What Are We Measuring in NLG? A Meta-Analysis of Evaluation Trends 2020-2025

Jing Yang[1,2] **Nils Feldhus**[7,3,*] **Salar Mohtaj**[3]
Leonhard Hennig[3] **Qianli Wang**[1] **Eleni Metheniti**[4]
Sherzod Hakimov[5] **Charlott Jakob**[1,3] **Veronika Solopova**[1]
Konrad Rieck[1,2] **David Schlangen**[3,5] **Sebastian Möller**[1,2,3] **Vera Schmitt**[1,3,6,8]
[1]Technische Universität Berlin
[2]BIFOLD – Berlin Institute for the Foundations of Learning and Data
[3]German Research Center for Artificial Intelligence (DFKI)
[4]Joint Research Center [5]University of Potsdam [6]CERTAIN
[7]University of Groningen [8]Johannes Gutenberg University Mainz

Abstract

As Natural Language Generation (NLG) dominates modern NLP, scalable evaluation remains a critical bottleneck. Consequently, LLM-as-a-judge (LaaJ) adoption has accelerated rapidly, appearing in more papers than human evaluation in 2025. This pivotal shift motivates a critical analysis of current evaluation practices. Overcoming the limits of rigid keyword filtering and manual review, we employ a multi-LLM information extraction pipeline to gather structured metadata from 14,171 papers across four major NLP conferences (2020–2025). Analyzing 3,334 filtered NLG papers, we identify three systemic challenges. (1) Metric inertia: despite the shift toward open-ended generation, legacy lexical metrics (BLEU, ROUGE) persist as primary indicators, typically used alongside rather than replaced by semantic alternatives. (2) Metric-criteria mapping problem: our paper-level co-occurrence data reveals that general-purpose automatic metrics are applied as broad proxies for quality, without specifying which dimension of text generation they are intended to evaluate. (3) Validation gap: LaaJ has grown rapidly without commensurate human validation (fewer than 8% of papers). Crucially, while LaaJ correlates with aggregate quality, alignment collapses on fine-grained criteria like fluency. To address these gaps, we distill our findings into a minimal Evaluation Checklist to guide metric selection, construct validity, and LaaJ deployment.[1]

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 flexible applications, ranging from poetry to medical report generation. This shift has reshaped evaluation paradigms and introduced new challenges, such as dealing with confabulations, ensuring factual consistency, and mitigating biases (Gehrmann et al., 2023). Consequently, metrics have evolved beyond simple lexical overlap. Newer semantic-aware metrics have emerged, including BERTScore (Zhang et al., 2019) and BLEURT (Sellam et al., 2020), and most notably, the surge of reference-free, prompt-based LaaJ methods (Gao et al., 2025; Gu et al., 2025; Li et al., 2024). Figure 1 shows that LaaJ adoption surged from less than 1% of papers before 2023 to 46% in 2025, surpassing human evaluation, which has remained flat-to-declining. Yet only 16% of 2025 papers explicitly validate LaaJ against human evaluation, exposing a validation gap that we

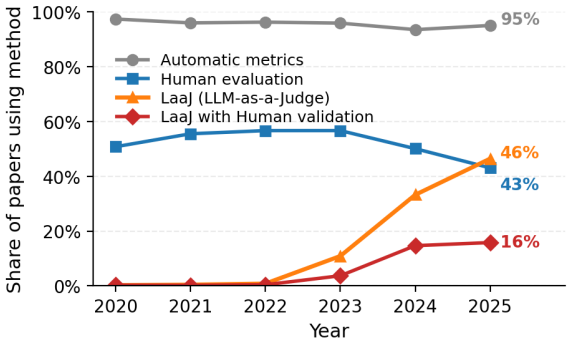

Figure 1: Evaluation method adoption across the 3,334 Top-30 NLG-task papers (2020–2025). Automatic metrics remain ubiquitous (∼95%). LaaJ adoption surged from less than 1% before 2023 to 46% in 2025, surpassing human evaluation, which declined from 51% to 43%. Yet only 16% of papers in 2025 explicitly validate LaaJ against human evaluation, indicating a validation gap.

[*]Work done at TU Berlin and BIFOLD.
[1]Code and data are available at: https://github.com/jingyng/NLG-Evaluation-Trends.

quantify throughout this work.

