors and time range. In addition to newly generated LLM outputs, we use outputs of the systems evaluated in the original paper, which include three summarization systems and two extractive baselines.

- **SAMSum** dataset (Gliwa et al., 2019) addresses the dialogue summarization task, containing dialogues with short summaries created by linguists. In addition to outputs from several different LLMs, we use pre-generated outputs of six systems from (Gao and Wan, 2022).

- **TL;DR** (Völske et al., 2017) is a collection of Reddit posts with user-created summaries. Unlike many popular datasets that focus on news articles, TL;DR contains informal and less structured texts spanning diverse topics. Similarly to CNN/DailyMail, we use outputs from Stiennon et al. (2020).

- **XSum** (Narayan et al., 2018) is a dataset for the *extreme summarization* task, designed for abstractive approaches. Unlike datasets such as Newsroom or CNN/DailyMail, which favor extractive summarization, XSum contains BBC articles paired with concise, single-sentence summaries. We utilize pre-generated outputs from five systems and baselines evaluated in the original paper, along with newly generated output from LLMs.

A.1.2 Aspects

- **Consistency** evaluates whether the summary is factually aligned with the source text. This involves determining if the facts in the summary can be entailed by the source. Consistency is closely tied to hallucinations, which may be categorized as factual or non-factual. In our approach, all information is required to be supported by the source text, making both types of hallucinations inconsistent with the source. The definition of consistency used in our prompts is: *Extent to which the facts in the summary are consistent with the source text. Factually consistent summary should not contain facts that are not supported by the source text.*

- **Accuracy** is largely synonymous with consistency. It evaluates whether the factual information from the source is accurately represented in the summary. Following Stiennon et al. (2020), we define accuracy as: *Extent to which the factual information in the summary accurately matches the source text. An accurate summary should not contain information that is not present in the source text, should not contradict the source text, and generally should not be misleading.*

- **Relevance** of a summary is concerned with content selection. A relevant summary should

include important points from the source text while omitting unimportant details. Compared to consistency, relevance is more subjective, as determining what information should or should not be included in a summary can sometimes be ambiguous. The definition used in our evaluation is: *Extent to which the summary captures important information of the source text. A relevant summary should include all and only important information from the source text.*

- **Coverage** evaluates how much of the important information from the source text is covered by the summary. In this sense, it is closely related to relevance, however, it does not include non-redundancy as a criterion. We use a definition adapted from Stiennon et al. (2020): *Extent to which the summary covers the important information in the source text. A summary has good coverage if it mentions the main information from the source text that is important to understand the events described in the text. A summary has poor coverage if someone reading only the summary would be missing several important pieces of information about the event in the source text.*

- **Coherence** refers to the structural quality of the summary and involves attributes such as cohesion, consistency and relevance (Reinhart, 1980). It is determined by both semantic and formal structure of the text. We define coherence as: *Extent to which the summary is well-structured and organized, presenting information in a logical order that flows naturally from sentence to sentence. Coherent summary forms a unified body of information and makes it easy to understand the main ideas.*

- **Fluency** focuses on the formal quality of the text, including grammaticality and naturalness. Unlike coherence, fluency is concerned with sentence–level quality rather than the overall structure of the text. We define fluency as: *Formal quality of individual sentences of the summary. A fluent sentence should be grammatical, natural and easy to understand.*

A.2 Data-to-text

For data-to-text category, we collected inputs from four datasets that represent four distinct tasks: table-to-text, RDF-to-text, attribute-value list to text and logical NLG. Since it is important for an NLG evaluation method to reliably evaluate outputs with respect to structured data in different formats, we include a number of different input formats in our training data, including JSON, CSV and linearized tables with markup. Table 5 provides an overview of the input formats. The list of evaluated systems in our dataset is shown in Table 7.

A.2.1 Datasets

- **E2E NLG** dataset (Dusek et al., 2020; Novikova et al., 2017b) was chosen as a representative of simple data-to-text tasks. In this dataset, models are tasked with generating descriptions of restaurant venues based on attribute-value-based meaning representations (MRs). Target descriptions were crowdsourced using textual and pictorial representations of MRs as stimuli. To ensure diversity, we selected pre-generated outputs from five systems described by (Dusek et al., 2020), considering their model architectures and evaluation results across various metrics and aspects. These were extended by new outputs from several LLMs. The inputs and pre-generated outputs were sampled from the test set. Three different input formats are used in the training data (see Table 5).

- **WebNLG 2020** (Ferreira et al., 2020) is an RDF-to-text dataset designed for generating natural language text from RDF triples collected from DBPedia knowledge base (Mendes et al., 2012). Each input contains between one and seven triples, where each triple represents a binary relation in the form (subject, property, object). Similarly to E2E NLG, we selected a diverse set of output systems based on model architectures and evaluation results from the WebNLG+ 2020 Challenge, and used additional LLMs to generate new outputs.

- **ToTTo** (Parikh et al., 2020) is an open-domain table-to-text generation dataset. It consists of Wikipedia tables with highlighted cells, and the task is to generate single-sentence descriptions of the data in these highlighted cells. Inputs are provided in two formats: full tables, which include indices to highlighted cells, and linearized tables, where only the highlighted data is presented in a linear order, while the

Format	Example	Datasets
JSON	`{"name": "The Phoenix", "eatType": "pub", "food": "Indian", ...}`	E2E NLG
attribute-value (1)	`name: The Phoenix\neatType: restaurant\nfood: Indian\n ...`	E2E NLG
attribute-value (2)	`name[The Phoenix], eatType[restaurant], food[Indian], ...`	E2E NLG
RDF (1)	`"The_Velvet_Underground | genre | Proto-punk" ...`	WebNLG 2020
RDF (2)	`(The_Velvet_Underground, genre, Proto-punk) ...`	WebNLG 2020
CSV	`united states,32,1,31,12\naustralia,5,0,5,3\n ...`	LogicNLG
linearized table	`<page_title> List of Norwegian fjords </page_title> <section_title> ...`	ToTTo

Table 5: Input data formats used in our dataset for data-to-text tasks.

structure is annotated with markup tags. As we observed that even medium-sized open-weight LLMs often struggle with hallucinations when using full tables, we restricted our evaluation to linearized tables.

- **LogicNLG** (Chen et al., 2020) introduced the task of logical NLG, where models generate statements that can be logically entailed from the data in a table. This task involves various aggregations and comparisons, making it more difficult than simply transforming structured data to free-form text. Although the task explicitly requires generating five logical statements per input, we sampled between one to five statements from each generated output to increase the diversity of output lengths. Inputs were formatted as CSV with "|" as a separator. In addition to new LLM outputs, we used existing system outputs from Zhao et al. (2023b).

A.2.2 Aspects

- **Faithfulness** measures whether all information in the generated text is supported by the data, making it equivalent to precision. Similarly to factual consistency in summarization, we consider both factual and non-factual hallucinations as errors. For our evaluations, we define faithfulness as: *Extent to which the information in the text is supported by the data.*

- **Correctness** evaluates whether the information from the data is accurately presented in the generated text. The output is maximally correct if it does not contain any incorrect statements with respect to the input. Correctness overlaps significantly with faithfulness, but its definition varies based on the specific task. For example, we define correctness for LogicNLG as: *Extent to which the statements are logically and factually correct with respect to the provided data.*

- **Coverage** refers to the degree to which the generated text covers the information in the data. The output has maximum coverage when all information from the data is included in the text, which makes it analogous to recall. For example, in WebNLG 2020, it evaluates whether all predicates and their arguments are mentioned, while in ToTTo, it determines whether all highlighted table cells are described. We apply coverage to evaluate all datasets except LogicNLG, where the task is to infer interesting observations, rather than fully cover the source data. We define coverage as follows, with slight variations depending on the task: *Extent to which the text includes description of all information presented in the data.*

- **Informativeness** is closely related to coverage, as it evaluates how much of the information the generated text provides. However, it does not require complete coverage of the data, therefore it is applicable to tasks such as LogicNLG, where full coverage of a table is not necessary. The definition used in our evaluation depends on the particular dataset. For example, the definition used for LogicNLG is: *Extent to which the statements provide interesting or useful information about the data.*

- **Fluency** refers to the formal quality of the generated text, and includes grammaticality, naturalness and readability. Some definitions also include coherence (Ferreira et al., 2020), although coherence is usually treated as a separate aspect. For our evaluation, we define fluency as: *Extent to which the text is grammatical, natural and easy to understand.*

- **Grammaticality** focuses on the correctness of grammar and spelling in the generated text. A text is fully grammatical if it contains no

grammatical or spelling errors. While grammaticality is often included as a sub-aspect of fluency, both aspects are commonly used in practice. Therefore, we include it to help the model learn differences between evaluation aspects on different levels of hierarchy. In our dataset, grammaticality is defined as: *Extent to which the text is grammatical (free of grammar and spelling errors).*

- **Naturalness** refers either to the human-likeness of the text, or the likelihood that it was produced by a native speaker. Like grammaticality, naturalness is often treated as a component of fluency. Additionally, its evaluation often includes assessment of grammaticality, as this can often be an indicator of whether the text was produced by a native speaker. This illustrates how evaluation aspects often overlap or have hierarchical relationships. For our purposes, naturalness is defined as: *Extent to which the text is likely to have been produced by a native speaker.*

A.3 Dialogue Response Generation

For dialogue response generation, we source the inputs from three dialogue datasets, focusing on open-domain non-task-oriented dialogue response generation. Table 8 provides an overview of all evaluated systems in the training dataset.

A.3.1 Datasets

- **Wizard of Wikipedia** (Dinan et al., 2019) consists of conversations grounded in one of 1365 topics and corresponding knowledge retrieved from Wikipedia. In these conversations, either participant may select the topic and initiate the discussion, although they have asymmetric roles. One participant takes on the role of the wizard, an expert with access to a topic-relevant knowledge, on which they can base their responses. The other participant acts as an apprentice, a curious learner that is eager to discuss the chosen topic. To create inputs of varying length, we randomly select a dialogue history length between two turns and the full conversation, and truncate the dialogue to this length. The last utterance is replaced by a system output, except when the reference is used as evaluated output. In addition to newly generated LLM responses, we include human responses from the original dataset in the outputs.
- **EmpatheticDialogues** (Rashkin et al., 2019) is a dataset of dialogues grounded in emotional situations, designed to train and evaluate dialogue models on empathetic response generation. Each conversation is associated with an emotional label, where one of the participants describes a situation in which they experienced a given emotion. We sample up to five turns from each dialogue and replace the last utterance with a generated response. The emotion label is used as additional context for annotator LLMs to evaluate the appropriateness and empathy of the responses, but is excluded from the prompts used for system output generation.
- **DailyDialog** (Li et al., 2017) includes conversations on various daily life topics, annotated with emotion labels and communicative intents. We include data from DailyDialog to represent diverse topics and scenarios in the training set. Alongside newly generated responses, we also collect pre-generated outputs from three sources (Gupta et al., 2019; Huang et al., 2020; Zhao et al., 2020) to represent older dialogue systems.

