

# Order in the Evaluation Court: A Critical Analysis of NLG Evaluation Trends

Jing Yang<sup>1,2</sup> Nils Feldhus<sup>1,2</sup> Salar Mohtaj<sup>3</sup>  
 Leonhard Hennig<sup>3</sup> Qianli Wang<sup>1,3</sup> Eleni Metheniti<sup>4</sup>  
 Sherzod Hakimov<sup>5</sup> Charlott Jakob<sup>1</sup> Veronika Solopova<sup>1,3</sup>  
 Konrad Rieck<sup>1,2</sup> David Schlangen<sup>3,5</sup> Sebastian Möller<sup>1,2,3</sup> Vera Schmitt<sup>1,3,6</sup>

<sup>1</sup>Technische Universität Berlin

<sup>2</sup>BIFOLD – Berlin Institute for the Foundations of Learning and Data

<sup>3</sup>German Research Center for Artificial Intelligence (DFKI), Berlin

<sup>4</sup>ANITI

<sup>5</sup>University of Potsdam

<sup>6</sup>CERTAIN

## Abstract

Despite advances in Natural Language Generation (NLG), evaluation remains challenging. Although various new metrics and LLM-as-a-judge (LaaJ) methods are proposed, human judgment persists as the gold standard. To systematically review how NLG evaluation has evolved, we employ an automatic information extraction scheme to gather key information from NLG papers, focusing on different evaluation methods (metrics, LaaJ and human evaluation). With extracted metadata from 14,171 papers across four major conferences (ACL, EMNLP, NAACL, and INLG) over the past six years, we reveal several critical findings: (1) Task Divergence: While Dialogue Generation demonstrates a rapid shift toward LaaJ (>40% in 2025), Machine Translation remains locked into n-gram metrics, and Question Answering exhibits a substantial decline in the proportion of studies conducting human evaluation. (2) Metric Inertia: Despite the development of semantic metrics, general-purpose metrics (e.g., BLEU, ROUGE) continue to be widely used across tasks without empirical justification, often lacking the discriminative power to distinguish between specific quality criteria. (3) Human-LaaJ Divergence: Our association analysis challenges the assumption that LLMs act as mere proxies for humans; LaaJ and human evaluations prioritize very different signals, and explicit validation is scarce (<8% of papers comparing the two), with only moderate to low correlation. Based on these observations, we derive practical recommendations to improve the rigor of future NLG evaluation.

## 1 Introduction

Over the past few years, significant developments have taken place in NLG. While the previous generation of models focused on a limited range of tasks like summarization and translation, the dominance of transformer-based Large Language Models (LLMs) has enabled substantially more flexi-

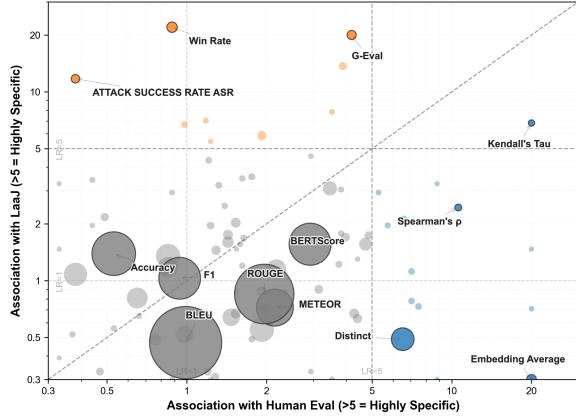


Figure 1: The fractured landscape of NLG evaluation, size of dots indicates usage frequency. Metrics are mapped by Likelihood Ratio (LR) to Human Eval ( $x$ ) vs. LaaJ ( $y$ ), in log-scale. The bottom left gray bubbles reveal “metric inertia” around generic metrics like BLEU. In contrast, highly specific metrics ( $LR > 5$ ) diverge: orange bubbles highlight metrics associated with highly LaaJ and blue bubbles metrics with highly human associations, indicating that LaaJ is not a direct proxy for human judgment.

ble applications, ranging from poetry to medical report generation. This shift has reshaped evaluation paradigms and introduced new challenges in evaluation, such as dealing with confabulations, ensuring factual consistency, and mitigating biases (Gehrmann et al., 2023). Consequently, metrics have evolved beyond simple co-occurrence. Newer semantic-aware metrics have emerged, including BERTScore (leveraging contextual embeddings) and BLEURT (fine-tuned on human judgments), and most notably, the surge of reference-free, prompt-based LaaJ methods (Figure 4 (b)) (Gao et al., 2025; Gu et al., 2025; Li et al., 2024). However, the adoption of these new methods has created a fragmented evaluation landscape. Figure 1 illustrates this complexity by visualizing the likelihood ratios of various metrics with LaaJ versus human evaluation. The distinct spread of metrics suggests that LaaJ and human evaluation are

not merely interchangeable validation steps but appear together with different metrics. Despite the exponential rise in LaaJ popularity since 2023, validating these automated judges remains a critical bottleneck, as human validation is not often performed.

Motivated by the need to better understand how evaluation is conducted in NLG research, we employ LLMs as research tools (Liao et al., 2025) to analyze the literature. This approach is critical to bridge the gap between scale and depth: it allows us to process a volume of papers impossible for manual review while extracting structured nuances that keyword search would miss. To the best of our knowledge, we present the largest study of its kind, applying multi-LLM extraction to 14,171 papers from four major NLP conferences. Unlike prior surveys that taxonomize evaluation methods (Gao et al., 2025) or review LaaJ techniques (Li et al., 2024), our work quantifies exactly how the community is (or is not) shifting its evaluation practices over time.

Our contributions are summarized as follows:

- We present the largest quantitative analysis of NLG evaluation to date, applying multi-LLM information extraction to 14,171 papers.
- We empirically identify two critical failures in current practice: “metric inertia”, where outdated metrics are misapplied to emerging tasks, and a validation gap, where the exponential rise of LaaJ is not matched by human validation.
- We demonstrate that LaaJ and human evaluators prioritize different signals, and provide actionable recommendations to better align automated evaluation with human judgment.

## 2 Related Work

In this section, we review some of the recent activities on using LLMs as a data annotator for scientific research and relevant surveys for NLG evaluation.

**LLMs for Annotation of Research Papers** An emerging trend in LLMs is using them for data annotation to reduce human labor. Tan et al. (2024) point out that LLMs can reliably annotate instructions and responses, even for specific domains, with human-level quality. However, LLMs are dependent on prompting, structure, and filtering to achieve good and trustworthy annotations. Only very few prior works have dealt with the automated processing of scientific literature (Agarwal et al., 2025; Du et al., 2024; Scherbakov et al., 2025). Ex-

isting survey papers making use of LLMs as annotators typically rely on prompts that assess whether a paper is relevant to the target topic (Alabi et al., 2025; Alyafeai et al., 2025) or simply collect metadata like citation counts (Bernasconi et al., 2025). Unlike previous work, we provide the first work to annotate research papers systematically on a large-scale, not only for paper’s relevance filtering, but for quantifying research trends and providing critical overviews.

**Surveys in NLG evaluation** Rapid progress in deep learning for NLG has spurred numerous new automatic evaluation metrics, leading to dedicated surveys. Sai et al. (2022) provide a taxonomy of NLG metrics and highlight that traditional reference-based metrics often correlate poorly with human judgment and fail to capture nuances like factual correctness. With the rise of LLMs, researchers have begun using them as evaluators (i.e., LaaJ) to assess generated text quality. This marks a new direction in NLG evaluation that earlier surveys did not address (Celikyilmaz et al., 2021; Sai et al., 2022). Recent surveys explicitly taxonomize these methods and highlight reliability challenges (Gao et al., 2025; Gu et al., 2025). While LaaJ can replicate human judgments on certain criteria, it exhibits substantial variability across tasks (Bavaresco et al., 2025) and often struggles with generalization (Huang et al., 2025) and reliability (Hu et al., 2024; Wang et al., 2024; Lee et al., 2025; Wang et al., 2025) compared to proprietary models. While humans remain the most reliable and trusted evaluators, several meta-analyses have highlighted the inconsistency in human evaluation protocols for NLG and the need for standardized guidelines (van der Lee et al., 2019; Howcroft et al., 2020; Celikyilmaz et al., 2021).

## 3 Paper Annotation

We include main conference papers from the ACL Anthology from 2020 to 2025, focusing on four conferences: ACL, EMNLP, NAACL, and INLG in consideration of the advancement of language models. In total, 14,171 papers were collected (full statistics in Table 12 of Appendix C). We can clearly see a surge of publications in 2025, with ACL 2025 nearly doubling the amount from ACL 2024. In the next part of this Section, we describe our paper annotation process as shown in Figure 2.

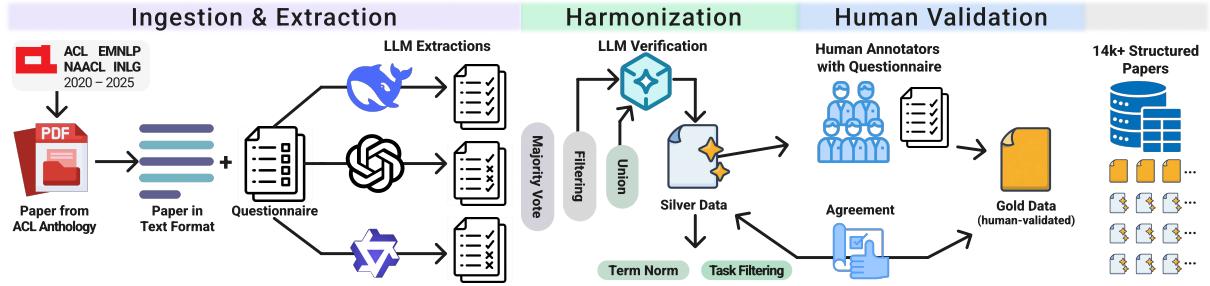


Figure 2: Our paper annotation pipeline (§3), including converting PDF to text (§3.1), extraction of metadata based on the NLG evaluation questionnaire (Table 9), harmonization based on majority vote, filtering, and merging (§3.2), and human validation (§3.4), yielding 14k structured papers including a subset of 110 papers with gold annotations.

### 3.1 LLM Annotation

We extract the full text of each paper from PDF format, using GROBID<sup>1</sup>. We then feed the full text (excluding title and abstract, as the full text should contain all information we need) into a model with our designed prompt (Appendix D.1).

For each paper, we prompted an LLM acting as an “expert NLP researcher with deep experience in Natural Language Generation (NLG)” to read the full text and return a structured JSON object answering four binary questions about NLG-related information:

1. Does the paper address NLG tasks?
2. Does the paper use automatic metrics to evaluate the generated outputs?
3. Does the paper use large language models (LLMs) as judges (i.e., after generation, LLMs are used to assess the outputs)?
4. Does the paper conduct human evaluations of the generated outputs?

Each question  $i$  is represented as a JSON object  $\text{answer}_i$  with a required binary field “answer” (“Yes” or “No”). If “answer” is “Yes”, we ask the LLM to extract further metadata, e.g., a verbatim excerpt as evidence, lists of datasets, languages, models, and criteria. (The full metadata list is provided in Appendix Table 9.) If the “answer” is “No”, all other fields in that section are set to empty strings or empty lists.

We categorize and define the evaluation methods by (1) automatic metrics: metrics used to measure generated text quality by providing a score; (2) LaaJ: evaluating generated text with LLM prompting, producing textual evaluation and a rating score; and (3) Human evaluation: evaluation produced by humans (either by experts or crowd-workers).

**Model Selection** Extracting using a single LLM may not be sufficiently reliable. Therefore, we performed the extraction using three different open-source models: DeepSeek-R1 (DeepSeek-AI et al., 2025), GPT-OSS-120B (OpenAI et al., 2025) and Qwen3-235B-A22B-Instruct (Yang et al., 2025).

### 3.2 Harmonization of LLM-based Extraction

With results from three different LLMs, we need to harmonize them before we can perform any analysis. To aggregate the results, we merge the answers to the four binary questions through majority voting to mitigate individual model biases. We filter out papers that have *no* to Answer 1 (A1). Afterwards, we are left with 8,665 initial NLG papers (61%). For these papers, we further employed another LLM (DeepSeek V3.1 Terminus) to verify the extraction with a different prompt. The prompt is designed to verify each extraction against the full paper text to validate the four binary questions, and perform four other actions: normalize, correct, add, and remove the other metadata from the same list for each field listed in Table 9 (see full prompt in Appendix D.2).