Motivated by the need to better understand how text quality evaluation is conducted in NLG research, we employ LLMs as research tools (Liao et al., 2025) to analyze the literature in major NLP conferences. This approach is critical to bridge the gap between scale and depth: it allows us to filter papers with a semantic accuracy that keyword searches lack, and extract structured metadata at a volume impossible for manual review. Our contributions are summarized as follows:

- We present the largest study of NLG evaluation to date, applying multi-LLM extraction to 14,171 papers, and analyzing 3,334 NLG papers. Our human validation of the results reveals strong agreement between human and LLM annotators.
- We identify three persistent challenges in NLG evaluation: metric inertia, where lexical metrics persist alongside semantic alternatives; a metric–criteria mapping problem, where general-purpose metrics appear broadly across tasks rather than systematically pairing with specific evaluation constructs; and a validation gap where LaaJ has expanded without proportional human validation.
- We provide an evaluation checklist grounded in our findings, intended as concrete starting points for more rigorous NLG evaluation.

2 Related Work

LLMs for Annotation of Research Papers. Tan et al. (2024) point out that LLMs can reliably annotate instructions and responses, even for specific domains, with human-level quality. A few prior works have dealt with the automated processing of scientific literature (Agarwal et al., 2025; Du et al., 2024; Scherbakov et al., 2025). Existing 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 systematic, large-scale annotation of research papers, not only for paper relevance filtering but also for quantifying research trends and providing critical overviews.

Surveys in NLG evaluation. 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). 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). While recent manual surveys identified metric shortcomings across 110 papers (Schmidtova et al., 2024), we use multi-LLM extraction to massively scale annotations to thousands of papers across a six-year span.

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 total, 14,171 papers were collected (full statistics in Table 18 of Appendix C). Next, we describe our paper annotation process as shown in Figure 2.

3.1 LLM Annotation

We extract the full text of each paper from PDF format using GROBID[2]. We then feed the text (excluding title and abstract, as the full text should contain all information we need) into a model with our designed prompt. For each paper, we prompted an LLM acting as an "expert NLP researcher with deep experience in Natural Language Generation" to read the text and return a structured JSON object answering four binary questions about NLG-related information: (Q1) Does the paper address NLG tasks? (Q2) Does the paper use automatic metrics to evaluate the generated outputs? (Q3) Does the paper use large language models (LLMs) as judges (i.e., after generation, LLMs are used to assess the outputs)? (Q4) Does the paper conduct human evaluations of the generated outputs?

[2] https://github.com/kermitt2/grobid

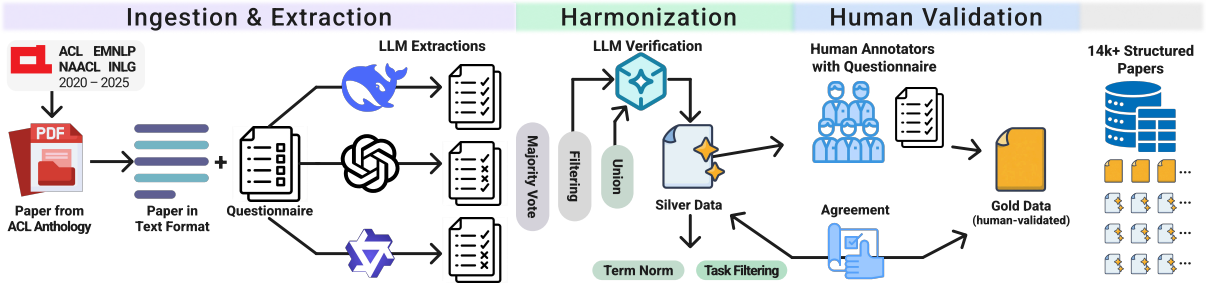

Figure 2: Our paper annotation pipeline (§3), including converting PDF to text (§3.1), extraction of metadata based on four NLG evaluation questions (Table 15), harmonization based on majority vote, filtering, and merging (§3.2), and human validation (§4), yielding 14k structured papers including a subset of 110 papers with gold annotations.

If the "`answer`" to a question is "`Yes`", we ask the LLM to extract further metadata, e.g., a verbatim excerpt as evidence, lists of datasets, languages, models, and criteria. If the "`answer`" is "`No`", all other fields in that section are set to empty. The prompt and full metadata list are provided in Appendix E.1 and Table 15.

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 following a given guideline.

Model Selection 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

We aggregate the results produced by the three LLMs to create a unified final output. We use majority voting on the four binary questions, and the union of three metadata lists. We filter out papers that have *no* to Q1, resulting in 8,665 initial NLG papers (61%). For these papers, we further employed another LLM (DeepSeek V3.1 Terminus) to verify the extraction with another prompt. The prompt is designed to verify each extraction against the full paper text to validate the four binary questions, and perform four actions: (1) **Normalize** metadata to use canonical forms (e.g., all variants of BLEU (bleu, Bleu, BLEU-4, etc.)); (2) **Correct** any incorrect items correct; (3) **Add** any missing important items; and (4) **Remove** any irrelevant or incorrect items from the list for each field listed in Table 15 (see full prompt in Appendix E.2).