A.3.2 Aspects

- **Coherence** in dialogue is a concept slightly different from coherence in tasks that involve generation of standalone texts, such as summarization or story generation. In dialogue, it measures how meaningful and logically consistent the response is with the preceding conversation. This includes not only alignment of the response with the last utterance, but also consistency with the dialogue participant's earlier responses in terms of logic and style. In our evaluation, coherence is broadly defined as: *Extent to which the response is a meaningful continuation of previous dialogue.*
- **Relevance** evaluates how closely a response aligns with the topic of conversation. In this sense, this aspect overlaps with coherence to some degree. We define relevance as: *Extent to which the response is relevant and on-topic given the dialogue history.*
- **Appropriateness** addresses whether the response is semantically and pragmatically appropriate in the given context. Depending on the definitions, appropriateness might overlap

to a large extent with coherence and relevance. For our evaluation, we define appropriateness as: *Extent to which the response is semantically and pragmatically appropriate given the conversation history.*

- **Empathy** is evaluated specifically on responses from the EmpatheticDialogues dataset, where the goal is to determine whether the response acknowledges and reflects the emotions of the other participant. We define empathy as: *Extent to which the response shows understanding of the feelings of the person talking about their experience.*

- **Interestingness** is concerned with the informational value of the response, specifically whether it presents stimulating ideas, facts or opinions. As it is one of the more subjective evaluation aspects, we define it vaguely and let the annotator models determine the criteria for interestingness: *Extent to which the response is interesting given the dialogue history.*

- **Engagingness** is closely related to interestingness but is sometimes treated as a distinct aspect (e.g., Mehri and Eskénazi, 2020a; See et al., 2019). While interestingness focuses on the context itself, engagingness emphasizes maintaining the user's attention and encouraging them to continue with the conversation. For our purposes, engagingness is defined as: *Extent to which the response captures and maintains the user's interest, encouraging further interaction. Engaging responses contain opinions, preferences, thoughts or interesting facts.*

- **Fluency** in dialogue response generation has a similar meaning to its use in other tasks and refers to the formal quality of the response. We define fluency as: *Extent to which the response is grammatically correct, natural and fluent.*

- **Understandability** evaluates both the content and form of a response, focusing on its clarity and ease of comprehension. The definition we use for our evaluation is: *Extent to which the response is easy to understand and comprehend given the dialogue history.*

A.4 Story Generation

The NLG tasks discussed so far generally contain inputs that are relatively longer compared to the outputs. This pattern is especially common in tasks like summarization and dialogue generation, although certain data-to-text tasks also share this characteristic. To represent scenarios with short inputs and long outputs, we include story generation in the training data. Table 9 lists all evaluated systems in our dataset.

A.4.1 Datasets

As the source of inputs, we use **WritingPrompts** (Fan et al., 2018), a story generation dataset derived from Reddit's WritingPrompts subreddit, where users submit prompts that can inspire other users to write stories. The dataset consists of a diverse range of topics, story lengths and writing styles. We reuse existing outputs from the OpenMEVA dataset (Guan et al., 2021) and generate additional outputs by four LLMs. To increase the diversity of generated stories in terms of their length, we generate the outputs with two different prompt versions, each requiring a different length of the story. The outputs are then randomly sampled from either the shorter or the longer set. As we observed a tendency of LLMs to generate a long list of errors for longer inputs, our evaluator models are instructed to limit the number of identified errors to a maximum of eight, and to prioritize the most severe ones if necessary.

A.4.2 Aspects

While relevance and coherence are two commonly used aspects for story generation, there is no consensus on which other evaluation aspects are the most relevant. Inspired by social sciences, Chhun et al. (2022) propose four additional aspects, aimed at providing a complete and non-redundant set of criteria. Following their work, we adopt the aspects defined in the HANNA benchmark, using our own definitions for most of them:

- **Relevance** measures the degree to which a story aligns with the given prompt (Chhun et al., 2022; Chiang and Lee, 2023b), title (Jhamtani and Berg-Kirkpatrick, 2020; Yao et al., 2019; Xie et al., 2023) or story beginning Wang et al. (2020b). In some cases, relevance also evaluates whether the story remains on-topic for its duration (e.g., Goldfarb-Tarrant et al., 2020). Since our inputs are

prompts, we define relevance as: *Extent to which the story is relevant to the writing prompt.*

- **Coherence** in story generation typically refers to logical consistency and narrative flow (e.g., Yao et al., 2019; Li et al., 2023; Jhamtani and Berg-Kirkpatrick, 2020). Other works define coherence more vaguely, such as how much the story "makes sense" (Chhun et al., 2022) or how well its sentences "fit together" (Xie et al., 2023). For our purposes, coherence is defined as: *Extent to which the story is logically consistent and coherent.*
- **Engagement** is a subjective and often vaguely defined aspect that evaluates how engaging the story is to the reader (e.g., Chhun et al., 2022; Li et al., 2023). Due to its inherent subjectivity, we apply a simple definition and leave its interpretation to the evaluator models: *Extent to which the story is engaging and interesting.*
- **Empathy** is related to emotional commentary and empathy, and refers to how well the story conveys character's emotions. We define empathy as: *The clarity and depth with which the character's emotions are conveyed in the story.*
- **Surprise** is concerned with the story's ending, and evaluates its unexpectedness and originality. We define surprise as: *How surprising the end of the story was.*
- **Complexity** measures how intricate and elaborate the story is. Complexity is not necessarily an aspect of quality, but rather a feature of the text. Whether greater complexity is desired or not depends on the audience. Our definition of complexity is: *How elaborate the story is.*

A.5 Question Answering

The question answering subset of our dataset consists of two distinct tasks: narrative question answering and table question answering. The inputs consist of a question and structured data in which the answer should be grounded. The evaluated systems in our training data are listed in Table 10.

A.5.1 Datasets

- **NarrativeQA** (Kociský et al., 2018) consists of human-written questions and free-form answers based on stories or their summaries. The stories include books and movie scripts and are provided either as full texts or as human-written summaries. Since full stories do not fit into the context window of many LLMs, we use only summaries to generate the outputs. Along with newly generated answers, we include human reference answers from the original datasets in the evaluated outputs.
- **FeTaQA** (Nan et al., 2022b) is a question answering dataset based on Wikipedia tables that requires models to aggregate and reason about the entities in the table and their relations. This places the task at the intersection of data-to-text and question answering task categories. In addition to new LLM outputs, we use pre-generated outputs from Zhao et al. (2023b).

A.5.2 Aspects

- **Correctness** evaluates if the answer to a question is correct with respect to the input. Since our models are instructed to assess the quality on an ordinal scale, we evaluate a *degree* of correctness – the answer should receive the maximum score if it is fully correct, while lower scores should reflect the number and severity of correctness issues. We define correctness as: *Extent to which the answer to the question is correct with respect to the input.*
- **Informativeness** addresses whether all information required by the question is provided in the answer. We define informativeness as: *Extent to which the answer provides all information that the question asked for.*
- **Completeness** evaluates comprehensiveness of the answer and the degree to which all aspects of the question are covered. The meaning is slightly different from informativeness, which is concerned with the information that the question explicitly asks for. In our dataset, completeness is defined as: *Extent to which the answer is comprehensive and ensures all question aspects are addressed.*
- **Conciseness** measures the degree to which an answer is focused and directly answers the question without unnecessary details and elaboration. Although the goal might often be to generate both complete and concise answers, these two aspects might correlate negatively.

Conciseness is defined as: *Extent to which the answer is concise and to the point.*

- **Relevance** is concerned with the specificity of an answer and measures the degree to which the answer addresses the particular question asked. Although it is related to conciseness, relevance is not that much concerned with the amount of detail in the answer. We define relevance as: *Extent to which the answer is specific and meaningful with respect to the question.*

- **Factuality** evaluates factual consistency of the answer with the provided context. In our dataset, this aspect is used in the narrative question answering task, and we use a similar definition as in summarization: *Extent to which the answer is supported by the summary.*

- **Faithfulness** is used for the table question answering task and is synonymous with factuality. We define faithfulness as: *Extent to which the information presented in the answer is supported by the input.*

- **Fluency** in question answering refers to the formal quality of the answer and is defined similarly as in the other tasks presented so far: *Extent to which the response is grammatical, natural and easy to understand.*

- **Naturalness** is interpreted in the same way as in data-to-text and dialogue response generation tasks, and the definition we apply is: *Extent to which the answer is likely to have been produced by a native speaker.*

- **Grammaticality** measures the grammatical quality of the answer and is defined as: *Extent to which the answer is grammatical (free of grammar and spelling errors).*

B Collected system outputs

Tables 6–10 list the evaluated systems for each task category in our dataset.

C LLMs used in the experiments

The specific versions of LLMs used in the experiments are presented in Table 11.

D Outliers

Table 12 shows the percentages of outliers detected for each annotator model in the ensemble and task category during synthetic data generation. To keep the merged evaluation outputs internally consistent, these outliers were removed before merging the outputs of the individual LLMs.

The proportions in the table indicate that the final annotation utilized most of the generated evaluations from the individual LLMs. The exceptions include Command R+ 104B and Gemma 2 27B on data-to-text tasks, with 29% and 10% outliers, respectively. Additionally, 10% evaluation outputs of Command R+ 104B were removed for question answering tasks. However, on average only 3.4% of evaluation outputs are detected as outliers across all model-task pairs.

E Meta-evaluation datasets

We used the following datasets for meta-evaluation. In the datasets where human scores are not already aggregated, we averaged the scores of all annotators.

- **SummEval** (Fabbri et al., 2021b) is a standard meta-evaluation dataset for summarization. It consists of summaries generated from CNN/-DailyMail articles, with 100 input articles and 16 different system outputs for each article. Human evaluations address four aspects: *(factual) consistency*, *relevance*, *coherence*, and *fluency*. Each output is scored by three expert annotators and five crowdworkers. We use only expert scores to measure correlations.

- **QAGS** (Wang et al., 2020a) contains annotations of *consistency* for summaries from CNN/DailyMail and XSum datasets. It includes 235 CNN/DailyMail summaries and 239 XSum summaries, each annotated by three evaluators. Annotators assign a binary factual consistency score (yes/no) for each sentence of the summary. We follow (Wang et al., 2020a) and apply majority voting for each sentence annotation, followed by averaging sentence-level scores to obtain the overall score for the summary.

- **HANNA** (Chhun et al., 2022) is a benchmark for story generation based on the WritingPrompts dataset. It contains three human scores for each of the 1056 stories across six

Table 6: Overview of the evaluated systems in the dataset for the summarization task. Sources: GW = Gao and Wan (2022), GR = Grusky et al. (2018), NA = Narayan et al. (2018), ST = Stiennon et al. (2020), new = newly generated.

System	Type	Sources
Qwen 2.5 0.5B (Yang et al., 2024)	Instruction-tuned LLM	new
Llama 2 7B Chat (Touvron et al., 2023)	Instruction-tuned LLM	new
Gemma 2 2B (Team et al., 2024)	Instruction-tuned LLM	new
Nous Hermes 2 Mixtral 8x7B DPO[7]	Instruction-tuned LLM	new
GPT-4o[8]	Instruction-tuned LLM	new
OpenAI summarization (pre-trained)	pre-trained LMs	ST
OpenAI summarization (supervised)	LMs trained with SFT	ST
OpenAI summarization (RLHF)	LMs trained with SFT+PPO	ST
T5 (Raffel et al., 2020)	pre-trained LM	ST
UniLM (Dong et al., 2019)	pre-trained LM	ST
CODS (Wu et al., 2021)	BART-based hybrid model	GW
ConvoSumm (Fabbri et al., 2021a)	BART-based model	GW
Ctrl-DiaSumm (Liu and Chen, 2021)	BART-based model	GW
ConvS2S (Gehring et al., 2017)	Convolutional seq-to-seq	NA
Topic-ConvS2S (Narayan et al., 2018)	Topic-conditioned ConvS2S	NA
S2S (Cho et al., 2014; Sutskever et al., 2014)	RNN-based seq-to-seq with attention	GR
PNG (Vinyals et al., 2015; Gülçehre et al., 2016)	Pointer-generator network	GR, GW, NA
TextRank (Barrios et al., 2016)	Ranking-based extractive summarization	GR
Reference	Human-written reference	ST
Title	Extractive baseline	ST
LEAD	Extractive baseline	NA
LEAD-2	Extractive baseline	ST
LEAD-3 (See et al., 2017)	Extractive baseline	GR, GW, ST
Ext-Oracle	Extractive oracle	NA
Fragments	Extractive oracle	NA

Table 7: Overview of the evaluated systems in the dataset for the data-to-text task. **Sources**: DU = Dusek et al. (2020), FE = Ferreira et al. (2020), ZH = Zhao et al. (2023b), new = newly generated.