- **Normalize** metadata to use canonical forms (for example, all BLEU (bleu, Bleu, BLEU-4, etc.) variance to BLEU)
- **Correct** any incorrect items
- **Add** any missing important items
- **Remove** any irrelevant or incorrect items

**Comparison of LLM annotations** To compare the difference among LLM annotations, we compute the agreement with Krippendorff’s  $\alpha$  (pairwise agreement between the LLM-harmonized one and the others, shown in Table 10 in Appendix). The overall agreements are high for all four answers (A1: 0.7101, A2: 0.6879, A3: 0.8048, and A4: 0.8124).

<sup>1</sup><https://github.com/kermitt2/grobid>

### 3.3 Post-processing Extractions

**Term Normalization** To group term variants representing the same value, we perform two normalization steps: (1) preprocessing and (2) fuzzy matching using SequenceMatcher<sup>2</sup> with a threshold of 0.9. For detailed term normalizations and examples, please check Appendix B.2 and B.3. We show the overall statistics of the unique count of normalized terms in Figure 3 (see exact numbers in Table 12 from Appendix C). All term categories show increasing counts, especially since 2023. Interestingly, the number of languages diminished in 2025; most 2025 papers include datasets that cover multiple languages but do not explicitly enumerate them. Thus, the actual language count shown in Figure 3 is likely underestimated.

**NLG Task Filtering** Our human annotations (§3.4) reveal that most disagreements on A1 occurred for papers that (1) involve generation tasks producing non-natural-language outputs; or (2) are likely classification tasks. Following task normalization and ranking based on their frequencies, we systematically exclude 15 tasks (full task list in Appendix B.4) and keep the top-30 tasks for our analysis after the filtering (3,334 papers).<sup>3</sup>

### 3.4 Human Validation

To further assess the quality of LLM annotations, we performed human annotation on 110 papers. We randomly select these papers identified as NLG-related by three LLMs’ majority voted first answer as “Yes”. We designed our guidelines similarly to the prompt for LLM annotations, with more details and examples (Appendix E). We conducted the annotation process in two rounds using 11 NLP researchers, comprising 3 Master’s students, 2 PhD students, and 6 Postdocs. In the first round, each annotator is assigned to 10 papers. In the second round, papers were shuffled and reassigned to ensure that each paper received two independent annotations. Disagreements were resolved by a third annotator who reviewed the conflicting labels and made the final determination.

We then compare the three LLMs and the harmonized results (§3.2) with the human ground truth (the four binary questions). The pairwise agreement is shown in Table 1. A3 and A4 have higher agreement than A1 and A2, similar to the results in Table 10, indicating that identifying NLG tasks

<sup>2</sup><https://docs.python.org/3/library/difflib.html>

<sup>3</sup>For the list of top-30 tasks, see Figure 8 in Appendix C.

and the usage of automatic metrics is harder than identifying LaAJ and human evaluation. LLM-harmonized results have overall the highest agreement with final human ground truth answers, showing the advantage of harmonization.

Pair	A1	A2	A3	A4
Human R1 vs. Human R2	81.82	80.00	89.09	91.82
Human vs. DeepSeek-R1	74.55	73.64	95.45	<b>92.73</b>
Human vs. GPT-OSS-120B	75.45	<b>75.45</b>	90.91	90.91
Human vs. Qwen3-235B	75.45	70.00	93.64	85.45
Human vs. Majority Voting	74.55	71.82	93.64	88.18
Human vs. LLM-harmonized	<b>77.27</b>	74.55	<b>95.45</b>	90.00

Table 1: Pairwise agreement (%) between human annotations and LLMs.

## 4 Results and Analysis

In this section, we organize the extracted results of the filtered NLG papers. To quantitatively analyze them, we use two measures Frequency ( $P(B|A)$ ) and Likelihood Ratio ( $LR(A \rightarrow B)$ ) as follows: Given terms  $A$  and  $B$ ,

$$P(B|A) = \frac{\text{Number of papers with } A \text{ and } B}{\text{Number of papers with } A} \quad (1)$$

$$LR(A \rightarrow B) = \frac{P(B|A)}{P(B|\neg A)} \quad (2)$$

Frequency represents empirical conditional probability: specifically, the proportion of papers containing both  $A$  (e.g.: metric BLEU) and  $B$  (e.g.: task MT). LR represents discrimination power: how much more likely term  $B$  is to appear in papers containing  $A$  than in papers not containing  $A$ . Terms are tasks, metrics, or LaAJ and human evaluation criteria (see Appendix B.5 for more specific definitions). A high LR means term  $A$  (e.g.: metric BLEU) is more associated with  $B$  (e.g.: task MT) than the rest of the corpus from the same term category as  $A$  (e.g.: DG).

### 4.1 Temporal Trend Analysis

We focus on the four most studied tasks, which collectively represent 78.1% of all filtered papers (see Figure 8): Dialogue Generation (DG; 871, 26.1%), Machine Translation (MT; 847, 25.4%), Text Summarization (TS; 734, 22.0%), and Question Answering (QA; 462, 13.9%). We illustrate the evolution of evaluation practices for these tasks in Figure 4. MT, despite being a traditional NLG

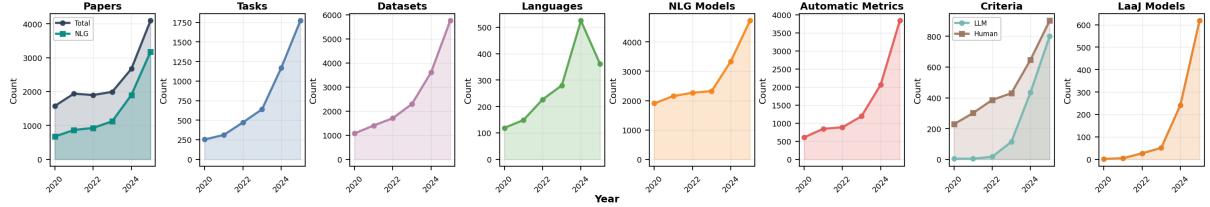


Figure 3: Distribution of different terms across years, numbers are counted are after normalization. The only abnormal trend is the number of languages, which decreased from 2024-2025.

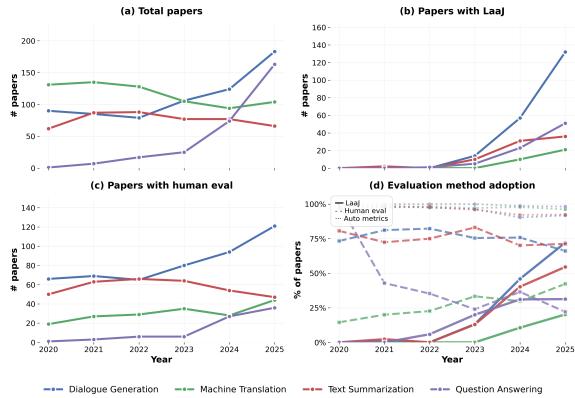


Figure 4: Distribution of papers (2020 – 2025) with different evaluation methods across the top-four tasks.

task, has been less frequently addressed over the years compared to tasks such as DG and QA (a). Across tasks, near 100% paper uses some type of automatic metrics. Comparing LaaJ and human evaluation, LaaJ has seen increasing adoption since 2023 (b), especially for DG, while human evaluation prevalence stays relatively stable except in QA, where it has reduced significantly since 2021. Notably, out of all MT papers, only 182 papers (26%) conducted human evaluation and LaaJ is not commonly used yet (31 papers), mostly only in 2025. In contrast to MT, TS and DG have much higher percentage of human evaluation, while decreasing from 2024 to 2025 when more LaaJ is being used.

We further examine each of the top four tasks and discuss how the evaluation methods differ. We restrict the analysis to single-task papers to find task-specific patterns, and visualize the top 10 most frequent metrics and criteria for that year. Figure 5 visualizes the evolution of evaluation methods across four major tasks. Each chart illustrates the rank of task association with evaluation metrics, human evaluation criteria, or LaaJ criteria over six years (2020-2025), and cycle size indicates their frequency.

**Dialogue Generation** The DG task has shifted from simple request fulfillment to complex interactions, therefore the metrics and criteria are

also more diverse. The most frequent metric is BLEU, although it has a low LR with DG. The top-ranked metric, Distinct, which measures diversity, is decreasing in frequency, while METEOR and BERTScore have a surging trend in LR. Performance-related measures such as *F1*, *Accuracy* and *Recall* have experienced a large drop in rank, likely due to increasing usage in other tasks and also overall better performance in DG. Throughout the years, the top-ranked criteria are *Empathy* and *Appropriateness*, even though their frequency is not as high as *Fluency*, *Relevance*, and *Coherence*. Comparing LaaJ with human evaluation, while human criteria focus on *Appropriateness* and *Naturalness*, LaaJ prioritizes alignment-focused criteria such as *Helpfulness* and *Harmlessness*. This indicates that LaaJ is introducing orthogonal safety checks, which reveals a gap in human evaluation.

**Machine Translation** MT primarily uses the BLEU metric (87.4%) for evaluation, with COMET also established as a major metric. The latter also has the highest association with MT since 2023, while BLEU has rather low association (since it is also heavily used in other tasks) and is decreasing in rank. In terms of evaluation criteria, translation (though notably vague) and adequacy are ranked high in association (meaning the criterion is unique to the MT task). The dominance of these n-gram and reference-based metrics illustrates deep-seated inertia. Despite the availability of LLMs, the LaaJ Criteria row is sparse, with almost no significant activity until 2024.

**Text Summarization** In TS, ROUGE is the dominant automatic evaluation metric, followed by BERTScore, but neither has the highest association with the task. FactCC was common during 2021-2023, then dropped out, while MoverScore saw an increasing use and reached the top spot in 2025 – similar to BARTScore. BLEU dropped a lot in the rank but rose again in 2024/25. In terms of criteria, *Redundancy* and *Coverage* are highly

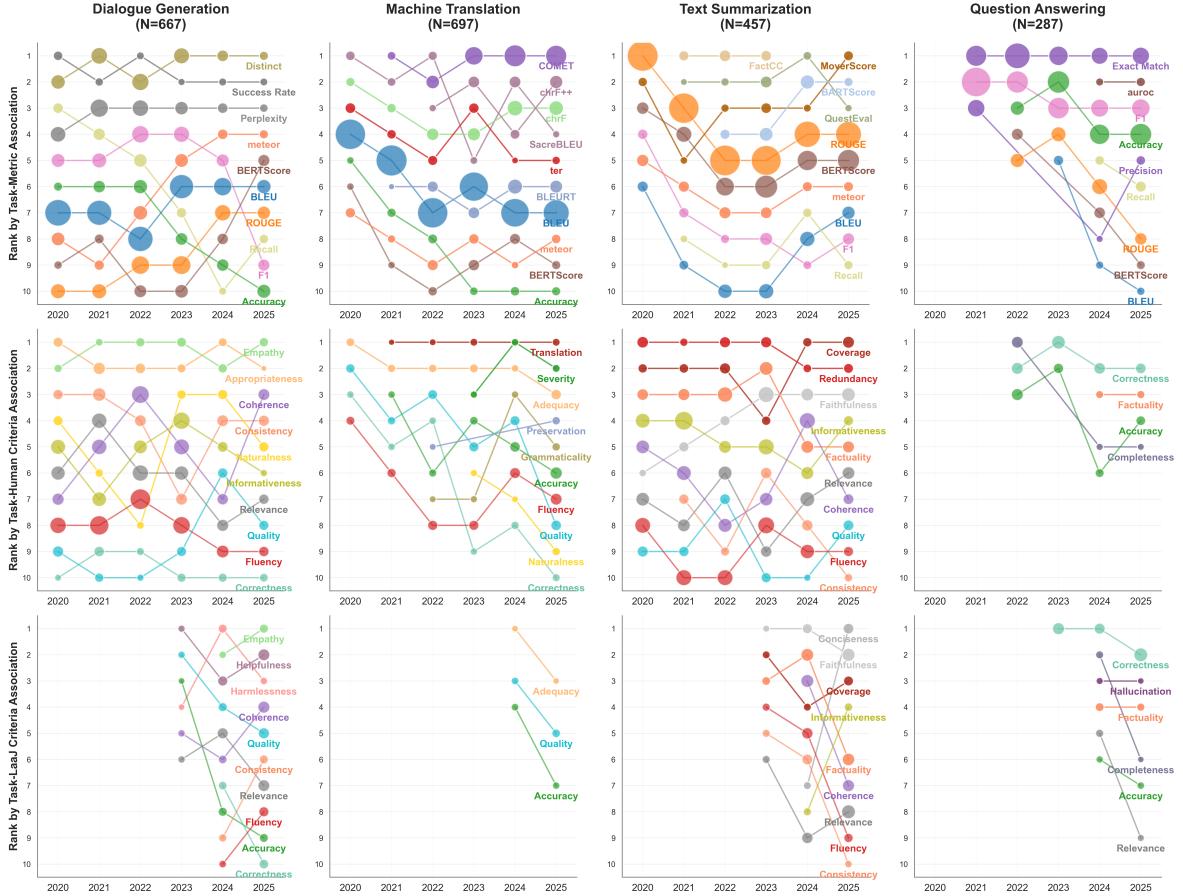


Figure 5: Bump charts of task-specific NLG evaluation trends across four tasks (left-to-right) and three paradigms (automatic, human, LaaJ) (top-down). The metrics and criteria are ranked by their task association, and size indicates the frequency of that metric among papers of that task. Only top-10 most frequent metrics and criteria are included.

associated with TS. *Faithfulness* is increasingly important, while *Factuality* decreases. *Relevance*, *Fluency*, and *Coherence*, while frequently used, are not strongly associated with TS. We also see a trend that human and LaaJ criteria ranks are more aligned in 2025, compared to 2024.