Comparison of LLM annotations To compare the difference among LLM annotations, we compute the agreement with Krippendorff's α (pairwise agreement between the LLM-harmonized one and the others, shown in Table 16 in Appendix). The overall agreements are high for all four answers (Q1: 0.710, Q2: 0.688, Q3: 0.805, and Q4: 0.812).

3.3 Extraction of LaaJ-Human Validation

To determine how many papers validated LaaJ against human evaluation, we analyzed 433 NLG papers (12.3% of 3,334 filtered NLG papers detailed in §3.4) that employed both methods. Because our initial pipeline did not extract comparison metadata, we designed a supplementary prompt using DeepSeek V3.1 Terminus (Appendix E.3). This prompt extracts: (1) a binary indicator of whether a comparison exists; (2) the comparison method (e.g., correlation, agreement); (3) the statistical metrics used (e.g., Pearson, Spearman, accuracy); and (4) the quantitative results (e.g., correlation scores, significance). Crucially, we required all extracted results to be mapped to specific evaluation criteria.

3.4 Post-processing Extractions

Term Normalization To group term variants referring the same value, we perform term normalization. For tasks, datasets, models, languages, and automatic metrics, we apply two normalization steps. (1) Spelling unification and (2) Term merging with fuzzy-matches to merge near-duplicates (e.g., *Dialog Generation* → *Dialogue Generation*). For detailed term normalizations and examples, please check Appendix B.2 and B.3.

Criteria Mapping via QCET Because fuzzy matching often over-merges distinct criteria (e.g., *factual correctness* vs. *factual consistency*), we instead classified each raw criterion into the QCET

taxonomy (Belz et al., 2024) (a 111-leaf lattice structured by frame-of-reference × type × aspect) using a two-step LLM pipeline. In Step 1, the LLM mapped each criterion to existing QCET leaves, assigning a match confidence of "strong", "partial", or "none". For criteria lacking a strong match, we clustered their embeddings using HDBSCAN to identify gaps in the taxonomy. This revealed four new criteria clusters, to which we added two more based on high overall frequency. We mapped these six new leaves to compatible parent branches (Table 3), expanding the taxonomy to 117 leaves. We also introduced two auxiliary categories: *Overall Quality* (for holistic judgments) and *Other / Unclassifiable* (which is dropped from analysis). In Step 2, the LLM re-classified all "partial" and "none" criteria against this extended taxonomy, using their Step 1 labels and embedding clusters as additional signals. To validate this mapping, two annotators evaluated a stratified random sample of 155 criteria on a 1–5 Likert scale (1=wrong, 5=perfect). The average scores from annotators were 4.45 and 4.81, indicating an acceptable mapping. More implementation details are in Appendix B.2.

NLG Task Filtering Our human annotations (§4) reveal that most disagreements regarding Q1 (NLG vs. non-NLG) come from papers involving either classification tasks or generation tasks that produce non-natural-language outputs. We narrow our scope of NLG as tasks that produce open-ended natural language outputs evaluated on linguistic and semantic criteria (e.g., fluency, coherence, and relevance), and deliberately exclude classification tasks and generation tasks where outputs are artifacts verifiable by execution or exact correctness (detailed reasons for these exclusions are provided in Table 11). After normalizing and ranking the tasks by frequency, we systematically exclude 15 tasks (see Appendix B.4 for the full list) and retain the top 30 tasks for our analysis (3,334 papers), shown in Figure 7.

4 Human Validation of Extractions

4.1 Validation of LLM extractions

Human annotations To assess the quality of the LLM extractions, we manually annotated 110 randomly selected papers that were identified as NLG research by a majority LLM vote on Q1. Our annotation guidelines were modeled after the LLM prompts, but included additional details and examples (Appendix F). The process was conducted by 11 NLP researchers (3 Master's students, 2 PhD students, and 6 Postdocs). To ensure reliability, each paper received two independent annotations, with any disagreements resolved by a third annotator. With the human-annotated ground truth, we compare LLM-extractions against it.

Pair	Q1	Q2	Q3	Q4
Human R1 vs. Human R2	81.82	80.00	89.09	91.82
Human vs. DeepSeek-R1	74.55	73.64	95.45	**92.73**
Human vs. GPT-OSS-120B	75.45	**75.45**	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	**77.27**	74.55	**95.45**	90.00

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

Binary-answer agreement When evaluated against the human ground truth on the four binary questions (Table 1), Q3 and Q4 demonstrate higher agreement than Q1 and Q2. This mirrors the inter-LLM agreement (Table 16) and suggests that the adoption of LaaJ and human evaluation is easier to identify than the use of NLG tasks and automatic metrics. Notably, the harmonized LLM output (§3.2) achieves the highest overall agreement with human, validating our harmonization step.