System	Type	Sources
Qwen 2.5 Coder 1.5B (Yang et al., 2024)	Instruction-tuned LLM	new
Gemma 2 2B (Team et al., 2024)	Instruction-tuned LLM	new
Llama 2 7B Chat (Touvron et al., 2023)	Instruction-tuned LLM	new
Solar 10.7B (Kim et al., 2024a)	Instruction-tuned LLM	new
DeepSeek Coder v2 16B (DeepSeek-AI et al., 2024)	Instruction-tuned LLM	new
Nous Hermes 2 Mixtral 8x7B DPO[9]	Instruction-tuned LLM	new
Claude 3.5 Sonnet[10]	Instruction-tuned LLM	new
GPT-4o[11]	Instruction-tuned LLM	new
NILC (Sobrevilla Cabezudo and Pardo, 2020)	Fine-tuned BART	FE
Orange-NLG (Montella et al., 2020)	Fine-tuned BART	FE
Amazaon AI (Shanghai) (Guo et al., 2020)	Graph CNN + T5	FE
GPT2-C2F (Chen et al., 2020)	Fine-tuned GPT-2	ZH
LoFT (Zhao et al., 2023a)	Fine-tuned BART	ZH
PLOG (Liu et al., 2022a)	Fine-tuned T5	ZH
R2D2 (Nan et al., 2022a)	Fine-tuned T5	ZH
Flan-T5 (Chung et al., 2024)	Instruction-tuned LM	ZH
Adapt (Elder et al., 2018)	RNN seq-to-seq	DU
Sheff2 (Chen et al., 2018)	RNN seq-to-seq	DU
Slug (Juraska et al., 2018)	RNN + convolutional seq-to-seq	DU
Forge1 (Mille and Dasiopoulou, 2018)	Rule-based	DU
TR2 (Smiley et al., 2018)	Template-based	DU
DANGNT-SGU (Tran and Nguyen, 2020)	Template-based	FE
RALI-Université de Montréal (Lapalme, 2020)	Template-based	FE

Table 8: Overview of the evaluated systems in the dataset for the dialogue response generation task. Sources: GU = Gupta et al. (2019), HU = Huang et al. (2020), ZH = Zhao et al. (2020), new = newly generated.

System	Type	Sources
Claude 3.5 Sonnet[12]	Instruction-tuned LLM	new
GPT-4o[13]	Instruction-tuned LLM	new
Tülu 3 (Lambert et al., 2025)	Instruction-tuned LLM	new
Dolphin 2.9 Llama 3 8B[14]	Instruction-tuned LLM	new
Vicuna 7B (Zheng et al., 2023)	Instruction-tuned LLM	new
BlenderBot-small (Roller et al., 2021)	Dialogue LM	new
DialoGPT-small (Zhang et al., 2020b)	Dialogue LM	new
GPT-Neo 125M (Gao et al., 2021)	Pre-trained LM	new
GPT-2 (Wolf et al., 2019)	Pre-trained LM	ZH
CVAE (Zhao et al., 2017)	Conditional variational autoencoder	GU
HRED (Serban et al., 2016)	Hierarchical recurrent encoder-decoder	GU, ZH
Transformer-generator (Dinan et al., 2019)	Transformer-based generative model	HU
Transformer-ranker (Urbanek et al., 2019)	Transformer-based ranking model	HU
DualEncoder (Lowe et al., 2015)	LSTM dual encoder	GU
VHRED (Serban et al., 2017)	Latent variable HRED	ZH
S2S (Cho et al., 2014; Sutskever et al., 2014)	RNN-based seq-to-seq with attention	GU, ZH

Table 9: Overview of the evaluated systems in the dataset for the story generation task. Sources: GU = Guan et al. (2021), new = newly generated.

System	Type	Sources
Gemma 2 2B (Team et al., 2024)	Instruction-tuned LLM	new
Dolphin 2.9 Llama 3 8B[15]	Instruction-tuned LLM	new
Nous Hermes 2 Mixtral 8x7B DPO[16]	Instruction-tuned LLM	new
GPT-4o[17]	Instruction-tuned LLM	new
GPT-2 (Radford et al., 2019)	Pre-trained LM	GU
GPT-KG (Guan et al., 2020)	Knowledge-enhanced GPT-2	GU
Fusion (Fan et al., 2018)	Convolutional seq-to-seq with attention	GU
Plan&Write (Yao et al., 2019)	Hierarchical RNN-based model	GU
S2S (Cho et al., 2014; Sutskever et al., 2014)	RNN-based seq-to-seq	GU

Table 10: Overview of the evaluated systems in the dataset for the question answering task. Sources: KO = Kociský et al. (2018), ZH = Zhao et al. (2023b), new = newly generated.

System	Type	Sources
Claude 3.5 Sonnet[18]	Instruction-tuned LLM	new
GPT-4o[19]	Instruction-tuned LLM	new
Nous Hermes 2 Mixtral 8x7B DPO[20]	Instruction-tuned LLM	new
DeepSeek Coder v2 16B (DeepSeek-AI et al., 2024)	Instruction-tuned LLM	new
Llama 2 7B Chat (Touvron et al., 2023)	Instruction-tuned LLM	new
Qwen 2.5 Coder 1.5B (Yang et al., 2024)	Instruction-tuned LLM	new
Qwen 2.5 0.5B (Yang et al., 2024)	Instruction-tuned LLM	new
GPT-Neo 125M (Gao et al., 2021)	Pre-trained LM	new
BART (Lewis et al., 2020)	Pre-trained LM	ZH
Flan-T5 (Chung et al., 2024)	Instruction-tuned LM	ZH
OmniTab (Jiang et al., 2022)	Table pre-trained LM	ZH
ReasTAP (Zhao et al., 2022)	Table pre-trained LM	ZH
TAPEX (Liu et al., 2022b)	Table pre-trained LM	ZH
Reference	Human-written reference	KO

Model	Quantization	Tag
Command R+ 104B	5-bit	`command-r-plus:104b-08-2024-q5_K_M`
Gemma 2 27B	8-bit	`gemma2:27b-instruct-q8_0`
Llama 3.1 Nemotron 70B	8-bit	`nemotron:70b-instruct-q8_0`
Llama 3.3 70B	8-bit	`llama3.3:70b-instruct-q8_0`
Mistral Large 2 123B	4-bit	`mistral-large:123b-instruct-2407-q4_K_M`
Qwen 2.5 72B	8-bit	`qwen2.5:72b-instruct-q8_0`

Table 11: Quantization levels and Ollama tags used for the models.

Model	Data-to-text	Summarization	Dialogue	Story Generation	Question Answering
Command R+ 104B	28.88	0.13	0.73	0.94	9.61
Gemma 2 27B	10.32	1.96	1.52	1.03	5.07
Llama 3.1 Nemotron 70B	4.27	1.48	4.02	1.03	0.85
Mistral Large 2 123B	4.89	0.93	0.63	0.25	1.22
Qwen 2.5 72B	1.39	1.36	1.43	1.12	0.47

Table 12: Proportions (%) of evaluation outputs detected as outliers for each model and task category. Across all model-task pairs, 3.4% of evaluation outputs are detected as outliers on average (median = 1.36%).

aspects: *relevance*, *coherence*, *engagement*, *empathy*, *surprise* and *complexity* (see Appendix A.4 for details on these aspects).

- **SFRES** and **SFHOT** (Wen et al., 2015) are used to perform a meta-evaluation for data-to-text. These datasets consist of dialogue acts (DAs) in structured format with generated responses providing information about restaurants and hotels in San Francisco. Human judgments are provided for *informativeness* and *naturalness*.

- **TopicalChat** (Gopalakrishnan et al., 2019) annotations from the USR dataset (Mehri and Eskénazi, 2020b) are used for meta-evaluation of dialogue response generation. The dataset includes human evaluations for five aspects: *groundedness*, *coherence*, *interestingness*, *naturalness* and *understandability*. As the source data differ from our training data, TopicalChat serves as an out-of-domain evaluation dataset. Additionally, since *groundedness* is not present in our training data, we evaluate it as an unseen aspect.

- **Wiki-DA** (Alva-Manchego et al., 2021) is a dataset of DA human ratings for text simplification, a task unseen by OpenNLGauge$_{ft}$ during training. Along with scores for *fluency*, this dataset also includes two unseen aspects: *meaning preservation* and *simplicity*.

F Prompt templates

Prompt templates for annotator and consolidator LLMs are presented in Listings 1–2. Prompt template used for fine-tuning and inference of the distilled model is shown in Listing 3.

G Full results

Tables 14–20 contain the meta-evaluation results for additional datasets described in Appendix E. Detailed results of human evaluation of error span quality are presented in Table 13.

G.1 Kendall's τ correlations

On QAGS, our method achieves lower Kendall's τ than Themis, although it surpasses the metric on Spearman's ρ. This discrepancy can be attributed to a difference in score precisions between these two methods. As a result of averaging, OpenNLGauge$_{ens}$ provides more granular floating point scores, while Themis predicts integer scores on a Likert scale. Generally, we can expect more tied pairs (i.e. pairs that are neither concordant nor discordant) in the calculation of Kendall's τ with integer scores, which can have a substantial effect on the correlations. In QAGS, 78% of human scores map to integer scores after normalization. Therefore, we consider Spearman's ρ more appropriate for comparing evaluation capabilities of different metrics.

Similarly, specific individual LLMs score higher on Kendall's τ than the ensemble on some datasets (see Tables 3, 14, 18 and 19). This is due to the same reason outlined above. Specifically, there are

System	Explanation	Error span	Error	No span given	Not an error	Hallucination	Total
OpeNLGauge$_{ens}$	Correct		218	9	3	0	230
	Partially correct		41	0	0	1	42
	Incomplete		20	0	1	0	21
	Vague		4	2	1	0	7
	Incorrect		20	3	32	1	56
	Not an error		0	0	10	0	10
	Total		303	14	47	2	366
	Span OK (%)		**83**				
	Exp. correct (%)		**63**				
OpeNLGauge$_{ft}$	Correct		108	3	4	4	119
	Partially correct		60	1	2	0	63
	Incomplete		9	0	0	0	9
	Vague		3	9	0	0	12
	Incorrect		41	7	42	6	96
	Not an error		0	1	8	0	9
	Total		221	21	56	10	308
	Span OK (%)		**72**				
	Exp. correct (%)		**39**				
TigerScore	Correct		8	17	0	0	25
	Partially correct		5	2	0	0	7
	Incomplete		2	2	0	0	4
	Vague		0	3	0	0	3
	Incorrect		5	16	2	0	23
	Not an error		1	2	0	0	3
	Total		21	42	2	0	65
	Span OK (%)		**32**				
	Exp. correct (%)		**38**				

Table 13: Detailed results of human evaluation of error-spans and their explanations. Each table shows absolute occurrence counts of different error span and explanation validity (see Section 7.2), with overall proportions of correct annotation given below. The reported differences between TigerScore and OPENLGAUGE$_{ens}$ are statistically significant (t-test, $p < 0.05$).

37% tied pairs for OPENLGAUGE$_{ens}$ and human scores, while the individual LLMs, which provide integer scores, show substantially larger proportions of tied pairs: between 46% and 54%.

H Ablation experiments

H.1 Prompt ablations

The detailed results of prompt ablation experiments are presented in Figure 11–14.

H.2 Ensemble size ablations

The effect of the ensemble size on Spearman correlations are presented in Figures 5–10.

I Inter-annotator agreement between ensemble models

To obtain additional insights into the variance between individual ensemble LLMs in their predicted overall scores, we compute pairwise inter-annotator agreements between the models on all meta-evaluation datasets.

Figure 16 shows the Spearman correlations for each LLM pair and dataset used in our meta-evaluation, where the correlations are calculated over all evaluation aspects in a given dataset. The agreement is generally high on the QAGS and Wiki-DA datasets, although substantially lower on other datasets. Note that individual models also achieved relatively high correlations with humans on these two datasets, which indicates that they might generally consist of examples (and possibly evaluation aspects) for which it is easier for LLMs to make decisions on their quality. In contrast, the models show relatively low agreement on HANNA, which could be explained by the more subjective nature of the story generation task and the corresponding evaluation aspects.[21] Across all tasks, Command R+ 104B tends to disagree more with other LLMs when compared to other pairwise agreements.

To assess the agreement of the models on exact overall score predictions, we also measure pairwise Cohen's κ (Figure 17), which treats the overall scores as categorical and therefore serves as a stricter measure. As expected, the agreements are lower compared to those measured by Spearman correlations, while the general trend is the same: the agreement on exact scores is higher on QAGS and Wiki-DA than on other datasets, while Command R+ 104B generally disagrees more with other LLMs across tasks.