**Question Answering** As a QA task usually generates relatively short answers, its evaluation also differs largely. This task often resembles classification, where the answers are directly compared to the ground truth via exact match. Compared to the other three tasks, QA has the least amount of human evaluation. When these two evaluations are employed, the main criterion is *correctness* (especially with LaaJ). This is concerning: while QA exhibits the largest growth, its evaluation is almost entirely based on string matching.

## 4.2 Metric Association by Evaluation Methods

We quantify the associations between automatic metrics and other evaluation methods (human eval-

uation vs. LaaJ) with LR defined in Equation (2). We restrict the analysis to metrics used in more than 10 papers.

Figure 1 shows how automatic metrics associate with human evaluation compared to LaaJ evaluation. The scatter plot reveals distinct clustering patterns that reflect different evaluation associations. Most frequently used metrics cluster at the bottom-left corner (BLEU, COMET,  $F_1$ , Accuracy, etc.), exhibiting low associations with both LaaJ and humans, indicating generic and independent usage. The top-left part displays metrics that often appear in papers performing LaaJ evaluation (Win Rate, ASR, AlignScore). The bottom-right side shows metrics that commonly appear in papers employing human evaluation (Embedding, FACTCC, DISTINCT, etc.). Very few metrics have high association rates with both LaaJ and human evaluation (top-right corner). The correlation between association rates with LaaJ and human is nearly zero (Spearman  $\rho = 0.007$ ), indicating orthogonal evaluation methods.

We further investigate the associations between

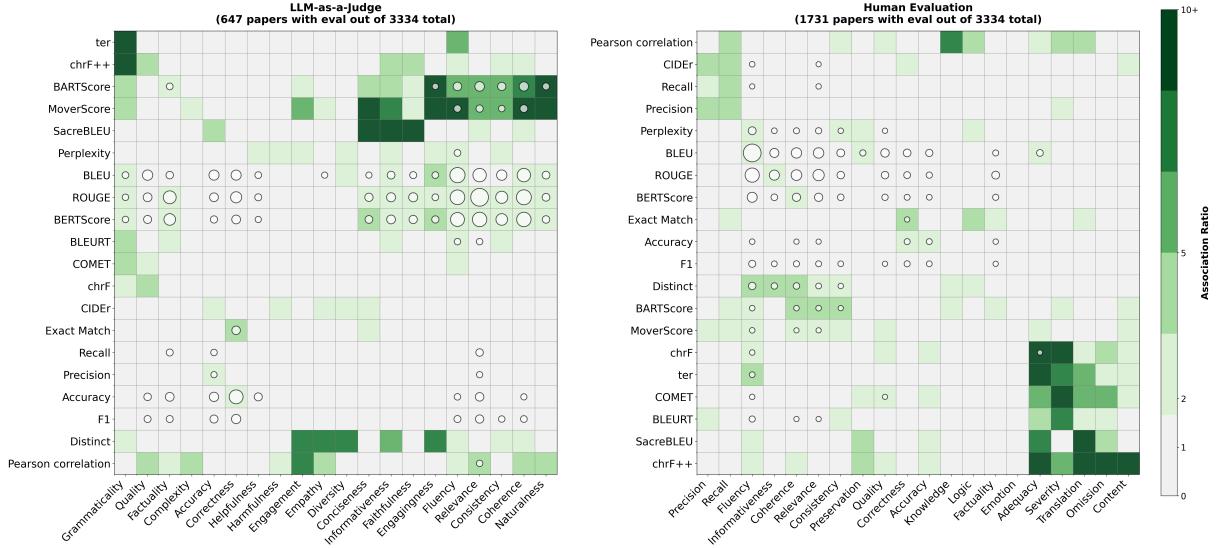


Figure 6: Metric-criteria association heatmaps for LaaJ (left) and human evaluation (right). Stronger color indicates higher LR, and bigger cycle size indicates higher frequency. Most frequent metrics and criteria have low LRs.

automatic metrics and evaluation criteria. To examine which automatic metrics appear together with different evaluation criteria, we compute frequencies and LR for metric-criterion pairs<sup>4</sup>. Figure 6 presents frequencies and LR of metric-criteria pairs from papers with LaaJ or human evaluations. We selected the ones with either high frequent or LR, and cluster similar rows/columns together. It is evident that higher frequency does not indicate higher association. Viewing the figure horizontally, we see that BLEU, ROUGE, and BERTScore have high co-occurrence with different criteria, but very low associations, meaning that they are general metrics that researchers commonly use without a link to specific criteria. Notably, a few metrics have both strong LR and co-occurrence with criteria, for example, in LaaJ: BARTScore - *Naturalness*, MoverScore - *Fluency/Coherence*, but these pairs are not observed similarly in human evaluation papers, confirming that LaaJ and human evaluations are not aligned.

### 4.3 LaaJ and Human Evaluation Comparison

To check whether papers evaluating the same dataset compare the evaluation results of LaaJ and humans, we investigate the 433 papers (12.3% of all papers) where both LaaJ and human evaluations are conducted. As our initial prompt does not extract metadata related to comparisons between LaaJ and human, we design a new prompt (LLM: DeepSeek V3.1 Terminus, exact prompt in

Appendix D.3) including: a binary yes/no question (whether there is a comparison), methods for comparison (e.g., correlation, agreement, ranking), metrics used (Pearson Correlation, Spearman Correlation, Accuracy, etc.) and comparison results (correlation scores, statistical significance, etc.). We require each result to be specific to every criterion and evaluated by both LaaJ and human.

Out of the 433 papers, only 254 (58.7%) explicitly compare their LaaJ results against humans, which is less than 8% of the NLG papers from the top-30 tasks (3,334). Figure 7 shows the comparison results, with most frequent metrics and criteria. Correlation metrics dominate, with Spearman  $\rho$  and Pearson  $r$  being most frequently reported. In terms of criteria, overall quality shows strongest validation ( $\mu = 0.47\text{--}0.71$  for correlation metrics), likely due to aggregation reducing noise. Specific criteria show more variable performance: *Coherence* demonstrates moderate validation ( $\mu = 0.35\text{--}0.54$ ), while *Fluency* exhibits lower scores ( $\mu = 0.35$ ), suggesting LLMs struggle to align with human judgments on this nuanced aspect. Sample sizes vary widely ( $n = 1$  to  $n = 36$ ), revealing uneven research coverage. While Overall quality has been extensively studied, most criteria (less frequent are not shown) including *Factuality* and *Adequacy* have limited validation evidence ( $n = 3\text{--}8$ ). In general, among the few papers that compared LaaJ with humans, the wide range of correlation scores ( $\mu = 0.35\text{--}0.71$ ) suggest LLM evaluations may capture general trends in human judgments to some degree, particularly for over-

<sup>4</sup>This is less accurate than task-specific LR as our data do not have exact map from metric to criterion

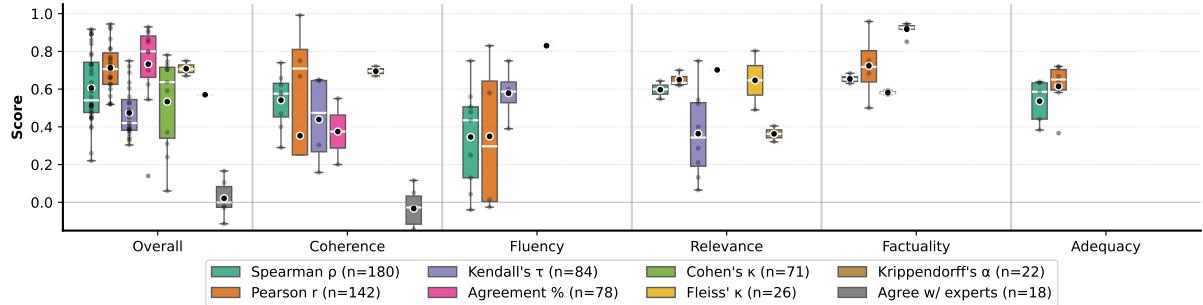


Figure 7: Distribution of LaaJ vs. human evaluation scores by metric and criterion. Individual observations are shown as gray dots.  $n$  denotes the number of studies included.

all quality, but also indicate substantial disagreements. The poor performance against expert human judgment especially raises questions about relying solely on LaaJ for high-stakes decisions.

## 5 Challenges and Recommendations

**Modernize Evaluation in Mature Fields** Our temporal analysis (Figure 4) and task-specific trends (Figure 5) reveal a metric inertia. MT remains dominated by BLEU (Reiter, 2018; Mathur et al., 2020) despite the availability of superior semantic metrics (Lavie et al., 2025). Tasks such as DG and Story Generation require creativity; metrics that rely on n-gram overlap correlate poorly with them. However, they are often used because these metrics are dominant for evaluation in the major tasks. Similarly, QA tasks heavily rely on Exact Match, ignoring semantic variability.

**Recommendation:** Mature fields must aggressively update their leaderboards to prevent overfitting to outdated metrics and couple them with standardized LaaJ evaluation protocols alongside traditional metrics.

**Formalize Metric-Criteria Mapping** Our association heatmap (Figure 6) shows that general metrics (BLEU, ROUGE) are indiscriminately applied across distinct semantic criteria, creating conceptual entanglement. This indicates that researchers select metrics based on popularity or tradition, confirming the lack of standardization noted by Belz et al. (2025).

**Recommendation:** Researchers should explicitly state the target criterion for each metric (e.g., “We use QAFactEval to measure *Factuality*.”), ensure construct validity, and adopt criterion-specific metrics over general-purpose proxies.

## Decouple Human-likeness from Judge-likeness

A critical finding of our study is that LaaJ and human evaluators are not interchangeable signals, but

distinct instruments that measure different dimensions of quality. As shown in Figure 1, metrics that associate strongly with LaaJ (e.g., *Win Rate*, *ASR*, *G-Eval*) form a distinct cluster from those associate with human evaluation (e.g., *Distinct*, *Embedding Average*).

**Recommendation:** Instead of framing LaaJ as a cheaper proxy for human preference, NLG research should adopt a stratified protocol: use LaaJ for technical adherence (instruction following, formatting constraints) and reserve human evaluation for high-level dimensions like *creativity* and *safety*, where LaaJ diverges from human perception.

## 6 Conclusion

In this work, we conducted a large-scale quantitative analysis based on 14,171 papers from 2020 to 2025, using a multi-LLM extraction pipeline validated against human annotations. Our analysis highlights a critical disconnect: the field has pivoted to open-ended generation tasks (e.g.: DG, QA), yet it relies on an evaluation ecosystem marked by “metric inertia” and unverified automation. We show that the exponential rise of LaaJ has not been matched by necessary rigor. The assumption that LLMs serve as effective proxies for human judgment is challenged by our data, which indicates that they prioritize different quality signals and lack strong correlation with human evaluators. Furthermore, the persistence of traditional n-gram metrics in tasks requiring semantic understanding suggests a reluctance to adopt newer, more appropriate tools. Moving forward, we recommend a paradigm shift away from “one-size-fits-all” metrics. Future research should prioritize a stratified evaluation approach, explicitly validating LaaJ against human judgment and reserving human resources for the high-level semantic dimensions that current automated methods fail to capture.