The lower agreements on Q1 and Q2 are mainly due to lower precision (Q1-Q4: 0.78, 0.72, 0.84, 0.75), as the recall is near-perfect on all four questions (0.98, 0.96, 1.0, 1.0). This means extraction errors are almost exclusively false positives, consistent with our high-recall design goal of not missing NLG papers. Moreover, these false positives are coupled: most Q2 and Q4 errors occur on papers the annotators classified non-NLG (which entails "no" downstream); among papers where human and LLM agree on Q1, agreement rises to 92.5% (Q2) and 96.2% (Q4), and the NLG task filtering removes many of the remaining Q1 false positives.

Metadata-level extraction quality To inspect extraction beyond the binary answer level, we sampled 60 of the 110 papers and asked annotators to score how closely each LLM-extracted metadata matches their own annotation, on a 1–3 Likert scale (1=mismatch, 2=partial match, 3=full match). All 11 fields exceed an average score of 2.5, ranging from 2.52 (NLG models, output formats) to 2.91 (LaaJ models). LaaJ models, languages, and LaaJ criteria score highest; NLG models and output formats lowest (Appendix Table 13).

4.2 Validation of non-NLG paper filtering

To test whether LLMs correctly filter out non-NLG papers, we randomly sampled 120 papers from the filtered-out "non-NLG" set; 6 NLP-researcher annotators each labelled 40 papers, and each paper received two annotations. Both annotators agreed with the rejection for 117 of the 120 papers (97.5%), with only three disagreements, confirming that LLMs correctly filter out non-NLG papers in almost all cases.

4.3 LaaJ-human validation extraction quality

To evaluate the supplementary LLM extractions for LaaJ-human validation, nine NLP researchers annotated a sample of 90 papers. For binary classification (presence of validation), human annotations achieved 88.9% agreement with the LLM. However, when extracting fine-grained criterion comparisons, humans identified an average of 13.86 distinct data points per paper compared to the LLM's 4.31. Our manual inspection indicates that under-extraction is caused by PDF-to-text conversion rather than the LLM. Tabular content is largely lost during parsing, while text mentions of aggregate scores survive. Therefore, we use LLM extraction for presence of validation and rely on human annotation for the per-criterion alignment analysis in §5.3. More evaluation details are in Appendix F.

5 Results and Analysis

To quantitatively analyze the extracted results, we use two co-occurrence measures defined over pairs of terms (a term being a task, a metric, or a LaaJ/human evaluation criterion). **Frequency** $P(B \mid A)$ is the proportion of papers containing term A that also contain term B; it captures how widely B is used within the subset of papers that use A. The **Likelihood Ratio** (LR) $LR(A \rightarrow B) = P(B \mid A)/P(B \mid \neg A)$ measures how much more likely B is to appear in papers using A than in papers not using A, capturing the distinctiveness of the co-occurrence. For example, in our data 81.5% of MT papers report BLEU ($P(\text{BLEU} \mid \text{MT}) = 0.815$), and BLEU is 2.6× more likely in MT papers than in non-MT papers ($LR(\text{MT} \rightarrow \text{BLEU}) = 2.63$). Despite the shared name, our LR is the relative-risk ratio from epidemiology, not the statistical likelihood-ratio test; full formulas for the three LR variants used in this paper appear in Appendix B.6.

Statistical significance testing Because LR can be inflated for rare co-occurrences, we pair it with the Dunning G^2 log-likelihood ratio test (Dunning, 1993). We then apply Benjamini–Hochberg FDR correction (Benjamini and Hochberg, 1995) across all tested pairs within each analysis family and retain only pairs with corrected $p \leq 0.05$.

5.1 Temporal Trend Analysis

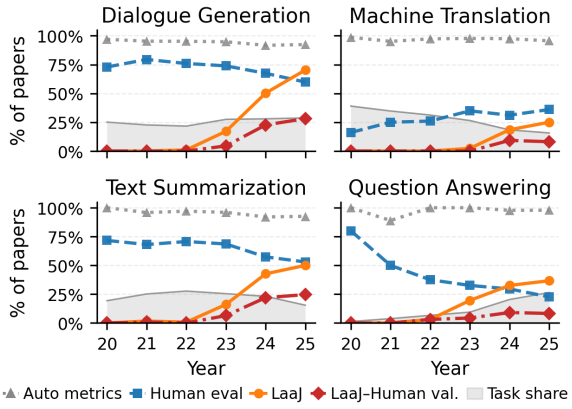

Figure 3: Per-task evaluation method adoption (%) of top-four NLG tasks per year. The *filled area* shows that task's share of all NLG papers each year.