Finally, we compute Krippendorff's α (Krippendorff, 2011), which allows us to measure the agreement between all ensemble models, while also being applicable to ordinal data. In general, the agreements are low to moderate. Similarly to the pairwise agreements, we observe substantially

[21] While such hypotheses could in principle be tested by comparing our results with the agreement between human annotators on a particular task, there are several issues that limit the reliability of such comparisons – for each meta-evaluation dataset we use, at least two of the following hold: (1) the number of human annotators is different from the number of LLMs in our ensemble, (2) either the sets of human annotators differ between evaluated outputs, or it is not made explicit in the dataset (or the corresponding paper) whether this is the case, (3) rating scales are different from ours (often three levels, or even binary).

higher agreements on QAGS and Wiki-DA, while the models agree the least on scoring generated stories in the HANNA dataset.

Overall, our analysis indicates that the predictions of individual models are sufficiently diverse to benefit from the combination in an ensemble without much redundancy.

J Score aggregation methods

Table 21 compares results of different approaches to aggregation of ensemble scores:

- **Average** computes the final score for the evaluated output as a simple average of scores from all individual LLMs.
- **Average w/o outliers** first removes the outliers before computing the average. The score is considered an outlier if it differs from the average of other scores for the same example by at least two standard deviations and this difference is at least 1.
- **Median** computes the final score as a median of the scores from individual LLMs to disregard the effect of extreme scores.
- **Majority voting** selects the most frequent score from the individual LLMs for the given output.
- **Min** selects the minimum of individual scores as the final score. This corresponds to the most strict evaluation, i.e. typically the evaluation that detected the most errors or assigned highest severity levels to the identified errors.

K False positives analysis

To estimate the extent of potential over-annotation by OPENLGAUGE$_{ens}$ and its components, we analyze overall score predictions for output-aspect pairs (y, a) which obtained maximum scores by all human annotators in a given dataset. We denote these examples Y_{max} and assume that if $(y, a) \in Y_{max}$, then y does not contain any errors with respect to aspect a.

Our analysis includes only those datasets and aspects that contain sufficiently fine-grained annotation scales (at least three levels), as it is unclear whether maximum scores on a binary rating scale reliably indicate a perceived lack of errors by human annotators. Additionally, we exclude datasets where the size of Y_{max} is too small or even zero[22]. Given an output-aspect pair $(y, a) \in Y_{max}$, we consider an LLM evaluation of y with respect to a an over-annotation if it contains one or more detected errors.

Figure 18 shows the distribution of error counts (top), mean severity levels (middle) and maximum severity levels (bottom) per output, as predicted by OPENLGAUGE$_{ens}$ and its components for the Y_{max} subset of SummEval. Most individual LLMs assess the outputs in Y_{max} as error-free, particularly Command R+ 104B, which agrees almost perfectly with human annotators in this subset. Note that in general, Command R+ 104B tends to disagree the most with other LLMs, as discussed in Appendix I, while also achieving the highest correlations with humans in evaluating factual consistency (Table 14), which represents approximately half of the examples in Y_{max}. In contrast, the ensemble shows a tendency to include at least one error in most of its annotations for Y_{max}. This could be attributed to the merging procedure by the consolidator model, which aggregates errors from five different models and could lead to accumulation of errors. As expected, an increasing number of errors detected by a model is also reflected in the mean and maximum severity levels.

The error annotations for TopicalChat (Figure 19) show a similar pattern, although a larger proportion of individual models have a tendency to detect one or more errors in this case.

L Output examples

Figures 20–22 show additional output examples for RDF-to-text, dialogue response generation, and summarization tasks.

[22] For example, Wiki-DA contains only already aggregated scores on a scale between 0 and 100, with none of them equal to the maximum possible score. Although we could allow some deviation from the maximum score (e.g., maximum 5 points) to select Y_{max}, any such threshold would be arbitrary. Therefore, we exclude this dataset from the analysis.

```
### Instructions
Your task is to evaluate an output of data-to-text task, where the model was
instructed to write a single-paragraph description of a venue based on the given
data. The data consist of a set of attribute-value pairs in the form 'attribute:
value'.
Based on the given data and the generated text, identify errors in the text with
respect to {{ aspect_name }} (described below).
For each error, determine its severity on a scale from 1 to 5, where 1 is the
least severe and 5 is the most severe.

Definition of {{ aspect_name }}:
{{ aspect_definition }}

Rules:
Do not make assumptions and do not bring in external knowledge not present in the
provided context.
Identify only the errors related to the {{ aspect_name }} of the text. Do not
consider other aspects like {{ negative_aspects }}!
If there are no errors related to {{ aspect_name }} in the text, you should output
'No Error' and provide 'Excellent' score.

Steps:
1. Carefully read the data and identify the main attributes and their values.
2. Read the generated text and compare it with the source data with respect to {{
aspect_name }}.
3. If the text contains any error that negatively affects its {{ aspect_name }},
identify its exact location (specific word or phrase), explain why it is
considered an error, and determine the severity of the error.
4. Finally, provide an overall score for the {{ aspect_name }} of the text. The
score should be a label on the following scale (lowest to highest):
'Unacceptable', 'Poor', 'Fair', 'Good', 'Excellent'. The score 'Unacceptable'
indicates that the text is {{ min_score_desc }}, while 'Excellent' indicates that
the text is {{ max_score_desc }}.

### Data
{{ input }}

### Generated Text
{{ output }}

### Output format:
Generate your output exactly in this format:
```
Error 1:
Location: <location of the error - the exact word or phrase in the response>
Explanation: <explanation for the error, including the reason why it is considered
{{ aspect_name }} issue>
Severity: <integer from 1 to 5>

Error 2:
...

Overall score: <one of: Unacceptable, Poor, Fair, Good, Excellent>
Explanation of the score: <explanation of the score>
```
```

Listing 1: Annotator prompt template for the data-to-text task.

```
### Instructions
You are given multiple error annotation sets for an AI model output. Your task is
to merge the annotation sets to a single final annotation set.
The result shouldn't contain any duplicates. If there are multiple error
annotations for approximately the same location that describe the same issue, you
should merge them into single location. Otherwise, the error annotations should be
as granular as possible. If there are multiple different locations with the same
issue, each should have its own error annotation. Likewise, if there are multiple
issues with respect to the same location, each should have its own error
annotation. Use the following guidelines:
* Each error annotation should describe a single issue.
* Merge only annotations where the locations have significant overlap.
* When merging multiple locations, choose a single span from the output text that
covers the locations from merged annotations.
* Never include multiple spans from the annotations under the same "Location" line.
* Do not include any other text in "Location" line than the text that is actually
in the output, except for annotations that mention omissions or similar issues.
* When merging explanations, combine the most relevant information from the merged
annotations.
* Severity levels range from 1 to 5 from least severe to most severe. Use the most
severe level from the merged annotations.
* Final annotation set should not include more than 8 error annotations. If there
are more than that, use only the most severe ones.
* Make the final annotation set as concise as possible in terms of number of error
annotations.

Don't use any markdown formatting. Generate merged error annotations in this
format, without any additional text:
```
Error 1:
Location: <span of text from the output, or None if not applicable>
Explanation: <explanation>
Severity: <severity level>
```

### Model output
{{ output }}

### Error annotations
{{ annotations }}
```

Listing 2: Consolidator prompt template for merging of individual annotations.

```
### Task
Your task is to evaluate a model output for {{ task_name }} task with respect to
{{ aspect_name }}. {{ extra_task_info }}

### Aspect Definition
{{ aspect_name }} - {{ aspect_definition }}

### Dialogue history
{{ input }}
{% if context %}
### Knowledge
{{ context }}
{% endif %}
### Response
{{ output }}

### Instructions
For any error in the output, identify its location, assign a severity level and
provide an explanation. Report at most 8 errors. If there are more errors, report
only the most severe ones. Finally, provide an overall score between 0 and 100 for
{{ aspect_name }} of the output.
```

Listing 3: Prompt template for the fine-tuned OPENLGAUGE$_{ft}$ metric.

Metric	Consistency		Coherence		Relevance		Fluency		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252	0.192	0.150
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219	0.205	0.161
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
BERTScore	0.151	0.122	0.285	0.220	0.302	0.232	0.186	0.154	0.231	0.182
MOVERSscore	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BARTScore	0.266	0.220	0.474	0.367	0.318	0.243	0.258	0.214	0.329	0.261
UniEval	0.446	0.371	0.575	0.442	0.426	0.325	0.449	0.371	0.474	0.377
G-Eval (GPT-3.5)	0.386	0.318	0.440	0.335	0.385	0.293	0.424	0.347	0.409	0.323
G-Eval (GPT-4)	0.507	0.425	0.582	0.457	**0.548**	**0.433**	0.455	0.378	0.523	0.423
LLM Evaluation (GPT-3.5)	0.393	0.331	0.459	0.371	0.455	0.363	0.355	0.296	0.415	0.340
LLM Evaluation (GPT-4)	0.531	0.464	0.540	0.434	0.491	0.395	0.480	0.409	0.511	0.426
X-Eval	0.428	0.340	0.530	0.382	0.500	0.361	0.461	0.365	0.480	0.362
Prometheus	0.150	0.137	0.150	0.126	0.164	0.138	0.189	0.168	0.163	0.142
AUTO-J	0.131	0.121	0.245	0.203	0.262	0.222	0.154	0.141	0.198	0.172
TIGERScore	0.427	0.387	0.381	0.318	0.366	0.304	0.363	0.327	0.384	0.334
InstructScore	0.232	0.213	0.328	0.276	0.211	0.179	0.260	0.237	0.258	0.226
Themis	0.600	0.566	0.566	**0.485**	0.474	0.412	**0.571**	**0.533**	**0.553**	**0.499**
OpenNLGauge$_{ens}$	0.548	0.470	**0.604**	0.462	0.513	0.389	0.470	0.389	0.534	0.427
• Command R+ 104B	**0.633**	**0.603**	0.239	0.203	0.360	0.302	0.347	0.320	0.395	0.357
• Gemma 2 27B	0.459	0.421	0.455	0.386	0.428	0.358	0.427	0.386	0.442	0.388
• Llama 3.1 Nemotron 70B	0.559	0.517	0.469	0.391	0.419	0.356	0.358	0.325	0.451	0.397
• Mistral Large 2 123B	0.627	0.590	0.528	0.434	0.456	0.375	0.398	0.359	0.502	0.439
• Qwen 2.5 72B	0.567	0.521	0.525	0.433	0.388	0.317	0.433	0.389	0.478	0.415
Llama 3.1 8B	0.181	0.165	0.176	0.150	0.156	0.127	0.232	0.218	0.186	0.165
OpenNLGauge$_{ft}$	0.527	0.453	0.561	0.441	0.514	0.408	0.407	0.349	0.502	0.413

Table 14: Segment-level Spearman (ρ) and Kendall (τ) correlations of different metrics on SummEval.

Metric	Groundedness		Coherence		Engagingness		Naturalness		Average	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
ROUGE-L	0.193	0.203	0.176	0.146	0.295	0.300	0.310	0.327	0.243	0.244
BLEU-4	0.131	0.235	0.180	0.175	0.232	0.316	0.213	0.310	0.189	0.259
METEOR	0.250	0.302	0.212	0.191	0.367	0.439	0.333	0.391	0.290	0.331
BERTScore	0.214	0.233	0.226	0.209	0.317	0.335	0.291	0.317	0.262	0.273
USR	0.416	0.377	0.337	0.325	0.456	0.465	0.222	0.447	0.358	0.403
x UniEval	0.602	0.455	0.455	0.330	0.573	0.430	0.577	0.453	0.552	0.417
G-Eval (GPT-3.5)	0.586	0.567	0.519	0.544	0.660	0.691	0.532	0.539	0.574	0.585
G-Eval (GPT-4)	0.531	0.551	0.594	0.605	0.627	0.631	0.549	0.565	0.575	0.588
LLM Evaluation (GPT-3.5)	0.653	0.581	0.550	0.531	0.651	0.648	0.515	0.550	0.592	0.578
LLM Evaluation (GPT-4)	**0.810**	0.786	**0.680**	**0.680**	**0.822**	**0.779**	**0.769**	**0.739**	**0.770**	**0.746**
X-Eval	0.734	0.728	0.558	0.622	0.449	0.593	0.417	0.478	0.539	0.605
Prometheus	0.437	0.412	0.451	0.465	0.495	0.473	0.355	0.384	0.435	0.434
AUTO-J	0.339	0.357	0.452	0.449	0.490	0.459	0.425	0.437	0.427	0.425
TIGERScore	0.137	0.138	0.417	0.438	0.328	0.333	0.455	0.477	0.334	0.346
InstructScore	0.140	0.102	0.299	0.297	0.264	0.233	0.374	0.332	0.269	0.241
Themis	0.778	0.761	0.639	0.644	0.790	0.766	0.727	0.729	0.733	0.725
OpenNLGauge$_{ens}$	0.704	0.697	0.622	0.621	0.675	0.692	0.599	0.604	0.649	0.653
• Command R+ 104B	0.383	0.368	0.463	0.453	0.262	0.259	0.421	0.398	0.386	0.374
• Gemma 2 27B	0.332	0.366	0.465	0.481	0.549	0.562	0.489	0.515	0.459	0.484
• Llama 3.1 Nemotron 70B	0.781	**0.791**	0.600	0.630	0.655	0.686	0.506	0.532	0.621	0.645
• Mistral Large 2 123B	0.658	0.648	0.541	0.554	0.645	0.659	0.509	0.529	0.586	0.596
• Qwen 2.5 72B	0.467	0.460	0.480	0.521	0.500	0.516	0.468	0.488	0.496	0.514
Llama 3.1 8B	0.374	0.362	0.225	0.228	0.415	0.408	0.222	0.238	0.309	0.309
OpenNLGauge$_{ft}$	0.485	0.538	0.531	0.575	0.622	0.650	0.522	0.547	0.540	0.578

Table 15: Segment-level Pearson (r) and Spearman (ρ) correlations of different metrics on TopicalChat.