## Limitations

Our agreements of LLM annotations were only computed for the four binary questions. Agreement on the list of terms extracted were not straightforward to compare, as there is not always a standard for different terms, for example, tasks, metrics and criteria. Future work could investigate how to better validate structure information extraction with LLMs. This directly leads to our second limitation: term normalization. Our method of normalizing the terms is limited. For example, our normalization identified 10+ variations of “Overall” criterion (“Overall”, “Overall Quality”, “Overall Aspects”, “Quality” etc.) that we normalized to one name, but we cannot verify if they actually measure the same thing. Third, tasks such as Dialogue Generation are still quite general; under them, there can be many subtasks in which very different metrics and criteria are used. Our association cannot represent these more specific tasks.

## Ethical Statement

The human annotators are NLP researchers that are either co-authors of this paper, or students that are hired by their host institutions.

## References

- Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy Dj Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025. [LitLLMs, LLMs for literature review: Are we there yet?](#) *Transactions on Machine Learning Research*.
- Jesujoba Oluwadara Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. [Charting the landscape of African NLP: Mapping progress and shaping the road ahead](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27795–27829, Suzhou, China. Association for Computational Linguistics.
- Zaid Alyafeai, Maged S. Al-Shaibani, and Bernard Ghanem. 2025. [Mole: Metadata extraction and validation in scientific papers using llms](#). *arXiv*, abs/2505.19800.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Julianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. [LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- Anya Belz, Simon Mille, and Craig Thomson. 2025. [Standard quality criteria derived from current NLP evaluations for guiding evaluation design and grounding comparability and AI compliance assessments](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26685–26715, Vienna, Austria. Association for Computational Linguistics.
- Eleonora Bernasconi, Domenico Redavid, and Stefano Ferilli. 2025. [Integrated survey classification and trend analysis via llms: An ensemble approach for robust literature synthesis](#). *Electronics*, 14(17).
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. [Evaluation of text generation: A survey](#). *arXiv*, abs/2006.14799.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhi-hong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Sri-nath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. [LLMs assist NLP researchers: Critique paper \(meta-\)reviewing](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.
- Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. [LLM-based NLG evaluation: Current status and challenges](#). *Computational Linguistics*, 51:661–687.
- Sebastian Gehrmann, Elizabeth Clark, and Thibault Sel-lam. 2023. [Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text](#). *J. Artif. Int. Res.*, 77.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on ILM-as-a-judge](#). *Preprint*, arXiv:2411.15594.
- David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. [Twenty years of confusion in human evaluation: NLG](#)

- needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.
- Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. 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.
- Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895, Vienna, Austria. Association for Computational Linguistics.
- Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archhana Sindhuwan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.
- Noah Lee, Jiwoo Hong, and James Thorne. 2025. Evaluating the consistency of LLM evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10650–10659, Abu Dhabi, UAE. Association for Computational Linguistics.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.
- Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2025. LLMs as research tools: A large scale survey of researchers' usage and perceptions. In *Second Conference on Language Modeling*.
- Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.
- OpenAI, ;, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sébastien Bubeck, and 108 others. 2025. *gpt-oss-120b & gpt-oss-20b model card*. *Preprint*, arXiv:2508.10925.
- Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.
- Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).
- Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A Lenert. 2025. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, 32(6):1071–1086.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Yicheng Wang, Jiayi Yuan, Yu-Neng Chuang, Zhuoer Wang, Yingchi Liu, Mark Cusick, Param Kulkarni, Zhengping Ji, Yasser Ibrahim, and Xia Hu. 2025. DHP benchmark: Are LLMs good NLG evaluators? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8079–8094, Albuquerque, New Mexico. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

## A License of Artifacts

Our data are based on open-source research papers from ACL Anthology, which are licensed to the general public under a liberal usage policy that allows unlimited reproduction, distribution and hosting of materials on any other website or medium, for non-commercial purposes. LLMs we used are all open-source models, which all permit the usage for research purposes.

## B Experimental Details

### B.1 Model Setup

Due to large model sizes, we used DeepSeek official API for DeepSeek-R1 and Novita API<sup>5</sup> for the other LLMs inferences. We set temperature as 1 for all models. Total cost is around 50 dollars.

### B.2 Term Normalization

**Task** To normalize the task names, we perform three steps: (1) removing punctuation and separators like /, & and -; (2) replacing acronyms with their standard forms (e.g., QA → Question Answering); and (3) applying fuzzy matching to merge near-duplicates (e.g., Dialog Generation → Dialogue Generation). Through this normalization process, the number of unique tasks was reduced from 3,626 to 3,203.

**Evaluation Metric** In preprocessing, we apply  $k$ -value normalization to treat metrics with different  $k$ -values as variants of the same base metric, e.g., BLEU-1,2,4 → BLEU; ROUGE-1,2,L → ROUGE.

**Evaluation Criteria** We normalize LaaJ and human evaluation criteria in the same way: (1) similar to task normalization, we remove accents, punctuations, and separators; (2) we apply manual mapping to convert common quality terms to their nominal forms, e.g., helpful → helpfulness, fair → fairness, grammar → grammaticality; (3) for multi-word criteria, we extract the key term from a list of pre-defined quality nouns, e.g., Language Fluency → Fluency; and (4) we employ fuzzy matching to consolidate minor variations of identical criterion, e.g., Emotion → Emotions.

**Language** We map the languages according to the list of ISO language codes<sup>6</sup> and keep the standard ISO language names; however, many mul-

tilingual datasets do not explicitly enumerate the investigated languages but instead provide only the designation “multilingual” with a language count.

**Model Names** To normalize the models, we preprocess the model names by using a uniform format: family\_version\_size\_extra(name) for all extracted models if applicable.

**Dataset Names** We perform similar preprocessing as other terms: lowercase all letters, replace separators (-, /, \_) with spaces and remove special characters (&). For WMT datasets, as there are many small variations, we grouped them based on years of release.

### B.3 Example of Term Normalization Mapping

We list our term normalization results on top-10 most frequent terms.

Term	Cnt	Top Variants (count)
Question Answering	986	Question Answering (978); question answering (5); Question-Answering (2)
Code Generation	708	Code Generation (707); code generation (1)
Mathematical Reasoning	219	Mathematical Reasoning (213); mathematical reasoning (5); Mathematical reasoning (1)
Instruction Following	211	Instruction Following (196); Instruction-Following (5); instruction following (3)
Language Modeling	163	Language Modeling (162); language modeling (1)
Data To Text Generation	136	Data-to-Text Generation (92); Data-to-Text (36); Data to Text (3)
Story Generation	112	Story Generation (111); Story generation (1)
Explanation Generation	101	Explanation Generation (100); explanation generation (1)
Semantic Parsing	88	Semantic Parsing (86); semantic parsing (2)
Commonsense Reasoning	81	Commonsense Reasoning (78); commonsense reasoning (2); Commonsense reasoning (1)

Table 2: Top-10 normalized tasks and their top variants

### B.4 NLG Task Filtering

The following tasks are considered as non-NLG and filtered out:

- Code Generation
- Mathematical Reasoning

<sup>5</sup><https://novita.ai/>

<sup>6</sup>[https://en.wikipedia.org/wiki/List\\_of\\_ISO\\_639\\_language\\_codes](https://en.wikipedia.org/wiki/List_of_ISO_639_language_codes)

Normalized Term	Cnt	Top Variants (count)
BLEU	2384	BLEU (2380); BLEU-C (1); BLEU-L (1); BLEU-P (1); BLEU@5 (1)
Accuracy	2248	Accuracy (1938); accuracy (265); Accuracy@1 (8); Accuracy@k (4); Weighted Accuracy (4)
ROUGE	1838	ROUGE (1774); ROUGE-L (53); ROUGE-2 (3); ROUGE-S (3); ROUGE-1 (2)
F1	1548	F1 (1136); F1-score (116); F1 Score (52); Macro-F1 (45); F1 score (28)
BERTScore	886	BERTScore (880); BERTSCORE (3); BertScore (3)
Recall	796	Recall (514); recall (54); Recall@k (44); Recall@10 (32); Recall@K (30)
Perplexity	745	Perplexity (664); perplexity (79); Δ Perplexity (1); $\epsilon$ -perplexity (1)
Exact Match	701	Exact Match (555); EM (94); exact match (26); Exact match (14); Exact-match (3)
Precision	590	Precision (447); precision (46); Precision@1 (23); Precision@k (17); Precision@5 (8)
CIDEr	247	CIDEr (242); CIDEr (3); CIDEr-D (1); CIDEr-R (1)

Table 3: Top-10 normalized automatic metrics and their top variants

Term	Cnt	Top Variants (count)
Correctness	451	correctness (173); Correctness (83); Answer correctness (4)
Relevance	390	relevance (187); Relevance (100); topic relevance (6)
Accuracy	250	accuracy (111); Accuracy (40); semantic accuracy (3)
Quality	237	overall quality (33); quality (23); Quality (20)
Helpfulness	217	helpfulness (137); Helpfulness (77); Helpful (2)
Factuality	193	Factuality (36); factuality (32); factual consistency (25)
Coherence	188	coherence (93); Coherence (67); semantic coherence (5)
Consistency	181	consistency (56); Consistency (37); semantic consistency (5)
Fluency	159	fluency (88); Fluency (62); Language Fluency (4)
Completeness	103	completeness (38); Completeness (36); content completeness (2)

Table 4: Top-10 normalized LaaJ criteria and their top variants

Term	Cnt	Top Variants (count)
Chinese	1142	Chinese (1106); Mandarin (18); Simplified Chinese (6)
German	1034	German (1031); Swiss German (1); Swiss-German (1)
Spanish	537	Spanish (536); spa (1)
Russian	402	Russian (401); rus (1)
Arabic	322	Arabic (288); Egyptian Arabic (13); Modern Standard Arabic (7)
Portuguese	219	Portuguese (210); Brazilian Portuguese (9)
Turkish	219	Turkish (218); tur (1)
Korean	210	Korean (209); kor (1)
Finnish	162	Finnish (161); fin (1)
Indonesian	145	Indonesian (144); ind (1)

Table 6: Top-10 normalized languages and their top variants

Term	Cnt	Top Variants (count)
Fluency	777	fluency (477); Fluency (269); Language Fluency (6)
Correctness	677	correctness (211); Correctness (87); answer correctness (13)
Relevance	643	relevance (280); Relevance (155); cultural relevance (7)
Quality	493	overall quality (108); quality (46); Overall Quality (36)
Coherence	486	coherence (248); Coherence (163); semantic coherence (8)
Accuracy	442	accuracy (139); Accuracy (71); Style Accuracy (4)
Consistency	324	consistency (90); Consistency (71); semantic consistency (10)
Informativeness	299	informativeness (183); Informativeness (97); Informative (5)
Factuality	262	factuality (58); Factuality (42); factual consistency (36)
Grammaticality	238	grammaticality (67); grammar (32); Grammaticality (30)

Table 5: Top-10 normalized human evaluation criteria and their top variants

Term	Cnt	Top Variants (count)
GSM8K	454	GSM8K (428); GSM8k (18); GSM-8K (6)
CNN/DailyMail	303	CNN/DailyMail (159); CNN/Daily Mail (70); CNN/DM (51)
WMT14	289	WMT14 English-German (40); WMT14 (35); WMT14 En-De (26)
XSUM	230	XSum (182); XSUM (46); Xsum (1)
TriviaQA	223	TriviaQA (221); TRIVIAQA (1); triviaQA (1)
HotpotQA	214	HotpotQA (192); HotPotQA (16); HOTPOTQA (6)
HumanEval	201	HumanEval (180); HumanEval+ (19); HUMAN-EVAL (2)
NATURAL QUESTIONS	173	Natural Questions (172); NATURAL QUESTIONS (1)
MATH	164	MATH (161); Math (3)
SQuAD	161	SQuAD (155); SQuAD (3); SQuAD-1 (2)