Figure 3 presents the evolution of evaluation practices across the four most prominent NLG tasks, which account for 78.1% of the papers: Dialogue Generation (DG; 26.1%), Machine Translation (MT; 25.4%), Text Summarization (TS; 22.0%), and Question Answering (QA; 13.9%). We observe three key trends. First, automatic metrics remain universally applied (near 100%) across all tasks. Second, traditional tasks like MT rely almost entirely on these automatic metrics, with only 26% conducting human evaluation and negligible LaaJ adoption until 2025. Third, in tasks like DG, TS, and QA, we observe a potential substitution effect: as LaaJ adoption has rapidly accelerated since 2023, human evaluation has either dipped (DG, TS) or plummeted entirely (QA). Importantly, across all tasks, the validation of LaaJ against human evaluation remains relatively low compared to its usage (§5.3).

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 8 most frequent metrics and criteria for that year. Figure 4 visualizes the evolution of evaluation methods across the four major tasks.

Dialogue Generation As Dialogue Generation (DG) shifts from simple request fulfillment to-

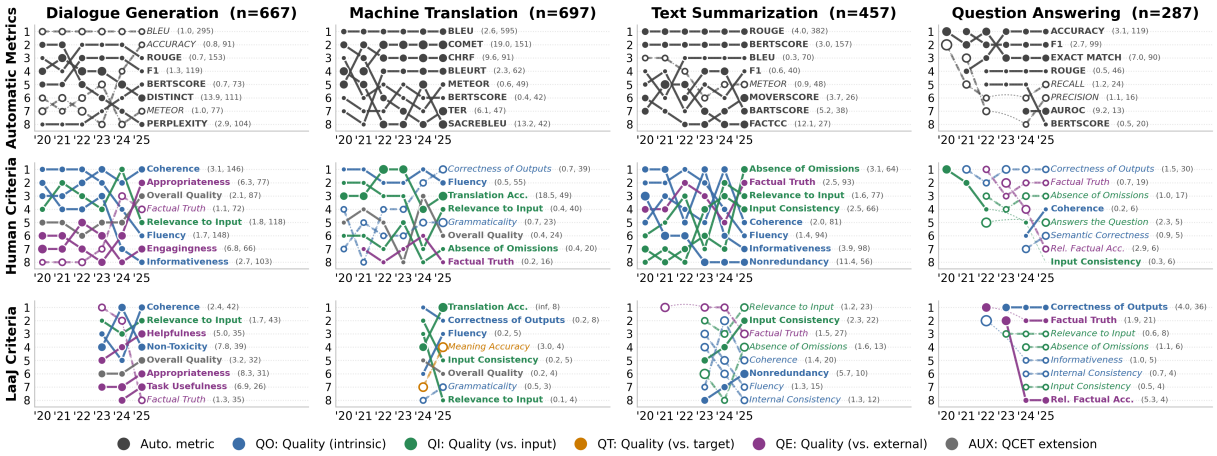

Figure 4: Evaluation trends for the four most frequent NLG tasks (columns) under three evaluation paradigms (rows). Each panel is a separate bump chart: the y-axis gives the annual frequency rank (1 = most used) of the eight most common metrics for that task, from 2020 to 2025 on the x-axis. Marker size scales with the metric's likelihood ratio (LR) for the task in that year. Solid lines with filled markers mark a significant overall association with the task; dashed lines with hollow markers mark metrics that are frequent but have a low LR. Endpoint labels give the metric name and `(overall LR, k)`, where k is the number of papers. COMET's `(19.0, 151)` in the MT/Automatic Metrics panel thus means it appears in 151 MT papers and is 19 times more likely to appear in MT than elsewhere; it enters at rank 4 in 2021 and holds rank 2 from 2022 onward, behind BLEU. Dotted arcs bridge years in which a metric drops out of the top eight; dashed connectors attach labels to metrics that are outside the top eight in 2025. A missing line in a given year means the metric was not used. Color denotes the QCET L1 dimension: QO (intrinsic to output), QI (relative to input), QT (relative to target reference), QE (relative to external standard), and AUX (QCET extensions).

ward complex, open-ended interactions, its evaluation landscape has correspondingly diversified (Figure 8). Among automatic metrics, *BLEU* remains the most frequent, though it exhibits a low task association (LR). Conversely, *Distinct*, which measures diversity, shows the highest LR with DG. Traditional performance metrics like *Accuracy* and *F1* also exhibit low LRs, likely diluted by their increasing use in other tasks. For human evaluation, *Engagingness* and *Appropriateness* are the most strongly associated criteria, even though *Coherence* is used more frequently overall. Notably, the reliance on *Fluency* and *Informativeness* has declined sharply, likely reflecting the baseline performance gains of modern LLMs. When comparing paradigms, both LaaJ and human evaluation emphasize *Appropriateness*. However, LaaJ uniquely prioritizes alignment-focused criteria such as *Helpfulness* and *Non-Toxicity*. This indicates that LaaJ is introducing novel safety dimensions, revealing a potential gap in human evaluation practices.