Metric	SFRES		SFHOT		Average
	Inf.	Nat.	Inf.	Nat.	
ROUGE-1	0.115	0.170	0.118	0.196	0.150
ROUGE-L	0.103	0.169	0.110	0.186	0.142
BERTScore	0.156	0.219	0.135	0.178	0.172
MOVERScore	0.153	0.190	0.172	0.242	0.189
BARTScore	0.238	0.289	0.235	0.288	0.263
UniEval	0.225	0.333	0.249	0.320	0.282
GPTScore	0.232	0.190	0.184	0.036	0.161
G-Eval (GPT-4)	0.189	0.351	0.198	0.338	0.269
LLM Evaluation (GPT-3.5)	0.304	0.385	0.242	0.294	0.306
LLM Evaluation (GPT-4)	0.213	0.405	**0.302**	0.359	0.320
Prometheus	0.161	0.150	0.211	0.169	0.173
Auto-J	0.179	0.084	0.176	0.127	0.141
TIGERScore	0.160	0.221	0.215	0.204	0.200
InstructScore	0.194	0.300	0.222	0.273	0.247
Themis	0.298	0.395	0.259	**0.380**	**0.333**
OpenNLGauge$_{ens}$	0.234	**0.415**	0.205	0.341	0.299
• Command R+ 104B	0.239	0.298	0.215	0.311	0.266
• Gemma 2 27B	0.254	0.334	0.275	0.317	0.295
• Llama 3.1 Nemotron 70B	0.178	0.284	0.047	0.209	0.179
• Mistral Large 2 123B	0.129	0.359	0.226	0.281	0.249
• Qwen 2.5 72B	0.226	0.347	0.245	0.315	0.283
Llama 3.1 8B	0.003	0.081	0.071	0.094	0.108
OpenNLGauge$_{ft}$	**0.354**	0.354	0.238	0.315	0.315

Table 16: Segment-level Spearman (ρ) correlations of different metrics on SFRES and SFHOT. **Inf.** = informativeness, **Nat.** = naturalness.

Metric	Coh.	Rel.	Eng.	Emp.	Sur.	Com.	Avg.
BLEU	0.539	0.514	0.483	0.410	0.471	0.516	0.489
ROUGE-1	**0.567**	0.518	0.529	**0.450**	**0.490**	**0.591**	**0.524**
METEOR	0.560	0.522	0.510	0.435	0.488	0.555	0.512
MoverScore	0.551	0.523	0.495	0.418	0.478	0.530	0.499
BERTScore	0.566	0.531	0.520	0.441	0.488	0.563	0.518
BARTScore	0.501	0.467	0.465	0.416	0.436	0.488	0.462
OpenNLGauge$_{ens}$	0.528	**0.559**	**0.538**	0.434	0.343	**0.591**	0.499
• Command R+ 104B	0.412	0.383	0.420	0.330	0.227	0.344	0.353
• Gemma 2 27B	0.453	0.445	0.461	0.356	0.323	0.521	0.427
• Llama 3.1 Nemotron 70B	0.500	0.508	0.497	0.380	0.332	0.565	0.464
• Mistral Large 2 123B	0.372	0.490	0.425	0.373	0.334	0.507	0.417
• Qwen 2.5 72B	0.407	0.484	0.427	0.277	-0.012	0.507	0.348
Llama 3.1 8B	0.119	0.344	0.154	0.094	0.080	0.212	0.167
OpenNLGauge$_{ft}$	0.448	0.523	0.517	0.444	0.404	0.555	0.482

Table 17: Segment-level Pearson (r) correlations of different metrics on HANNA. **Coh.** = coherence, **Rel.** = relevance, **Eng.** = engagement, **Emp.** = empathy, **Sur.** = surprise, **Com.** = complexity, **Avg.** = average.

Metric	Coh.	Rel.	Eng.	Emp.	Sur.	Com.	Avg.
BLEU	0.339	0.292	0.356	0.315	0.299	0.414	0.336
ROUGE-1	0.389	0.330	0.416	0.354	**0.355**	0.503	0.391
METEOR	0.378	0.310	0.412	0.366	0.354	**0.505**	0.387
MoverScore	0.392	0.385	0.420	0.331	0.321	0.473	0.387
BERTScore	0.372	0.355	0.415	0.356	0.320	0.469	0.381
BARTScore	0.259	0.249	0.291	0.287	0.227	0.294	0.268
OpenNLGauge$_{ens}$	0.393	0.474	0.452	0.367	0.276	0.489	0.409
• Command R+ 104B	0.325	0.362	0.372	0.320	0.215	0.348	0.324
• Gemma 2 27B	0.381	0.383	0.400	0.310	0.278	0.445	0.366
• Llama 3.1 Nemotron 70B	**0.429**	0.445	0.427	0.328	0.263	0.464	0.393
• Mistral Large 2 123B	0.294	0.415	0.389	0.349	0.337	0.474	0.376
• Qwen 2.5 72B	0.344	0.423	0.364	0.262	-0.006	0.416	0.301
Llama 3.1 8B	0.093	0.334	0.140	0.086	0.064	0.184	0.150
OpenNLGauge$_{ft}$	0.416	**0.498**	**0.464**	**0.391**	0.319	0.460	**0.425**

Table 18: Segment-level Spearman (ρ) correlations of different metrics on HANNA. **Coh.** = coherence, **Rel.** = relevance, **Eng.** = engagement, **Emp.** = empathy, **Sur.** = surprise, **Com.** = complexity, **Avg.** = average.

Metric	Coh.	Rel.	Eng.	Emp.	Sur.	Com.	Avg.
BLEU	0.248	0.209	0.260	0.230	0.220	0.305	0.245
ROUGE-1	0.287	0.237	0.306	0.260	0.262	0.376	0.288
METEOR	0.278	0.224	0.303	0.269	0.261	0.377	0.285
MoverScore	0.289	0.280	0.308	0.242	0.236	0.353	0.285
BERTScore	0.273	0.257	0.304	0.260	0.234	0.348	0.279
BARTScore	0.185	0.177	0.209	0.206	0.164	0.212	0.192
OpenNLGauge$_{ens}$	0.307	0.367	0.350	0.280	0.208	0.378	0.315
• Command R+ 104B	0.270	0.297	0.300	0.253	0.171	0.277	0.261
• Gemma 2 27B	0.329	0.328	0.343	0.267	0.241	0.384	0.315
• Llama 3.1 Nemotron 70B	**0.369**	0.377	0.364	0.272	0.221	0.388	0.332
• Mistral Large 2 123B	0.253	0.354	0.334	0.301	**0.284**	**0.405**	0.322
• Qwen 2.5 72B	0.294	0.360	0.307	0.222	-0.003	0.353	0.255
Llama 3.1 8B	0.079	0.285	0.116	0.073	0.055	0.152	0.127
OpenNLGauge$_{ft}$	0.341	**0.400**	**0.377**	**0.323**	0.257	0.377	**0.346**

Table 19: Segment-level Kendall (τ) correlations of different metrics on HANNA. **Coh.** = coherence, **Rel.** = relevance, **Eng.** = engagement, **Emp.** = empathy, **Sur.** = surprise, **Com.** = complexity, **Avg.** = average.

Metric	Fluency	Meaning	Simplicity	Average
BLEU	0.460	0.622	0.438	0.507
SARI	0.335	0.534	0.366	0.412
BERTScore	0.636	0.682	0.614	0.644
LENS	0.816	0.662	0.733	0.737
OpenNLGauge$_{ens}$	**0.840**	**0.864**	**0.770**	**0.825**
• Command R+ 104B	0.704	0.787	0.601	0.697
• Gemma 2 27B	0.755	0.769	0.688	0.737
• Llama 3.1 Nemotron 70B	0.778	0.822	0.660	0.753
• Mistral Large 2 123B	0.705	0.744	0.735	0.728
• Qwen 2.5 72B	0.771	0.829	0.730	0.776
Llama 3.1 8B	0.373	0.528	0.313	0.405
OpenNLGauge$_{ft}$	0.801	0.851	0.716	0.789

Table 20: Segment-level Pearson (r) correlations of different metrics on Wiki-DA. **Meaning** = Meaning preservation.

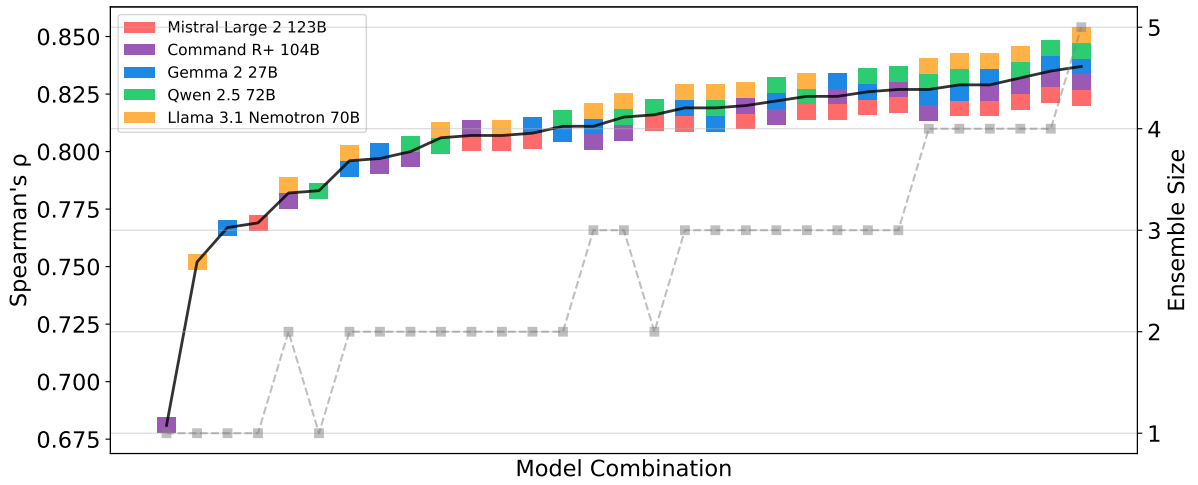

Figure 5: Effect of ensemble size on Spearman's ρ correlations with human scores for the Wiki-DA dataset. Specific model combinations are represented by the colored patches.