Table 7: Top-10 normalized datasets and their top variants

Term	Cnt	Top Variants (count)
GPT-4	918	GPT-4 (916); ChatGPT (GPT-4) (1); SyncTOD (GPT-4) (1)
GPT-4o	905	GPT-4o (892); GPT-4O (12); gpt-4o (1)
Transformer	601	Transformer (594); TRANSFORMER (2); transformer (2)
BART	552	BART (550); GEE (BART) (1); SOV-MAS (BART) (1)
GPT-3.5-turbo	519	GPT-3.5-turbo (261); GPT-3.5-Turbo (176); GPT-3.5 Turbo (54)
LLaMA-2-7B	507	LLaMA-2-7B (126); Llama-2-7B (123); LLaMA2-7B (77)
GPT-2	506	GPT-2 (493); GPT2 (10); CALM (GPT-2) (2)
BERT	469	BERT (467); PACSUM (bert) (1); Transformer (BERT) + MLP (1)
GPT-3.5	427	GPT-3.5 (422); ChatGPT (GPT-3.5) (4); ChatGPT-API (GPT-3.5) (1)
T5	427	T5 (426); SOV-MAS (T5) (1)

Table 8: Top-10 normalized NLG models and their top variants

- Instruction Following
- Language Modeling
- Semantic Parsing
- Natural Language Inference
- Text Classification
- Text-To-SQL Generation
- Math Problem Solving
- Math Reasoning
- Multiple Choice Question Answering
- Automatic Speech Recognition
- Named Entity Recognition
- Sentiment Analysis
- Text-To-Speech Generation

## B.5 LR Formulations

We list each of our LR below:

Metric (or criteria)-task LR:  $A$  represents a set of papers with a specific task, and  $B$  represents papers with a specific metric (or criteria) (note that  $B$  is not just a subset of  $A$ )

$$LR_{metric|task}(A \rightarrow B) = \frac{P(B|A)}{P(B|\neg A)} \quad (3)$$

Metric-LaaJ (or human) LR:  $A$  represents papers with LaaJ (or human evaluation), and  $B$  represents a metric:

$$LR_{metric|C_{LaaJ}}(A \rightarrow B) = \frac{P(B|A)}{P(B|\neg A)} \quad (4)$$

Metric-criteria LR:  $A$  represents papers with a specific metric, and  $B$  represents papers with a specific

criteria

$$LR_{criteria|metric}(A \rightarrow B) = \frac{P(B|A)}{P(B|\neg A)} \quad (5)$$

## C Additional Results

Key	Description
<b>Q1: NLG task presence (answer_1)</b>	
answer	“Yes” if the paper addresses any NLG task.
quote	Verbatim excerpt supporting the decision.
tasks	NLG tasks (e.g., Summarization, MT, Paraphrase Gen, or Other:<task>).
datasets	List of datasets used for generation/evaluation.
languages	List of languages (e.g., English, Chinese).
models	List of models used to generate outputs.
outputs	Short description of the generated output type.
<b>Q2: Automatic evaluation metrics (answer_2)</b>	
answer	“Yes” if automatic metrics are used.
quote	Excerpt mentioning automatic evaluation.
metrics	List of metrics (e.g., BLEU, BERTScore).
<b>Q3: LaaJ (answer_3)</b>	
answer	“Yes” if an LLM is used <i>after</i> generation.
quote	Excerpt describing LLM-based evaluation.
models	Names of LLMs used as judges.
methods	Procedure (e.g., pairwise, scoring prompt).
criteria	Rubric properties (e.g., fluency, coherence).
<b>Q4: Human evaluation (answer_4)</b>	
answer	“Yes” if humans evaluate generated outputs.
quote	Excerpt mentioning human evaluation.
guideline	Instructions given to human raters.
criteria	Explicit criteria (e.g., fluency, coherence).

Table 9: Schema of the JSON object returned by the LLM for each paper.

Model	A1	A2	A3	A4
DeepSeek-R1	91.97	93.69	95.28	94.32
GPT-OSS-120B	80.68	89.86	95.52	94.93
Qwen3-235B	92.42	94.58	95.01	95.14
Krippendorff’s $\alpha$	0.7101	0.6879	0.8048	0.8124

Table 10: Pairwise agreement (%) between each LLM and LLM-harmonized results (row 1-3), and Krippendorff’s  $\alpha$  among the three LLMs.

Model	A1	A2	A3	A4
DeepSeek-R1	63.5	69.0	16.0	27.1
GPT-OSS-120B	48.5	59.9	15.4	26.7
Qwen3-235B	65.1	70.3	18.7	34.4
Majority Voting	60.8	65.9	15.9	28.3
LLM-harmonized	56.6	56.5	14.0	26.5

Table 11: Yes percentage across all models and questions, for the four binary questions from three LLMs, along with their majority votes and LLM-harmonized results.

Conference	Total	NLG	NLG %	Tasks	Datasets	Languages	NLG Models	Auto Metrics	LaaJ Crit.	Human Crit.	LaaJ Models
ACL-2020	778	297	38.2	115	561	92	1029	314	0	132	0
ACL-2021	571	243	42.6	124	582	76	751	277	0	150	1
ACL-2022	603	275	45.6	196	686	88	842	365	7	172	5
ACL-2023	910	463	50.9	319	1202	162	1211	582	28	239	22
ACL-2024	864	591	68.4	506	1546	190	1392	841	223	334	97
ACL-2025	1600	1096	68.5	832	2596	233	2221	1667	422	484	309
EMNLP-2020	751	291	38.7	159	578	86	910	323	1	140	1
EMNLP-2021	847	326	38.5	151	674	131	1004	410	3	143	4
EMNLP-2022	828	381	46.0	247	950	193	1062	436	14	212	23
EMNLP-2023	1047	553	52.8	395	1401	231	1297	727	96	261	37
EMNLP-2024	1268	823	64.9	634	1985	339	1735	1122	311	374	160
EMNLP-2025	1809	1374	76.0	994	3073	225	2438	2027	478	519	341
INLG-2020	46	44	95.7	31	96	11	137	78	3	47	1
INLG-2021	45	44	97.8	29	89	28	116	72	0	61	0
INLG-2022	25	23	92.0	24	57	12	89	44	0	42	0
INLG-2023	36	36	100.0	27	94	23	107	77	8	44	2
INLG-2024	53	51	96.2	43	100	28	212	103	27	59	7
INLG-2025	50	43	86.0	36	121	15	146	101	48	58	41
NAACL-2021	477	177	37.1	105	394	55	554	279	1	110	2
NAACL-2022	442	184	41.6	125	462	62	569	262	0	130	0
NAACL-2024	486	288	59.3	253	850	234	740	445	97	193	42
NAACL-2025	635	416	65.5	331	1131	153	1023	770	228	252	114
<b>Total</b>	<b>14171</b>	<b>8019</b>	<b>56.6</b>	<b>3308</b>	<b>11289</b>	<b>801</b>	<b>12995</b>	<b>6805</b>	<b>939</b>	<b>1412</b>	<b>783</b>

Table 12: Statistics of total number of papers, NLG papers, and unique number of term counts of tasks, datasets, languages, NLG models, automatic metrics, LaaJ criteria, human criteria, and LaaJ models. NLG papers are counted after LLM harmonization. The number of total papers in ACL and EMNLP increased significantly in 2025.

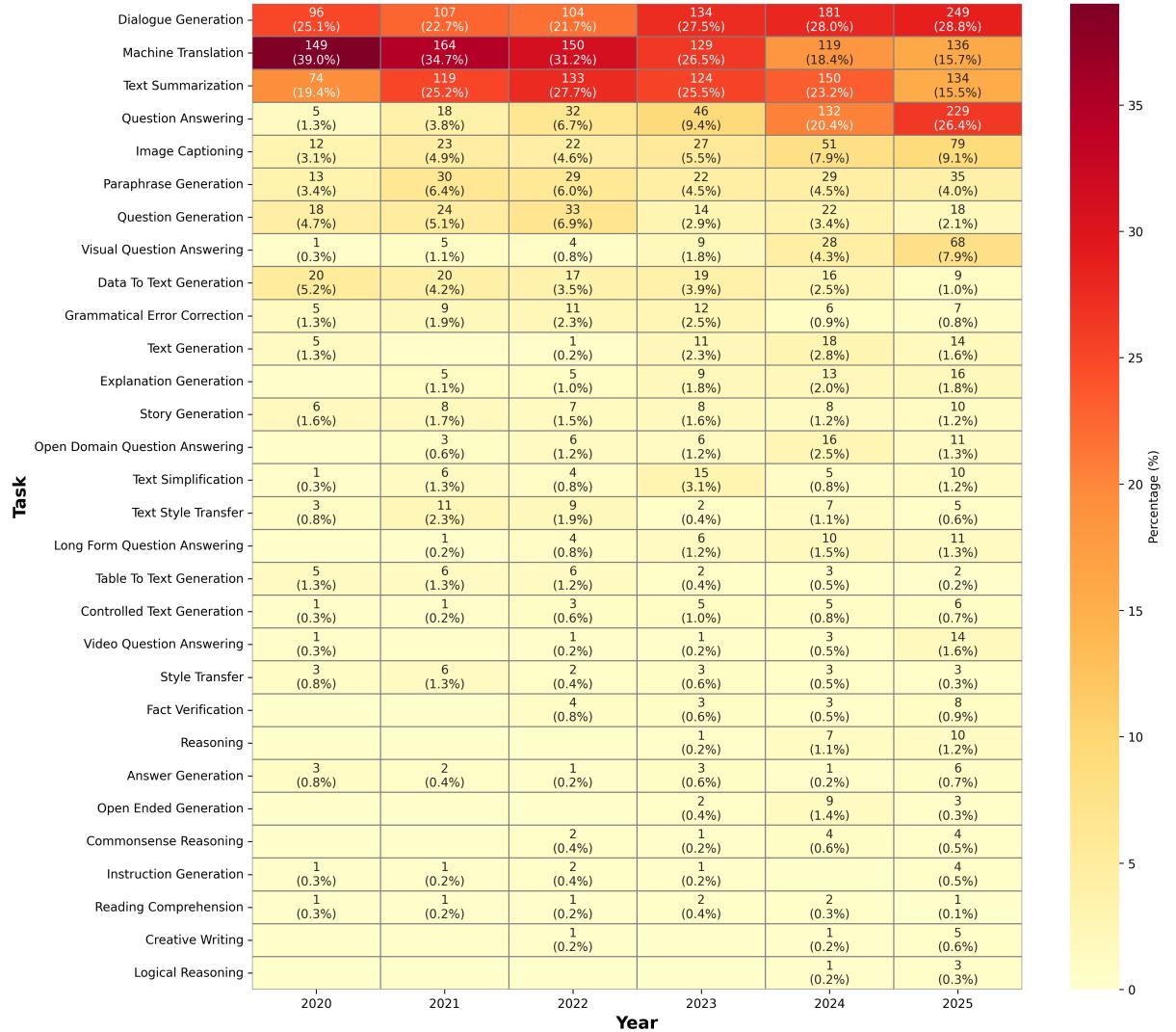


Figure 8: Heatmap of task-year distribution of the top-30 tasks. Both the total number and percentage per year are shown. The percentage of MT is decreasing while QA-related tasks increase significantly (QA, VQA).