Machine Translation MT primarily uses *BLEU* (87.4%) for evaluation (while rather low association (since it is also heavily used in other tasks), with *COMET* also established as a major metric. The latter also has the highest association (LR=19) with MT since 2023. In terms of evaluation criteria, *Correctness of Outputs* and *Fluency* are among the most frequent ones, despite their low LR. *Translation Accuracy* is ranked highest in association (the criterion is unique to the MT task). The dominance of these n-gram and reference-based metrics illustrates deep-seated inertia (formally defined in §6). Despite the availability of LLMs, the LaaJ Criteria row is sparse, without any activity until 2024. Overall, the evaluation paradigm in MT has not shifted as much as the other tasks.

Text Summarization In TS, *ROUGE* and *BERTScore* dominate automatic evaluation, both showing high task associations (LRs). *FactCC* has an increased 2021–2022 frequency for evaluating hallucinations and retains the highest LR despite a recent drop. Notably, *BLEU* is frequently used but lacks significant task association. In human evaluation, *Nonredundancy* exhibits the highest LR despite low usage, while *Absence of Omissions* and *Factual Truth* are increasingly prioritized over declining criteria like *Fluency* and *Informativeness*. Lastly, most LaaJ criteria currently show no significant statistical association with TS.

Question Answering Because QA typically generates relatively short responses, its evaluation paradigm diverges substantially from the other

three tasks. It often resembles a classification problem, relying heavily on exact-match comparisons against a ground truth. Consequently, QA exhibits the least diverse set of evaluation metrics and criteria. Most criteria have no significant LR with QA, suggesting a lack of task-specific evaluation designs. The predominant criterion is *Correctness of Outputs*, particularly when employing LaaJ. This trend is concerning: although QA has experienced the most rapid growth, its evaluation remains heavily reliant on simplistic string matching, misaligned with the complexities of open-ended QA.

5.2 Metric-criteria Association by Evaluation Methods

Although task-specific LRs identify the metrics and criteria most associated with each NLG task (Figure 4), this term-level association does not imply paper-level pairing. A highly associated metric and criterion may both be common in a task's corpus without ever co-occurring in the same paper. To study their co-occurrence and association level, we compute the within-task pair association LR for every (metric, criterion) pair, as most metrics do not explicitly mention the criterion it measures[3]. We calculate separately for the LaaJ and human evaluation subsets of the task's papers (Figure 5).

Significant metric-criterion associations concentrate almost exclusively in the frequency $\times$ frequency quadrant. Across DG and TS, *BLEU*, *ROUGE*, *BERTScore*, and *METEOR* pair significantly with *Coherence*, *Fluency*, and *Relevance to Input*, indicating a default evaluation set shared across tasks. The distinctiveness $\times$ distinctiveness quadrant contains no significant cells: task-distinctive metrics and criteria are well attested individually but rarely co-reported. A few distinctive metrics do reach high LR against frequent criteria. In TS, *FactCC* pairs near-exclusively with *Factual Truth*, consistent with its design as a factuality metric. The orange-versus-blue split of these cells is itself informative: *UniEval*'s associations reach significance only among LaaJ papers, while MT's four significant cells are all human-side and all involve *Fluency*; indicating a divergence between LaaJ and human evaluation paradigms. MT and QA show much weaker structure: no metric in MT reaches significance against the task-distinctive criteria (*Translation Accuracy*, *Meaning Accuracy*, *Grammaticality*); QA yields no significant pairings under either method.

These results indicate that NLG papers choose their metric and criterion largely independently, each defaulting to frequent measures rather than pairing task-specific ones. This decoupling raises a general construct-validity concern: the chosen metrics rarely reflect a deliberate methodological match to the reported criteria.