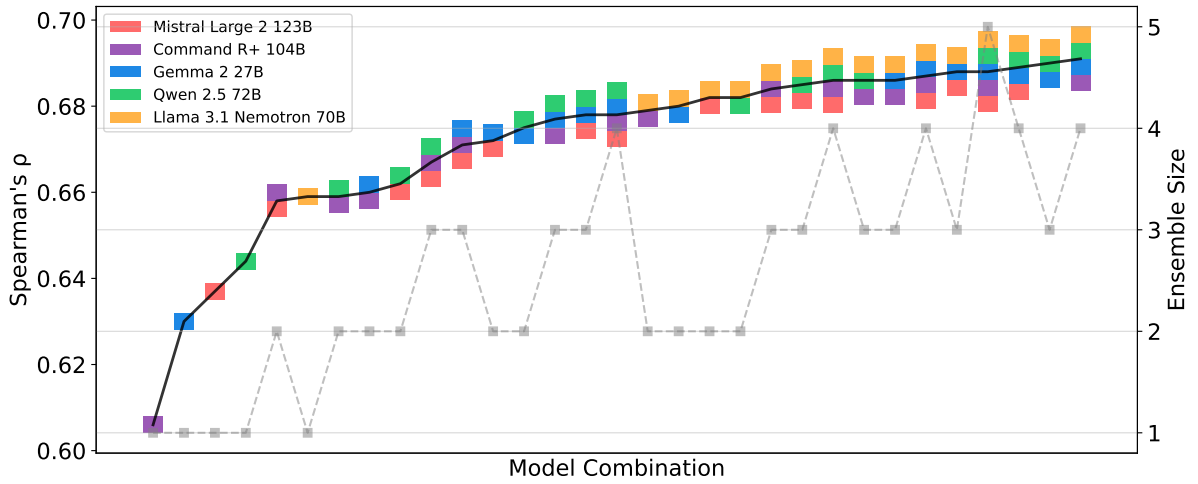

Figure 6: Effect of ensemble size on Spearman's ρ correlations with human scores for the QAGS dataset.

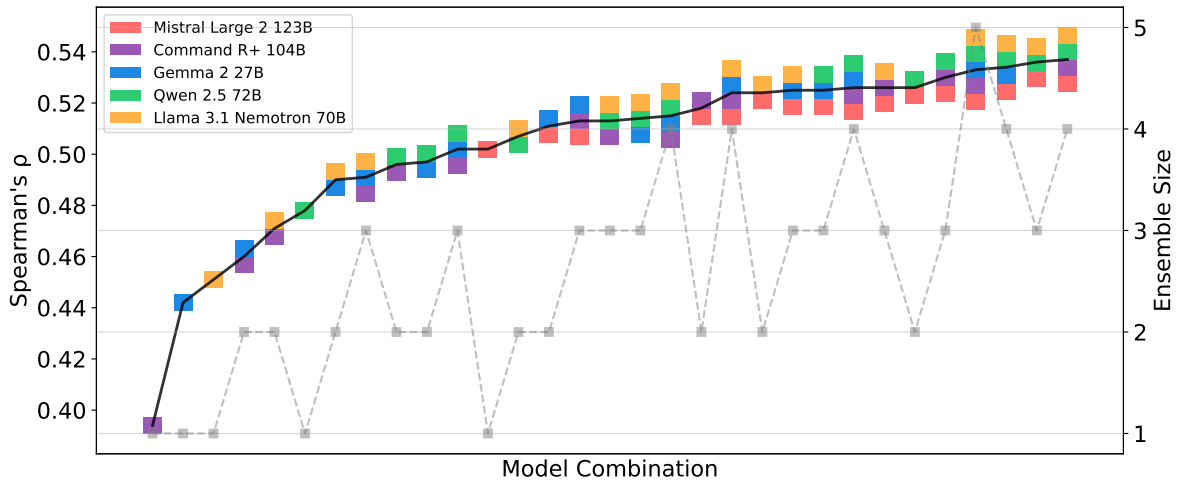

Figure 7: Effect of ensemble size on Spearman's ρ correlations with human scores for the SummEval dataset.

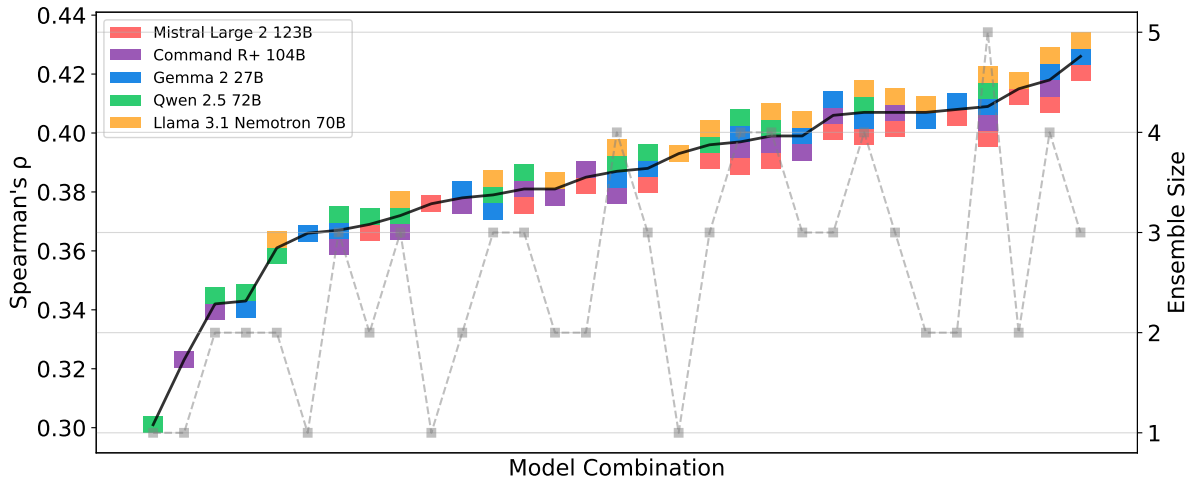

Figure 8: Effect of ensemble size on Spearman's ρ correlations with human scores for the HANNA dataset.

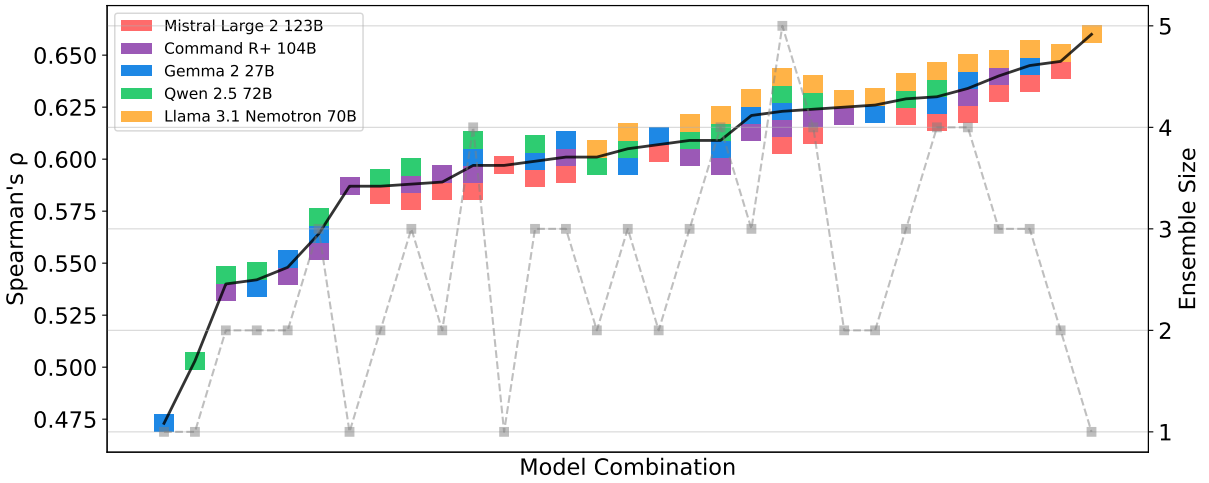

Figure 9: Effect of ensemble size on Spearman's ρ correlations with human scores for the TopicalChat dataset.

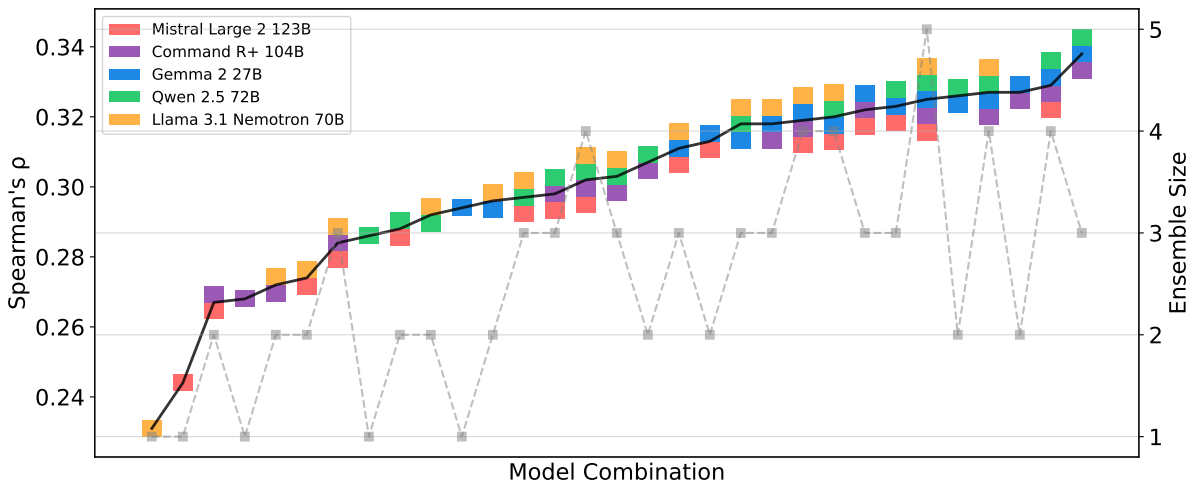

Figure 10: Effect of ensemble size on Spearman's ρ correlations with human scores for the SFRES/SFHOT dataset.

Figure 11: Results for Ablation 1 on QAGS and TopicalChat. The LLMs are instructed to provide both integer overall scores (1–5) and integer severity levels (1–5). The plotted values represent differences in Spearman's ρ correlations with human scores between the original prompt and the ablation. For TopicalChat, **Coh.** = coherence, **Eng.** = engagingness, **Gro.** = groundedness, **Nat.** = naturalness, **Avg.** = average of all aspects.

Figure 12: Results for Ablation 2 on QAGS and TopicalChat. The LLMs are instructed to provide integer overall scores (1–5), and categorical severity levels on the following scale: *Neutral*, *Minor*, *Moderate*, *Major*, *Critical*.

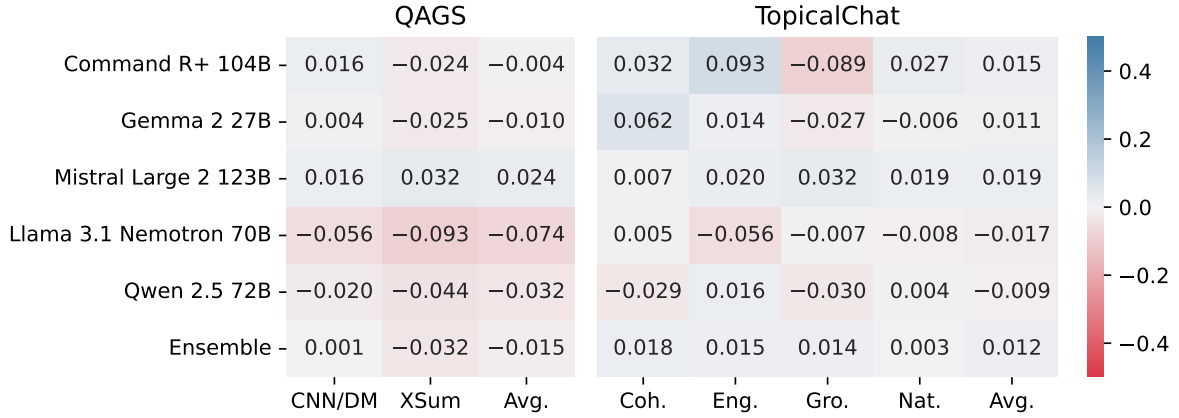

Figure 13: Results for Ablation 3 on QAGS and TopicalChat. The LLMs are instructed to provide categorical overall scores on the scale described in Section 4, and categorical severity levels on the following scale: *Neutral*, *Minor*, *Moderate*, *Major*, *Critical*.