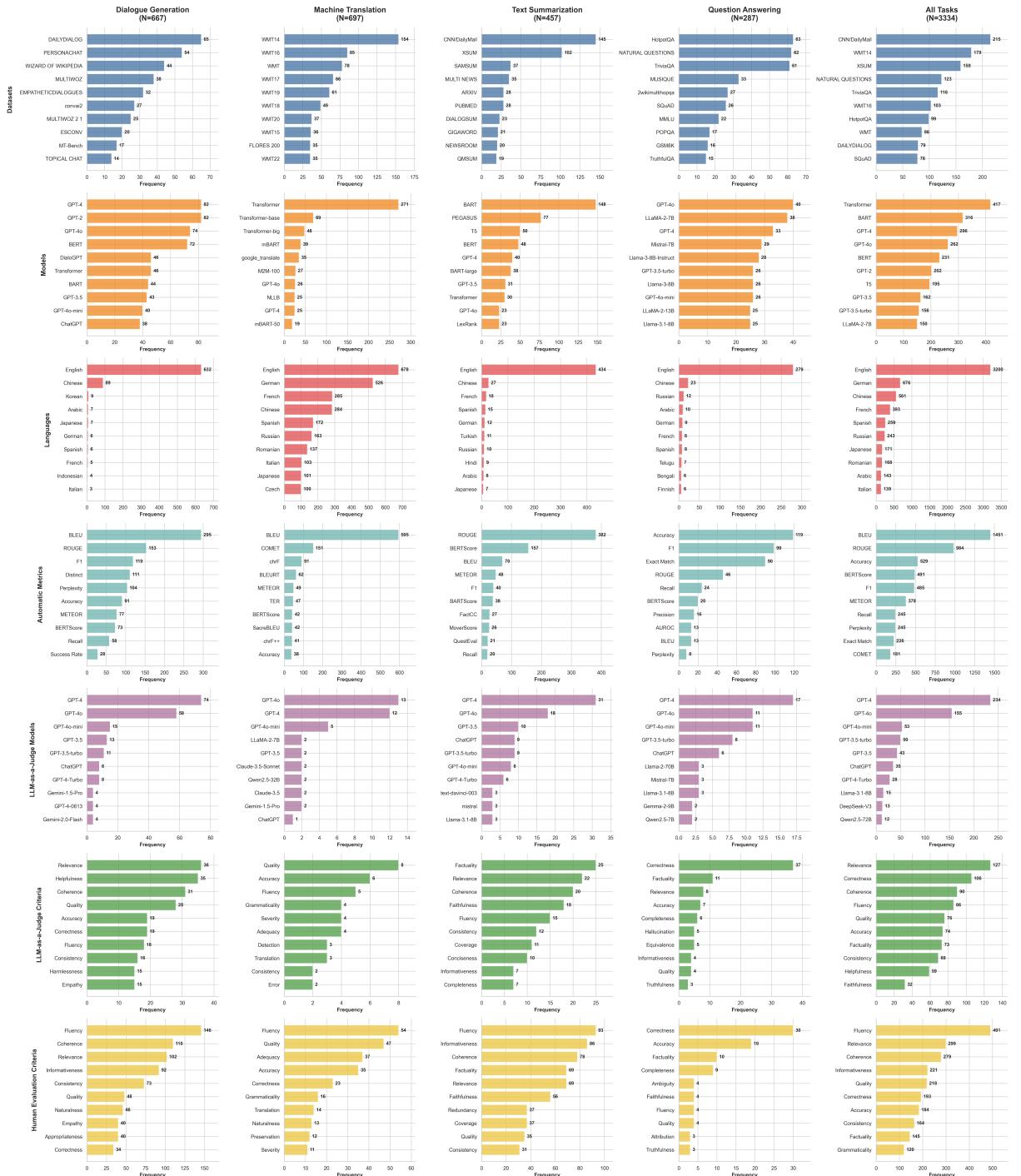


Figure 9: Distribution of metadata by tasks, columns include top-four and all-tasks, and rows are the metadata extracted (counts are after term normalization).

## D Prompts for annotation

### D.1 Prompt 1: Initial Extraction

You are an expert NLP researcher with deep experience in Natural-Language Generation (NLG).

TASK

Read the paper provided below and answer the four numbered questions. Return **only** a single, valid JSON object (no markdown, no comments, no trailing commas).

PAPER

{full\_paper\_text}

QUESTIONS

1. Does the paper address NLG tasks?
2. Does the paper use automatic metrics to evaluate the generated outputs?
3. Does the paper use Large-Language Models (LLMs) as judges  
(i.e., *after* generation, an LLM is used to judge/assess the outputs)?
4. Does the paper conduct *human* evaluations of the generated outputs?

ANSWER FORMAT (strict)

```
{  
    "answer_1": {  
        "answer": "Yes|No",  
        "quote": "...",  
        "tasks": ["Text Summarization", "Machine Translation", "Other:<task>"],  
        "datasets": [...],  
        "languages": ["English", "Chinese", "German", "..."],  
        "models": [...],  
        "outputs": "..."  
    },  
    "answer_2": {  
        "answer": "Yes|No",  
        "quote": "...",  
        "automatic_metrics": [...]  
    },  
    "answer_3": {  
        "answer": "Yes|No",  
        "quote": "...",  
        "models": [...],  
        "methods": ["pairwise evaluation", "..."],  
        "criteria": ["fluency", "coherence", "..."]  
    },  
    "answer_4": {  
        "answer": "Yes|No",  
        "quote": "...",  
        "guideline": "...",  
        "criteria": ["fluency", "coherence", "..."]  
    }  
}
```

INSTRUCTIONS & CONSTRAINTS

\* If the answer is "No", set all other fields in that section to an empty string ("") or an empty list  $\hookrightarrow ([])$ .

\* For **answer\_1.tasks** choose one or more from: {"Text Summarization", "Dialogue Generation", "Paraphrase Generation", "Machine Translation", "Image Captioning", "Code Generation"}. If  $\hookrightarrow$  none apply, use "Other:<task name>".

\* **Answer-2 guidance (automatic evaluation metrics)**\*

\* The **automatic\_metrics** must be a list of automatic metrics used to evaluate the generated outputs.

\* **Answer-3 guidance (LLM as judge)**\*

1. Answer **Yes** only if an LLM is used *after* generation to assess the outputs.

```

2. **methods** - short name/description of the evaluation procedure or prompt.
3. **criteria** - list the rubric properties the LLM is asked to score (e.g.,
↳ "fluency", "relevance", "helpfulness"). If the prompt does not specify criteria, leave as an empty list
↳ [].

* **Answer-4 guidance (human evaluation)**
  * The **quote** must mention humans, annotators, raters, a crowdsourcing platform, or a similar
    ↳ human-evaluation indicator.
  * The **guideline** must mention questions or criteria for the evaluation.
  * The **criteria** must be explicitly mentioned in the human evaluation, list all criteria. If the
    ↳ paper does not specify criteria, leave as an empty list [].

* The **quote** fields must be verbatim excerpts from the paper (use ellipses ... to shorten if needed).
* Use double quotes for all JSON strings; do **not** use backticks.
* Do not add any keys, text, or formatting other than the JSON object.

```

## D.2 Prompt 2: Verification and Normalization

You are verifying and improving the extracted metadata from a research paper about natural language  
 ↳ generation (NLG) evaluation. Your task is to:

1. \*\*Verify\*\* the extracted yes/no answers are correct
2. \*\*Normalize\*\* metadata to use canonical forms (e.g., "BLEU" instead of "bleu")
3. \*\*Correct\*\* any incorrect items
4. \*\*Add\*\* any missing important items
5. \*\*Remove\*\* any irrelevant or incorrect items

## Paper Information

\*\*Paper ID:\*\* {paper\_id}  
 \*\*Title:\*\* {title}  
 \*\*Abstract:\*\* {abstract}

\*\*Full Paper Text:\*\*  
 {full\_text}

---

## Extracted Metadata to Review

### Question 1: Does the paper address NLG tasks?

\*\*Extracted Answer:\*\* {answer\_1\_answer}  
 \*\*Extracted Metadata:\*\*  
 - \*\*Tasks:\*\* {answer\_1\_tasks}  
 - \*\*Datasets:\*\* {answer\_1\_datasets}  
 - \*\*Languages:\*\* {answer\_1\_languages}  
 - \*\*Models:\*\* {answer\_1\_models}  
 - \*\*Outputs:\*\* {answer\_1\_outputs}

---

### Question 2: Does the paper use automatic metrics to evaluate the generated outputs?

\*\*Extracted Answer:\*\* {answer\_2\_answer}  
 \*\*Extracted Metadata:\*\*  
 - \*\*Automatic Metrics:\*\* {answer\_2\_metrics}

---

### Question 3: Does the paper use Large-Language Models (LLMs) as judges (i.e., \*after\* generation, an  
 ↳ LLM is used to judge/assess the outputs)?

\*\*Extracted Answer:\*\* {answer\_3\_answer}  
 \*\*Extracted Metadata:\*\*  
 - \*\*Models:\*\* {answer\_3\_models}

```

- **Methods:** {answer_3_methods}
- **Criteria:** {answer_3_criteria}

---

### Question 4: Does the paper conduct *human* evaluations of the generated outputs?

**Extracted Answer:** {answer_4_answer}
**Extracted Metadata:** 
- **Guideline:** {answer_4_guideline}
- **Criteria:** {answer_4_criteria}

---

## Your Task

For each question above:
1. **Verify the Yes/No answer** - Is it correct based on the full paper text?
2. **Review the metadata lists** - For each item:
   - Is it correctly extracted from the paper?
   - Is it relevant to the specific question?
   - Should it be normalized? (e.g., "BLEU" vs "bleu", "GPT-3" vs "gpt-3")
3. **Add missing items** - Are there important items mentioned in the paper that are missing?
4. **Remove incorrect items** - Are there items that shouldn't be there?

## Guidelines

### Normalization Rules
- Use canonical/standard forms (e.g., "BLEU" not "bleu", "GPT-3" not "gpt-3")
- Use consistent capitalization for metrics, models
- Use title case for tasks (e.g., "Machine Translation")
- **For metrics**: Simplify to base form (e.g., "BLEU-1", "BLEU-2", "BLEU-4" → all become "BLEU");
  ↳ "ROUGE-1", "ROUGE-2", "ROUGE-L" → all become "ROUGE")
- **For models**: Keep version numbers distinct (e.g., "GPT-3", "GPT-4", "BERT-base", "BERT-large" are
  ↳ different)
- Merge case variations and abbreviations that refer to the same thing

### Verification Rules
- Only include items **explicitly mentioned** in the paper
- Focus on the **main contributions** - don't include every model/dataset mentioned in passing
- For tasks: Only include NLG tasks that are actually evaluated/studied
- For metrics: Include all automatic metrics used to evaluate NLG outputs
- For criteria: Include evaluation criteria used for human eval or LLM-as-evaluator
- Be accurate over complete - it's better to miss minor details than include wrong information

### Answer-Specific Guidelines

**Question 1 (Does the paper address NLG tasks?):**
- Answer "Yes" only if the paper studies/evaluates natural language GENERATION (not just
  ↳ understanding/classification)
- **Tasks**: Choose from {"Text Summarization", "Dialogue Generation", "Paraphrase Generation", "Machine
  ↳ Translation", "Image Captioning", "Code Generation"}. If none apply, use "Other:<task name>".
- **Datasets**: NLG datasets used
- **Languages**: Languages of the generated outputs (e.g., "English", "Chinese", "German")
- **Models**: NLG models being evaluated
- **Outputs**: Description of what is being generated

**Question 2 (Does the paper use automatic metrics to evaluate the generated outputs?):**
- Answer "Yes" if the paper uses any automatic metrics to evaluate generated text
- **Automatic Metrics**: List of automatic evaluation metrics (e.g., BLEU, ROUGE, METEOR, BERTScore)
- **Important**: Simplify metric variants to base form

**Question 3 (Does the paper use LLMs as judges?):**
- Answer "Yes" only if an LLM is used *after* generation to assess the outputs (not just as generation
  ↳ model)
- **Models**: Which LLMs are used for evaluation (e.g., "GPT-4", "Claude-3")
- **Methods**: Short name/description of the evaluation procedure or prompt (e.g., "pairwise
  ↳ evaluation", "direct scoring")

```

- **Criteria**: List the rubric properties the LLM is asked to score (e.g., "fluency", "relevance", "helpfulness"). If not specified, use empty list.

**Question 4 (Does the paper conduct human evaluations?)**  
 - Answer "Yes" if humans, annotators, raters, or crowdsourcing are used to evaluate generated outputs  
 - **Guideline**: Description of questions or criteria for the evaluation  
 - **Criteria**: List all criteria explicitly mentioned (e.g., "fluency", "coherence", "relevance"). If not specified, use empty list.