5.3 LaaJ with Human Validation

Using the extraction methodology from §3.3, we extracted validation results from 433 papers employing both LaaJ and human evaluation. Only 254 of these (58.7%) explicitly validate LaaJ against human judgments, representing 37.4% of the 679 NLG papers using LaaJ, under 8% of total 3,334 NLG papers. Among the remaining 179 papers without an explicit comparison, 76.3% still measure at least one shared criterion across LaaJ and human, suggesting comparison was generally feasible but unreported, rather than the two methods deliberately covering disjoint aspects. Figure 6 visualizes per-criterion alignment using the 70 human-annotated subset of papers (§4), restricted to QCET-mapped criteria with more than 10 comparison values. As shown in Figure 6, evaluating our human-annotated subset reveals a disparity in LaaJ reliability. While automated judges achieve moderately high median correlations for aggregate *Overall Quality*, their alignment with human annotators degrades significantly on nuanced, fine-grained criteria. Dimensions such as *Fluency* and *Coherence* exhibit median correlations near or below 0.4, accompanied by large variance, demonstrating that LaaJ struggles to reliably capture targeted linguistic constructs. Furthermore, as often cautioned in the NLG community, evaluating surface-level quality is insufficient for determining real-world utility; our data confirms that LaaJs fail to align with humans on *task-specific usefulness*.

5.4 Analysis on the Decline of Human Evaluation

To investigate the decline of human evaluation, we disentangle which factors explain for the presence of human evaluation with a regression analysis. We model presence of human evaluation as a dependent variable; year, task (top-four tasks), venue and presence of LaaJ as the independent variables.

[3]We manually annotated 50 randomly sampled papers containing automatic metrics (183 metric mentions, two annotators, $\kappa = 0.49$) to confirm this. Only 21–36% of metric mentions explicitly state which criterion the metric measures, and 6% of papers state a mapping for every metric (Appendix B.5).

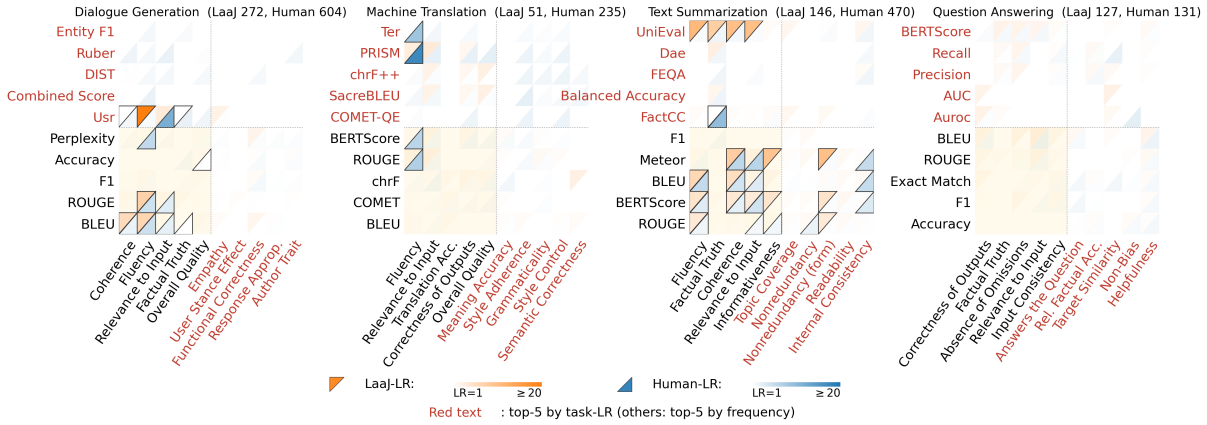

Figure 5: Within-task co-occurrence LRs for metric–criterion pairs, one panel per task. Rows list metrics and columns list criteria; each axis contains the five most frequent items (black labels) and the five with the highest task-LR (red labels), so the quadrant where frequent metrics meet frequent criteria (yellow background) is distinct from the rest. Each cell is split along the diagonal: the upper-left orange triangle covers the task's LaaJ papers and the lower-right blue triangle its human-evaluation papers. Colour intensity encodes the co-occurrence LR, and an outline marks a significant pair. In Dialogue Generation, BLEU with Fluency is outlined and shaded on both halves (133 human-evaluation papers, LR = 2.02; 19 LaaJ papers, LR = 2.59), indicating that the pairing is associated with the task under both paradigms, whereas a cell shaded on one side only reflects a paradigm-specific pairing. Across tasks, papers pair metrics and criteria along frequently used terms rather than task-distinctive ones.