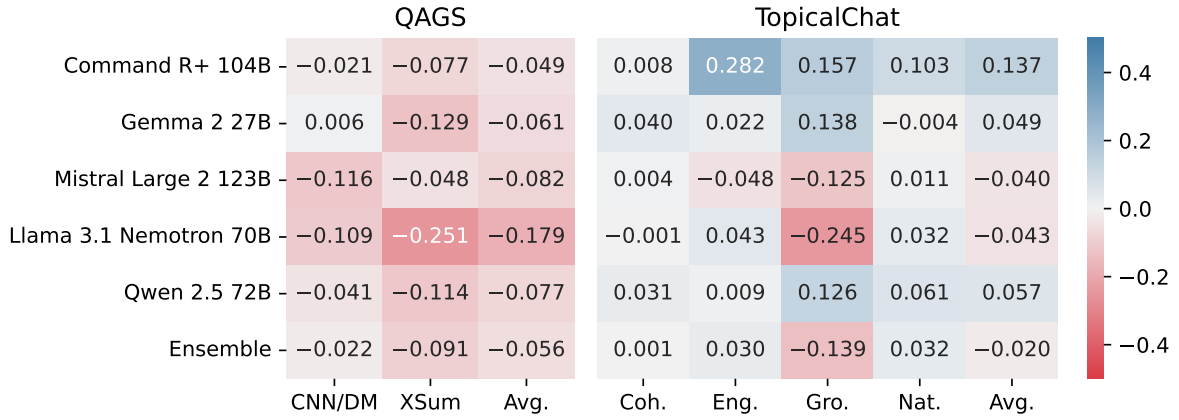

Figure 14: Results for Ablation 4 on QAGS and TopicalChat, where the rules section is removed from the prompt.

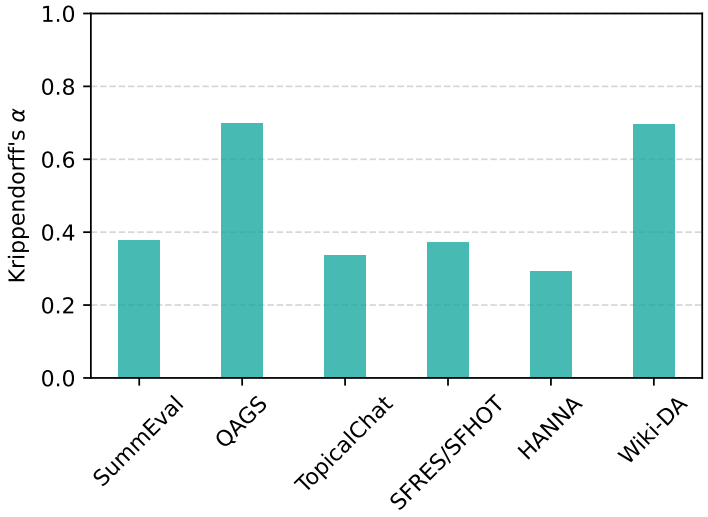

Figure 15: Inter-annotator agreement (Krippendorff's α) between all LLMs in the ensemble for all meta-evaluation datasets. For each dataset, the coefficient is computed over all evaluation aspects.

Model/Method	QAGS	SummEval	TopicalChat	SFRES/SFHOT	HANNA	Wiki-DA
Command R+ 104B	0.681	0.394	0.332	0.266	0.323	0.681
Gemma 2 27B	0.643	0.442	0.481	0.295	0.366	0.767
Llama 3.1 Nemotron 70B	0.669	0.451	0.660	0.179	0.393	0.752
Mistral Large 2 123B	0.645	0.502	0.598	0.249	0.376	0.769
Qwen 2.5 72B	0.651	0.478	0.496	0.283	0.301	0.783
Average	**0.688**	0.533	**0.652**	**0.299**	0.409	**0.837**
Average w/o outliers	0.677	**0.538**	0.623	0.289	**0.411**	0.825
Majority vote	0.654	0.504	0.556	0.290	0.381	0.785
Median	0.668	0.509	0.594	0.296	0.401	0.803
Min	0.641	0.467	0.622	0.265	0.389	0.776

Table 21: Segment-level Spearman (ρ) correlations of different score aggregation methods. For each dataset, correlations are averaged across aspects. Individual LLMs are included for comparison. Best *ensemble* results for each dataset are highlighted in bold.

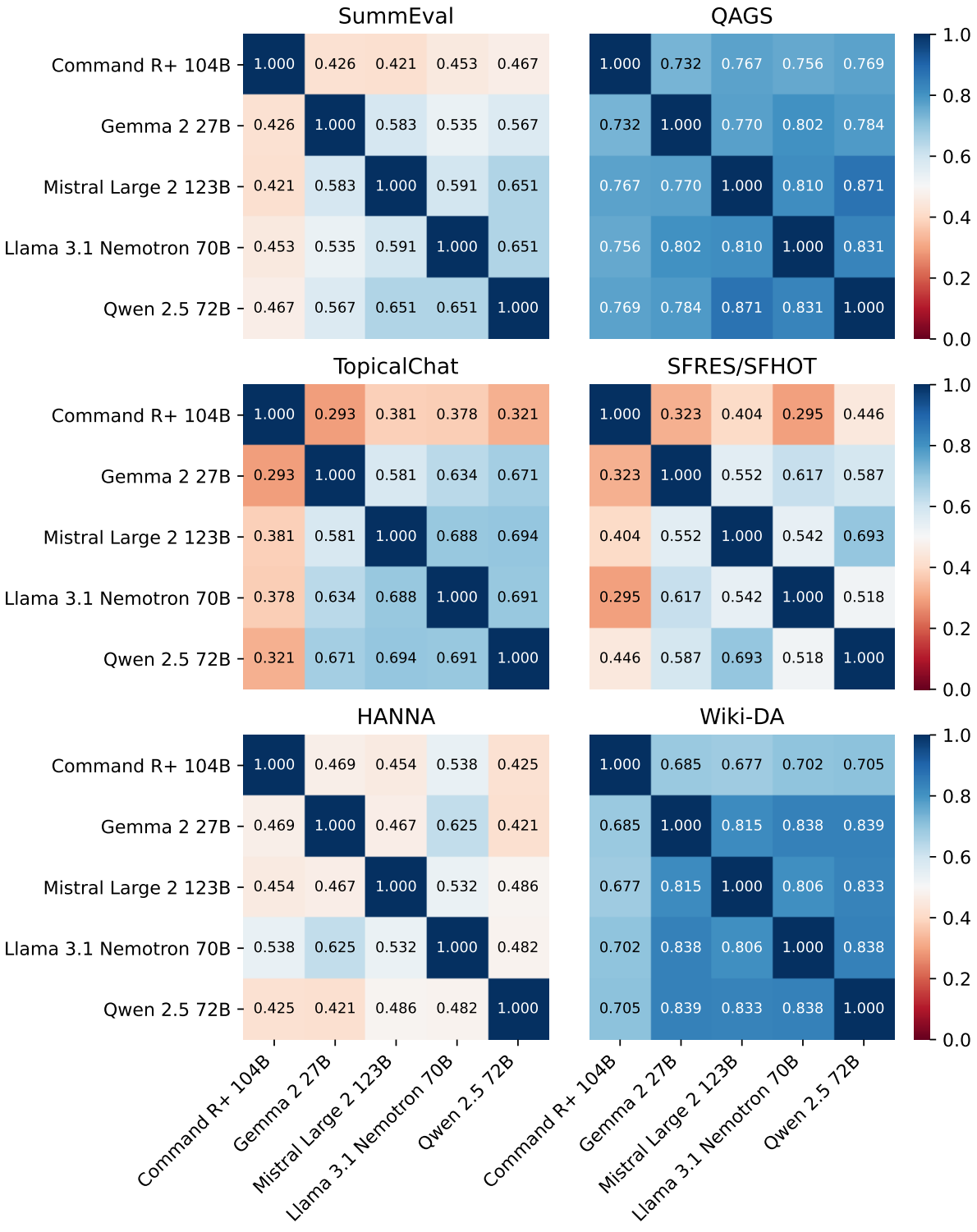

Figure 16: Pairwise Spearman (ρ) correlations of individual LLM scores for all meta-evaluation datasets. For each dataset, the correlations are computed over all evaluation aspects.

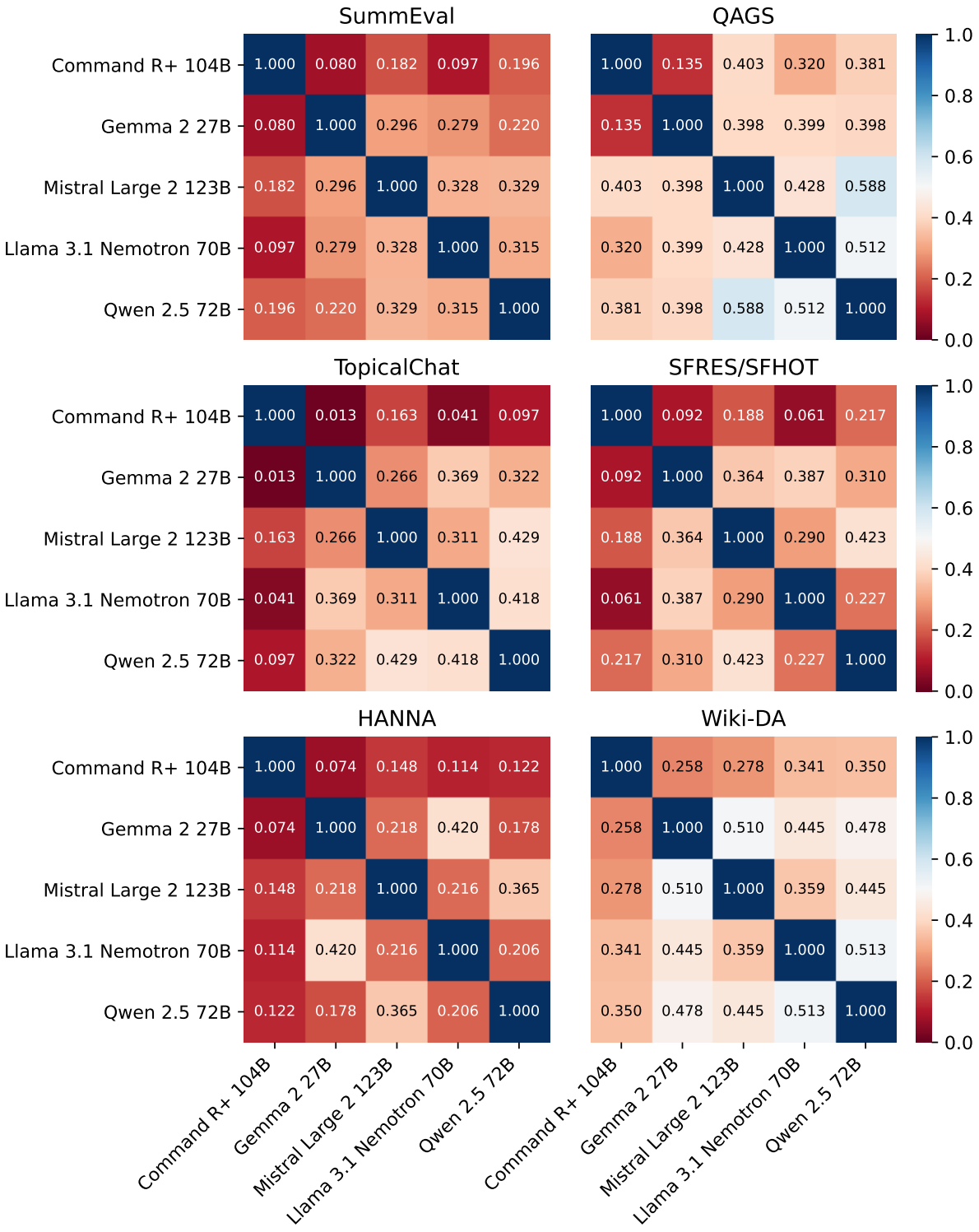

Figure 17: Pairwise inter-annotator agreements (Cohen's κ) of individual LLM scores for all meta-evaluation datasets. For each dataset, the coefficients are computed over all evaluation aspects.