**## Output Format**

Please return ONLY a JSON object with the following structure:

```
{
  "paper_id": "{paper_id}",
  "answer_1": {
    "answer": "Yes"/"No",
    "answer_changed": true/false,
    "tasks": ["normalized_task1", ...],
    "datasets": ["normalized_dataset1", ...],
    "languages": ["normalized_language1", ...],
    "models": ["normalized_model1", ...],
    "outputs": ["output_description1", ...],
    "changes_made": {
      "added": {"tasks": [...], "datasets": [...], ...},
      "removed": {"tasks": [...], "datasets": [...], ...},
      "normalized": {"original_item": "normalized_item", ...},
      "explanation": "Brief explanation of major changes"
    }
  },
  "answer_2": {
    "answer": "Yes"/"No",
    "answer_changed": true/false,
    "automatic_metrics": ["normalized_metric1", ...],
    "changes_made": {
      "added": {"automatic_metrics": [...]},
      "removed": {"automatic_metrics": [...]},
      "normalized": {"original_metric": "normalized_metric", ...},
      "explanation": "Brief explanation of major changes"
    }
  },
  "answer_3": {
    "answer": "Yes"/"No",
    "answer_changed": true/false,
    "models": ["normalized_model1", ...],
    "methods": ["normalized_method1", ...],
    "criteria": ["normalized_criterion1", ...],
    "changes_made": {
      "added": {"models": [...], "methods": [...], "criteria": [...]},
      "removed": {"models": [...], "methods": [...], "criteria": [...]},
      "normalized": {"original_item": "normalized_item", ...},
      "explanation": "Brief explanation of major changes"
    }
  },
  "answer_4": {
    "answer": "Yes"/"No",
    "answer_changed": true/false,
    "guideline": ["guideline1", ...],
    "criteria": ["normalized_criterion1", ...],
    "changes_made": {
      "added": {"guideline": [...], "criteria": [...]},
      "removed": {"guideline": [...], "criteria": [...]},
      "normalized": {"original_item": "normalized_item", ...},
      "explanation": "Brief explanation of major changes"
    }
  },
  "overall_notes": "Any general observations"
}
```

```

## Important Notes

- **Be conservative with changes** - Only modify if you're confident
- **Prioritize accuracy** - Better to keep existing correct items than to add uncertain ones
- **Normalize consistently** - Use standard naming conventions
- **Document major changes** - Explain why you added/removed important items
- **Use the full paper text** - Read the complete paper to verify all metadata is accurate and complete

```

### D.3 Prompt 3: LLM-Human Validation Extraction

You are analyzing a research paper that uses **both** LLM-as-a-judge (LLM evaluators) and human evaluation to assess natural language generation outputs. Your task is to extract detailed information about how (or whether) the paper validates LLM evaluation against human evaluation.

**## Paper Information**

```

Paper ID: {paper_id}
Title: {title}
Abstract: {abstract}
Full Paper Text: {full_text}

```

---

**## Previously Extracted Metadata**

```

LLM-as-a-Judge (Answer 3)
- Models: {answer_3_models}
- Methods: {answer_3_methods}
- Criteria: {answer_3_criteria}

```

```

Human Evaluation (Answer 4)
- Guideline: {answer_4_guideline}
- Criteria: {answer_4_criteria}

```

---

**## Your Task**

Extract information about **validation** of LLM-as-a-judge against human evaluation. Answer the following questions based on the full paper text:

**### Question 1: Is there explicit validation?**

**Does the paper explicitly compare LLM-as-a-judge results with human evaluation results?**

Answer "Yes" only if the paper:

- Compares LLM and human judgments on the same set of instances
- Reports quantitative metrics of agreement/correlation between LLM and human
- Discusses the relationship between LLM and human evaluation results

Answer "No" if:

- Both LLM and human evaluations are conducted but never compared
- LLM and human evaluate different sets of instances or different aspects
- Only qualitative discussion without any comparison

---

**### Question 2: LLM-as-a-Judge Details**

**A. Number of LLM Models:**

- How many different LLM models were used as judges?
- List the models (from Answer 3 metadata)

**B. LLM Prompts\*\* (CRITICAL - Extract exact prompts if available):**

- Does the paper show the exact prompt(s) used for LLM evaluation?
- If yes, extract the full prompt text verbatim

- If no, describe what information is provided about the prompts
- Note: Look for prompts in main text, tables, figures or appendices.

---

### ### Question 3: Human Evaluation Details

#### **\*\*A. Number of Human Evaluators\*\*:**

- How many human annotators/evaluators were used?

#### **\*\*B. Evaluator Type\*\*:**

- "expert": Domain experts, researchers, or trained annotators
- "crowdsourced": Crowd workers (MTurk, Prolific, etc.)
- "mixed": Combination of both
- "unclear": Not specified

#### **\*\*C. Inter-Annotator Agreement\*\* (CRITICAL):**

- Was inter-annotator agreement (IAA) reported?
- If yes, extract:
  - Metric used (Cohen's kappa, Fleiss' kappa, Krippendorff's alpha, percentage agreement, etc.)
  - Value(s) reported
  - Interpretation if provided (e.g., "substantial agreement")
- If no, note "Not reported"

#### **\*\*D. Human Evaluation Guidelines\*\*:**

- Does the paper provide detailed evaluation guidelines/instructions?
- Are example annotations or scoring rubrics shown?
- Where are guidelines described (main text, appendix, supplementary)?

---

### ### Question 4: Validation Setup (if Q1 = Yes)

#### **\*\*A. Validation Type\*\* (select all that apply):**

- "correlation\_analysis": Correlation between LLM and human scores
- "agreement\_analysis": Agreement between LLM and human labels/judgments
- "ranking\_comparison": Compare rankings produced by LLM vs human
- "error\_analysis": Analyze disagreements between LLM and human
- "other": Other types of validation

#### **\*\*B. Validation Metrics\*\*:**

Examples: "Pearson correlation", "Spearman correlation", "Kendall's tau", "Cohen's kappa", "Accuracy", ↳ "F1", "Krippendorff's alpha", etc.

#### **\*\*C. Shared Evaluation Criteria\*\*:**

List only criteria that both LLM and human evaluate.

#### **\*\*D. Sample Size\*\*:**

- How many instances/examples were used for validation?
- Note if validation uses subset or all evaluated instances

---

### ### Question 5: Validation Results (if Q1 = Yes)

#### **\*\*A. Correlation/Agreement Scores\*\* (Extract ALL reported values):**

For EACH metric reported, extract:

- Metric name (e.g., "Spearman correlation", "Cohen's kappa")
- Value (numerical - extract exact value as reported)
- Which criterion it applies to (e.g., "fluency", "coherence")
- Which LLM model (if multiple LLMs were compared)
- Statistical significance if reported (p-value, confidence intervals)

#### **\*\*B. Correlation Strength Interpretation\*\*:**

- Does the paper interpret correlation strength?
- What threshold do they use for "strong" correlation?
- Do they compare to prior work?

---

```

## Output Format

Return ONLY a JSON object with this structure:
{
  "paper_id": "{paper_id}",
  "explicit_validation": {
    "answer": "Yes"/"No",
    "explanation": "Brief explanation"
  },
  "llm_judge_details": {
    "num_models": 1,
    "models": ["GPT-4", ...],
    "prompts": {
      "provided": "yes"/"no"/"partial",
      "location": "appendix"/"main_text"/"figure"/"not_provided",
      "notes": "Additional notes"
    }
  },
  "human_eval_details": {
    "num_evaluators": 3,
    "evaluator_type": "expert"/"crowdsourced"/"mixed"/"unclear",
    "inter_annotation_agreement": {
      "reported": "yes"/"no",
      "metric": "Fleiss' kappa",
      "value": 0.68,
      "interpretation": "substantial agreement"
    },
    "guidelines": {
      "detailed_guidelines_provided": "yes"/"no"/"partial",
      "location": "appendix"/"main_text"/"not_provided"
    }
  },
  "validation_setup": {
    "validation_types": ["correlation_analysis", ...],
    "validation_metrics": ["Pearson correlation", ...],
    "shared_criteria": ["fluency", ...],
    "sample_size": {
      "total_generated": 100,
      "validated_by_both": 50
    }
  },
  "validation_results": {
    "quantitative_scores": [
      {
        "metric": "Spearman correlation",
        "value": 0.87,
        "criterion": "fluency",
        "llm_model": "GPT-4"
      }
    ],
    "summary_finding": "1-2 sentence summary"
  },
  "criteria_mapping": {
    "llm_only_criteria": [...],
    "human_only_criteria": [...],
    "shared_criteria": [...]
  }
}

## Important Notes

**If explicit_validation.answer = "No":**
- Set validation_setup and validation_results to `null`
- Still fill out llm_judge_details and human_eval_details
- Still fill out criteria_mapping

**Always extract (regardless of validation):**

```

```
- llm_judge_details: Always extract LLM setup information
- human_eval_details: Always extract human evaluation information
- criteria_mapping: Always show which criteria each method uses

**Guidelines:**
- Be precise: Only mark as validated if there's explicit comparison
- Extract exact values: Copy numerical results exactly as reported
- Distinguish correlation types: Pearson vs Spearman vs Kendall
- Note statistical significance: If p-values or confidence intervals are reported
- Consider multi-criterion scenarios: LLM and human might evaluate different criteria even in same
  ↳ paper

Common Scenarios:
1. Full validation: Paper uses LLM to evaluate all outputs, validates on human-annotated subset,
  ↳ reports correlation
2. Parallel evaluation: Both LLM and human evaluate the same outputs, direct comparison
3. Sequential validation: Human labels used as ground truth, LLM accuracy measured
4. Independent streams: Both methods used but never compared (answer "No")
5. Qualitative only: Paper discusses differences but no quantitative comparison (answer "No")

Read the full paper text carefully and extract all validation-related information accurately.
```

## E Human Annotation Guideline

### Overview

This document provides detailed instructions for manually annotating research papers about natural language generation (NLG) evaluation. You will read each paper and extract structured metadata to answer four main questions about the paper's approach to NLG evaluation.

**Important:** The papers you are annotating have been pre-filtered as potential NLG papers (with an initial “Yes” answer to Question 1). However, this filtering may not be perfect. You should verify this classification and change the answer to “No” if, after reading the paper, you determine it does not actually address NLG tasks.

### Annotation Process

#### Step 1: Read the Paper

1. Download and read the paper using the PDF link provided in the spreadsheet
2. Take notes as you read to identify relevant information for each question

#### Step 2: Answer Four Main Questions

For each paper, you will answer four yes/no questions and extract relevant metadata.

---

#### Question 1: Does the paper address NLG tasks?

##### Definition:

Natural Language Generation (NLG) refers to tasks where a system produces/generates natural language text as output. This is distinct from Natural Language Understanding (NLU) tasks where the system only reads/analyzes text.

##### Important Context:

These papers have been pre-filtered as NLG papers (initially classified as “Yes”). However, the automatic filtering may have made mistakes. Your job is to verify this classification by carefully reading the paper.

##### How to Answer:

###### *Answer “Yes” if:*

- The paper generates text, sentences, or natural language as output
- The paper evaluates or studies systems that produce natural language
- Examples: summarization systems, dialogue systems, machine translation, paraphrase generation, image captioning

###### *Answer “No” if:*

- The paper only does classification, tagging, or understanding tasks
- No text is generated as output
- Examples: sentiment analysis, named entity recognition, question answering with extractive answers (just selecting existing text)
- The paper was incorrectly classified during pre-filtering

###### *If you change the answer to “No”:*

- Explain your reasons for “No”, for example, specify the tasks addressed in this paper
- Skip this paper entirely and move to the next paper
- You do not need to fill in any other fields (Q1 metadata, Q2, Q3, Q4)

##### Metadata to Extract (if answer is “Yes”):

## 1. Tasks (List of NLG task types)

*What to include:*

- The main NLG task(s) that the paper addresses
- Use standardized task names from this list:
  - Text Summarization
  - Dialogue Generation
  - Paraphrase Generation
  - Machine Translation
  - Image Captioning
  - Code Generation
- If the task doesn't fit any category, use: "Other: [specific task name]"

*Examples:*

- **[GOOD]** "Text Summarization", "Dialogue Generation"
- **[BAD]** Don't include: "NLP", "Generation" (too vague)

*Instructions:*

- Include only the primary task(s) being studied/evaluated
- Don't include tasks mentioned only in related work or background
- If a paper studies multiple NLG tasks, list all of them

## 2. Datasets (List of dataset names)

*What to include:*

- Names of NLG datasets used for experiments or evaluation
- Include datasets that are central to the paper's contribution
- Use the official dataset name as cited in the paper

*Examples:*

- **[GOOD]** "CNN/DailyMail", "XSum", "WMT14", "MultiWOZ"
- **[BAD]** Don't include: Generic terms like "news articles", "dialogue data"

*Instructions:*

- Only include datasets that are actually used in the paper's experiments
- Don't include datasets only mentioned in related work
- If a paper creates a new dataset, include its name, if it is not named, use "Proposed dataset for <task name>"
- Use the exact name from the paper

## 3. Languages (List of languages)

*What to include:*

- The language(s) of the generated outputs
- Use standard language names in English

*Examples:*

- **[GOOD]** "English", "Chinese", "German", "French"
- **[BAD]** Don't use: ISO codes like "en", "zh" (use full names)

*Instructions:*

- Include all target languages for generation
- For multilingual papers, list all languages mentioned
- If the paper doesn't specify but uses English datasets, annotate as "English"

## 4. Models (List of model names)

*What to include:*

- Names of NLG models used or proposed for GENERATION (not evaluation)
- These are models that produce/generate the text outputs

- Include both models proposed by the authors and baseline generation models

*IMPORTANT: This is for generation models only, NOT evaluation models:*

- [GOOD] Include: Models that generate the summaries, translations, dialogue responses, etc.
- [BAD] Don't include: Models used to evaluate/judge outputs (those go in Q3)
- Example: If GPT-4 generates text → Q1. If GPT-4 judges/evaluates text → Q3.