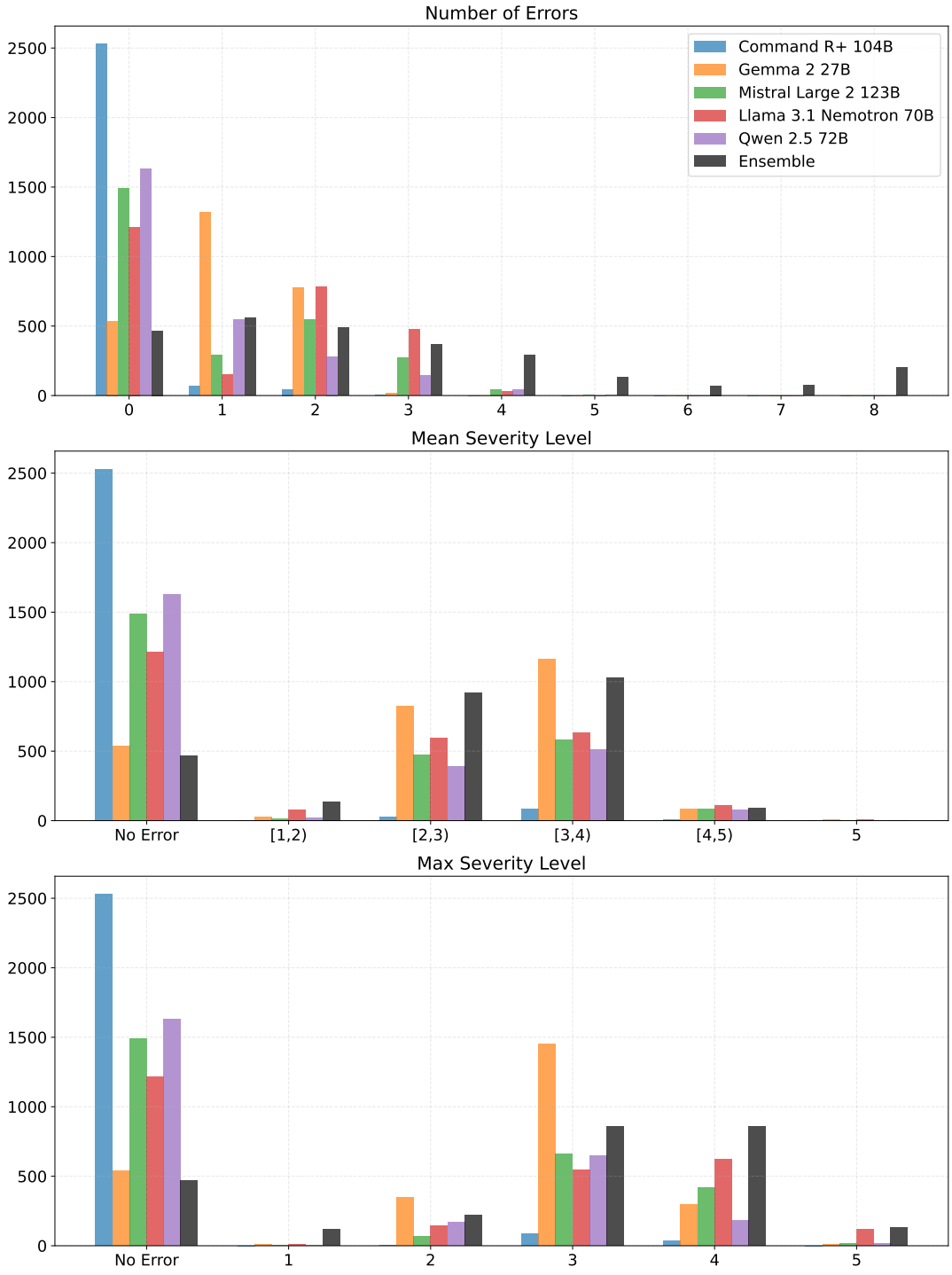

Figure 18: Distribution errors detected by the ensemble LLMs in outputs rated with maximum score by human annotators in SummEval. **Top:** Frequencies of numbers of detected errors per evaluated output. **Middle:** Frequencies of mean severity levels assigned to detected error per output. Values larger than 0 are binned to ranges $[a, b)$, where $0 < a <= 5$ and $b = a + 1$. **Bottom:** Frequencies of maximum severity levels assigned to detected errors per output.

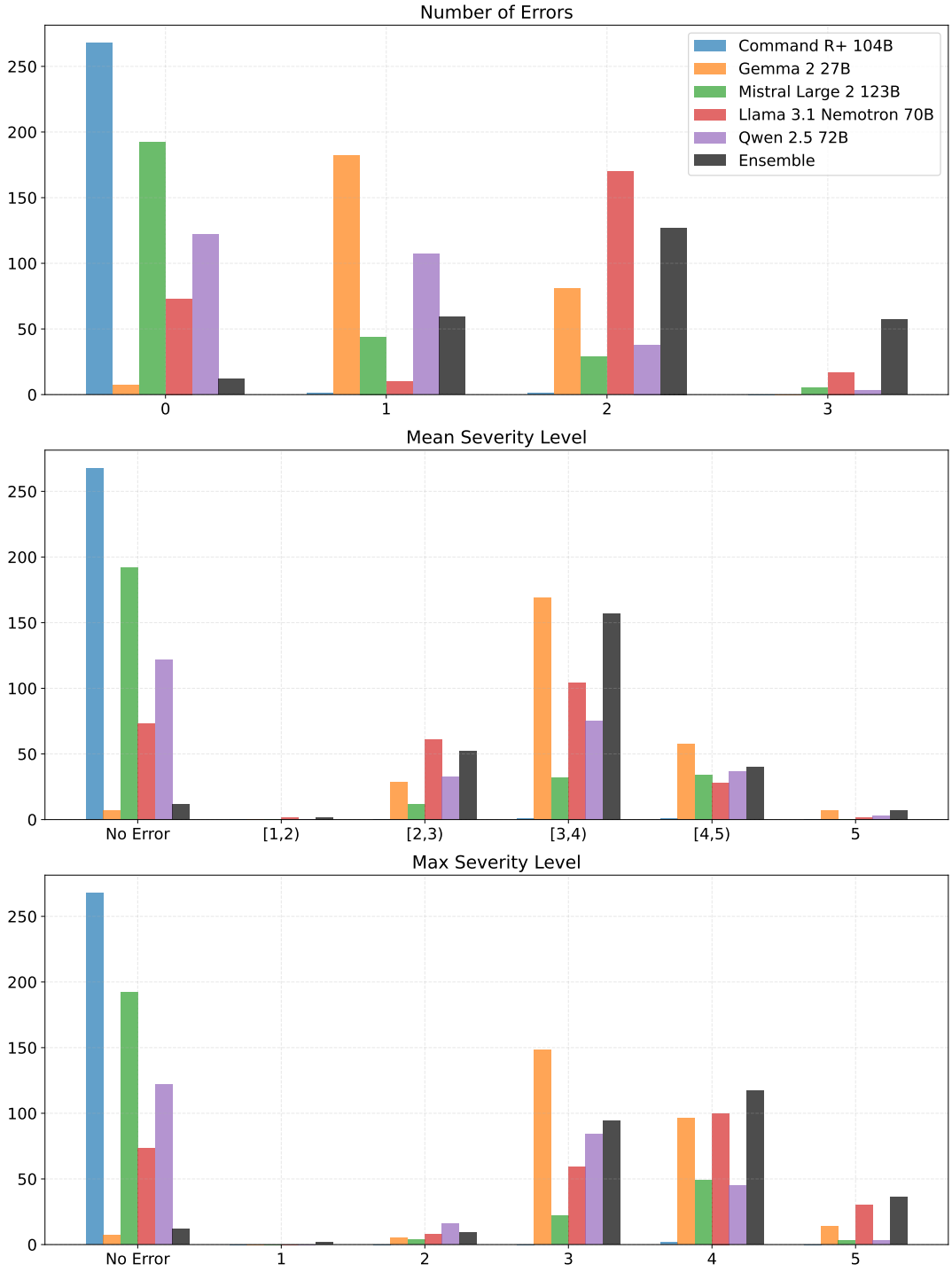

Figure 19: Distribution errors detected by the ensemble LLMs in outputs rated with maximum score by human annotators in TopicalChat. Note that groundedness evaluations are excluded from the analysis, as the dataset contains only binary ratings for this aspect. **Top:** Frequencies of numbers of detected errors per evaluated output. **Middle:** Frequencies of mean severity levels assigned to detected error per output. Values larger than 0 are binned to ranges $[a, b)$, where $0 < a <= 5$ and $b = a + 1$. **Bottom:** Frequencies of maximum severity levels assigned to detected errors per output.

Data:
`Adolfo_Suárez_Madrid–Barajas_Airport | operatingOrganisation | ENAIRE`
`Adolfo_Suárez_Madrid–Barajas_Airport | runwayLength | 3500.0`
`Adolfo_Suárez_Madrid–Barajas_Airport | elevationAboveTheSeaLevel | 610.0`
`Adolfo_Suárez_Madrid–Barajas_Airport | runwayName | "14L/32R"`

Text: Adolfo Suárez Madrid–Barajas Airport is a major international airport located near the heart of the city[1], offering a wide range of air services[2].

Evaluation aspect: *Faithfulness* – Extent to which the information in the text is supported by the data

[1] Explanation: This phrase is considered an error because the provided data does not mention the airport's size (major/international), location relative to the city (near the heart), or any information that could imply these details. The text introduces unsupported information, negatively affecting faithfulness.
Severity: 4

[2] Explanation: Similar to Error 1, this phrase introduces information not present in the provided data. There is no mention of the types or variety of air services offered by the airport, making this an unsupported claim that affects faithfulness.
Severity: 3

Overall score: Poor (2/5)

Explanation of the score: The generated text contains significant errors related to faithfulness, as it introduces substantial information not supported by the provided data. While the text does not contradict any data points (since it doesn't directly address them), its lack of adherence to the given facts and introduction of unsubstantiated claims warrant a 'Poor' rating. For an 'Excellent' score, the text should have strictly adhered to the provided data, which includes specific details about the airport's operating organisation, runway length, elevation, and runway name, none of which were accurately represented or utilized in the generated text.

Figure 20: Example error span annotation provided by OpenNLGauge for the RDF-to-text task. The generated text is evaluated for *faithfulness* with respect to the input data.

Emotional Context: A is guilty
Dialogue History:
A: One time I skipped my friend's birthday party to watch TV
B: I hope something good was on
A: There was! But I felt so bad afterwards. I'm a monster

Response: It's okay[1], everyone makes mistakes sometimes. Next time, try to find a balance between social commitments and personal enjoyment[2].!

Evaluation aspect: *Empathy* – Extent to which the response shows understanding of the feelings of the person talking about their experience

[1] Explanation: While the intention is to comfort, the phrase "It's okay" might come across as dismissive of A's expressed guilt and remorse, potentially undermining their emotional experience. A more empathetic approach could acknowledge the negative feelings before offering reassurance.
Severity: 2

[2] Explanation: This part of the response shifts focus from acknowledging and validating A's current emotional state (guilt) to providing advice for future actions. While helpful, it doesn't directly address or show understanding of A's feelings about their past action, slightly lacking in empathy.
Severity: 3

Overall score: Fair (3/5)

Explanation of the score: The response demonstrates some level of empathy by attempting to reassure and offer a solution, indicating an understanding that mistakes happen. However, it falls short in fully acknowledging and validating the speaker's emotional experience (guilt), hence not scoring higher than 'Fair'. To achieve a 'Good' or 'Excellent' score, the response would need to more directly address and empathize with A's feelings of guilt and self-perceived monstrosity.

Figure 21: Example error span annotation provided by OpenNLGauge for the dialogue response generation task. The response is evaluated for *empathy*. In addition to dialogue history, the input also includes the emotional context of the person talking about their experience.

Source: *[NOT PROVIDED]*
Summary: By A new study that college are more likely to engage in energy - binge drinking when alcohol than non - .[1]

Evaluation aspect: *Fluency* – Formal quality of individual sentences of the summary. A fluent sentence should be grammatical, natural and easy to understand.

[1] Explanation: The sentence is fragmented and lacks proper structure. It starts with an unnecessary preposition ("By") and contains several grammatical errors, such as missing words and incorrect phrasing. This makes the sentence difficult to understand and unnatural.
Severity: 5

Overall score: Unacceptable (1/5)

Explanation of the score: The summary is very disfluent due to significant grammatical errors and lack of coherence, making it extremely difficult to comprehend.

Figure 22: Example error span annotation provided by OpenNLGauge for the summarization task. The summary is evaluated for *fluency*. Note that source text is not included, as it is not relevant for evaluation of fluency.