*Examples:*

- [GOOD] “GPT-3”, “BART”, “T5”, “Seq2Seq”, “Transformer” (if used for generation)
- [GOOD] “GPT-3.5”, “GPT-4” (keep versions distinct when specified)
- [BAD] Don't include: Models only used for evaluation/judging outputs

*Normalization rules:*

- Use canonical model names with proper capitalization
- Keep version numbers distinct: “GPT-3” vs “GPT-4” are different models
- Normalize case variations: “gpt-3” → “GPT-3”, “bert” → “BERT”
- Include size variants if specified: “BERT-base”, “BERT-large”

*Instructions:*

- Focus on models that generate the NLG outputs being evaluated
- Don't list every model mentioned in passing in related work
- If a paper proposes a new generation model with a name, include it
- For papers proposing unnamed approaches, describe briefly: “Proposed model with [backbone] architecture”
- Remember: Evaluation models go in Q3, not here!

## 5. Outputs (List of output descriptions)

*What to include:*

- Brief descriptions of what text is being generated
- Focus on the actual output artifacts, not the process

*Examples:*

- [GOOD] “News article summaries”, “Task-oriented dialogue responses”, “English-to-German translations”, “Image captions”
- [BAD] Don't use long sentences: “The approach generates task-oriented dialogue responses”

*Instructions:*

- Use natural language descriptions
- Be specific but concise (3-7 words typically)
- If multiple types of outputs, list the main ones
- Focus on what is generated, not how

## Question 2: Does the paper use automatic metrics to evaluate the generated outputs?

**Definition:**

Automatic metrics are computational measures that evaluate the quality of generated text without human involvement. These metrics compare generated text against reference texts or use logics, rules or learned models to score outputs.

**How to Answer:**

*Answer “Yes” if:*

- The paper reports scores from any automatic evaluation metrics
- Metrics are used to compare different systems or configurations
- Common examples: BLEU, ROUGE, METEOR, BERTScore, BLEURT, ChrF

*Answer “No” if:*

- No evaluation of generated outputs is performed, or only uses human evaluation

**Metadata to Extract (if answer is “Yes”):**

**Automatic Metrics (List of metric names)**

*What to include:*

- All automatic evaluation metrics used to assess the generated outputs
- Use standardized metric names with proper capitalization, if unsure, check online sources
- A list of common evaluation metric names: <https://huggingface.co/evaluate-metric>

*Examples:*

- [GOOD] “BLEU”, “ROUGE”, “METEOR”, “BERTScore”, “BLEURT”, “ChrF”, “TER”, “PARENT”
- [BAD] Don’t use: “bleu”, “Blue” (wrong capitalization)

*Critical normalization rule:*

Simplify metric variants to their base form:

- “BLEU-1”, “BLEU-2”, “BLEU-4” → all become “BLEU”
- “ROUGE-1”, “ROUGE-2”, “ROUGE-L” → all become “ROUGE”
- “F1-score”, “Exact Match” → keep as separate metrics
- “BERT-F1”, “BERTScore”, “Bertscore” → use “BERTScore”

*Why normalize variants?*

- We want to know which metric families are used, not every variant
- This simplifies analysis and prevents overcounting similar metrics

*Instructions:*

- Include all metrics actually used in the evaluation section
- Don’t include metrics only mentioned in related work
- Use the base metric name (BLEU not BLEU-4)

---

### Question 3: Does the paper use Large Language Models (LLMs) as judges?

**Definition:**

This refers to using LLMs after generation to automatically evaluate or judge the quality of generated outputs. The LLM is used as an evaluator, not as the generation model itself (except both use the same model).

**How to Answer:**

*Answer “Yes” if:*

- An LLM (like GPT-4, Claude, Llama) is used to score, rank, or judge generated outputs
- The paper describes using LLM prompts to assess quality
- Examples: “GPT-4 as a judge”, “LLM-based evaluation”, “using ChatGPT to rate fluency”

*Answer “No” if:*

- LLMs are only used for generation, not evaluation
- No LLM-based evaluation is performed
- Only traditional automatic metrics or human evaluation is used

**Metadata to Extract (if answer is “Yes”):**

**1. Models (List of LLM names used as judges)**

*What to include:*

- Names of specific LLMs used for evaluation

- Include version numbers when specified

*Examples:*

- [GOOD] “GPT-4”, “GPT-3.5”, “Claude-3”, “PaLM-2”, “Llama-2-70B”
- [BAD] Don’t use: “ChatGPT” (use “GPT-3.5-Turbo” or “GPT-4” if version is known)

*Normalization rules:*

- Use official model names with proper capitalization
- Keep versions distinct: “GPT-3” vs “GPT-4”
- If paper says “ChatGPT” without version, keep as “ChatGPT”
- Keep consistent format: “ModelName-Version-Size” (e.g., “Claude-3-Opus”, “Llama-2-70B”)

## 2. Methods (List of evaluation methods/approaches)

*What to include:*

- Brief description or name of the evaluation procedure
- How the LLM is prompted or used

*Examples:*

- [GOOD] “Pairwise comparison”, “Direct scoring”, “Likert scale rating”, “Binary preference”, “Multi-aspect scoring”
- [GOOD] “Chain-of-thought evaluation”, “Self-consistency”
- [BAD] Don’t include: Full prompt text (too detailed)

*Instructions:*

- Use short descriptive names (2-5 words)
- If the paper gives a name to their method, use it
- If not, describe the approach briefly

## 3. Criteria (List of evaluation criteria)

*What to include:*

- The specific aspects or dimensions that the LLM is asked to evaluate
- The rubric properties being scored

*Examples:*

- [GOOD] “Fluency”, “Relevance”, “Coherence”, “Factuality”, “Helpfulness”, “Safety”
- [BAD] Don’t include: The scores themselves (like “1-5 scale”)

*If criteria are not specified:*

- Leave the field empty if the paper doesn’t mention specific criteria
- Don’t guess or infer criteria

*Normalization rule:*

- Use criteria names with proper capitalization: “Fluency” instead of “fluency”
- Use nouns instead of adjectives: “Naturalness” instead of “natural”

*Instructions:*

- List all criteria mentioned
- Use the exact terminology from the paper when possible
- If paper uses very general terms like “quality”, still include it

## Question 4: Does the paper conduct human evaluations of the generated outputs?

**Definition:**

Human evaluation means that real people (not LLMs or automatic metrics) are asked to read and assess the generated outputs. This includes crowdsourcing, expert annotations, or user studies.

## **How to Answer:**

*Answer “Yes” if:*

- Human annotators / raters evaluate the generated text
- User studies with human participants assess outputs after the generation, not for human annotated datasets used to training the generation model

*Answer “No” if:*

- Only automatic metrics or LLM judges are used
- Humans are only used for data collection, not evaluation

## **Metadata to Extract (if answer is “Yes”):**

### **1. Methods (List of evaluation methods/approaches)**

*What to include:*

- Brief description or name of the evaluation procedure
- How human evaluators are asked to assess the outputs
- The type of evaluation task (rating, ranking, comparison, etc.)

*Examples:*

- [GOOD] “Pairwise comparison”, “Direct scoring”, “Likert scale rating”, “Binary preference”, “Multi-aspect rating”
- [GOOD] “Ranking”, “Best-worst scaling”, “Magnitude estimation”
- [GOOD] “A/B testing”, “Adequacy and fluency rating”
- [BAD] Don’t include: Full instruction text or specific questions (too detailed)

*Instructions:*

- Use short descriptive names (2-5 words)
- If the paper gives a name to their evaluation method, use it
- If not, describe the approach briefly (e.g., “5-point Likert scale rating”)
- If multiple different evaluation methods are used, list them separately

### **2. Criteria (List of evaluation criteria)**

*What to include:*

- The specific aspects or dimensions that humans are asked to evaluate
- Evaluation categories or rubric items

*Examples:*

- [GOOD] “Fluency”, “Adequacy”, “Coherence”, “Informativeness”, “Naturalness”, “Relevance”, “Grammaticality”, “Readability”, “Factuality”
- [BAD] Don’t include: The scores themselves

*If criteria are not specified:*

- Leave it empty if the paper only describes a general “quality” rating without specific dimensions
- Don’t infer criteria if not explicitly stated

*Normalization rule:*

- Use criteria names with proper capitalization: “Fluency” instead of “fluency”
- Use nouns instead of adjectives: “Naturalness” instead of “natural”

*Instructions:*

- Use exact terminology from the paper
- List all criteria mentioned in the evaluation setup
- If the paper uses “overall quality” as the only criterion, include it

---

## **General Annotation Guidelines**

## Quality Standards

1. **Be accurate:** Only annotate information that is explicitly stated in the paper
2. **Be complete:** Try to find all relevant information for each question
3. **Be consistent:** Use standardized names and formats
4. **Be conservative:** When in doubt, don't guess—leave it out or mark as uncertain

## Handling Edge Cases

*If you're unsure about something:*

- Add a comment/note in your annotation
- Mark items you're uncertain about
- It's better to be cautious than incorrect

*If information is ambiguous:*

- Use your best judgment based on context
- Add a note explaining your interpretation

*If a field should be empty:*

- Leave it empty, don't put placeholder text

## Normalization Standards

*Capitalization:*

- Metrics: Follow standard conventions (BLEU, ROUGE, BERTScore)
- Models: Use official capitalization (GPT-4, BERT, T5)
- Tasks: Use title case (Text Summarization, Machine Translation)
- Languages: Capitalize (English, Chinese, German)

*Naming:*

- Use official, canonical names when available
- Be consistent across all annotations
- Merge obvious duplicates (e.g., "MT" and "Machine Translation" → use "Machine Translation")

## Common Mistakes to Avoid

### [DON'T INCLUDE]:

- Information from related work sections (unless actually used in the paper)
- Background or motivation content (focus on what the paper does)
- Every model/dataset mentioned (focus on what's evaluated)

### [DO INCLUDE]:

- Information from experiments and evaluation sections
- Main contributions and findings
- All metrics, criteria, and methods actually used
- Clear, specific terminology from the paper

---

## Annotation Workflow Summary

### For each paper:

1. Read the paper (especially abstract, methodology, and evaluation sections)
2. Answer Question 1: Does it address NLG tasks?
  - Verify the pre-filtered classification - the paper was initially classified as "Yes"
  - Change to "No" if it doesn't actually address NLG tasks
  - If No: Skip to the next paper (we only annotate NLG papers)
  - If Yes: Continue to extract all metadata
3. Extract Q1 metadata: tasks, datasets, languages, models, outputs

4. Answer Question 2: Does it use automatic metrics?
  - If Yes: Extract and normalize metric names
5. Answer Question 3: Does it use LLMs as judges?
  - If Yes: Extract LLM models, methods, and criteria
6. Answer Question 4: Does it conduct human evaluation?
  - If Yes: Extract evaluation methods and criteria
7. Review your annotations for completeness and consistency
8. Add any notes about difficult decisions or uncertainties

#### **Questions or Issues?**

If you encounter any problems during annotation:

- Document unclear cases in the notes section
- Flag papers that are ambiguous or difficult to categorize
- Ask for clarification on systematic issues