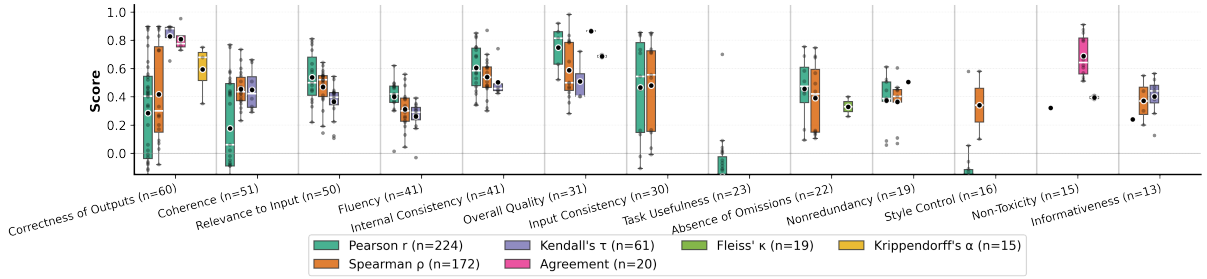

Figure 6: Distribution of LaaJ vs. human evaluation scores by metric and criterion. Individual observations are shown as gray dots. n denotes the number of studies included. Values are based on human annotated 70 papers.

Specifically, we report LR with 95% confidence intervals (a CI interval excluding 1 indicates significance) with modified Poisson regression with robust standard errors, as the presence of human evaluation is binary. With a single variable, the regression reduces exactly to LR; with the several variables listed, it yields each variable's LR adjusted for the others.

From the regression results, we can explain how each variable explains the presence of human evaluation: (1) "Task" strongly affects the presence of human evaluation, MT and QA papers are half as likely to conduct it compared to TS and DG, consistent with our findings; (2) "Venue" has no significant impact on human evaluation trends; (3) "Year" has a significant impact in 2025, consistent with our claim that human evaluation declines in 2025; (4) "LaaJ" results show that papers using LaaJ are more likely to report human evaluation, thus increase of LaaJ usage does not mean its substitution of human evaluation.

Variable (reference)	**LR [95% CI]**
LaaJ use	**1.28** [1.16, 1.41]
Year 2021 (vs. 2020)	1.11 [0.98, 1.27]
Year 2022	1.08 [0.95, 1.23]
Year 2023	1.12 [0.99, 1.27]
Year 2024	0.97 [0.84, 1.11]
Year 2025	**0.86** [0.74, 1.00]
Machine Translation	**0.49** [0.42, 0.56]
Text Summarization	**1.15** [1.02, 1.30]
Dialogue Generation	**1.16** [1.02, 1.31]
Question Answering	**0.51** [0.43, 0.61]
Venue EMNLP (vs. ACL)	0.97 [0.90, 1.05]
Venue INLG	1.08 [0.91, 1.28]
Venue NAACL	1.04 [0.94, 1.15]

Table 2: Regression result. Bold: 95% CI excludes 1, significant. Year and venue are exclusive categories, compared to a reference (2020 / ACL); task indicators are not exclusive (a paper can address several tasks), so each task LR compares papers with vs. without that task

6 Challenges and Guidelines

Metric Inertia A persistent challenge in NLG evaluation is the reluctance to discontinue lexical-overlap based metrics, a pattern we term metric inertia. Part of this inertia is driven by the structure of legacy datasets, which typically provide a single "gold standard" reference. Researchers default to legacy metrics because they offer a fast, deterministic way to calculate overlap against this ground truth. However, modern LLMs excel at stylistic variation, paraphrasing, and generating valid alternative answers that do not share the exact vocabulary of a single human annotator (Liu et al., 2021). By continuing to anchor evaluations to lexical overlap, NLG field risks optimizing for surface-level reference mimicry rather than actual semantic quality. Even when a ground-truth reference is available, researchers should prioritize reference-based semantic metrics (e.g., BERTScore, BLEURT) that reward meaning over rigid vocabulary matching.

Metric-Criteria Mapping Beyond temporal inertia, automatic metrics are applied across both tasks and criteria. Cross-task, legacy lexical metrics like BLEU and ROUGE are used across all four focus tasks with weak LR (Figure 4); within a task, they pair with multiple frequent quality criteria (Figure 5). Researchers therefore select metrics based on popularity or tradition rather than the construct under evaluation, creating conceptual entanglement and confirming the lack of standardization noted by Belz et al. (2025). The widespread reporting of established metrics as generic proxies for "quality" limits their interpretability (Sai et al., 2022); without explicitly binding a metric to a specific linguistic or semantic property, it is difficult to determine which distinct aspect of generation is driving an observed improvement.

Validation Gap The accelerating adoption of LLM-as-a-Judge (LaaJ) has outpaced its empirical validation. Our data indicates that while LaaJ usage has now surpassed human evaluation (Figure 1), direct comparative validation on the same data remains infrequent. More notably, the existing validation evidence skews heavily toward aggregate measures like *Overall Quality* and *Correctness of Outputs*, while demonstrating high variance and weaker alignment on fine-grained criteria. This validation gap introduces a methodological vulnerability: by increasingly relying on automated evaluators without construct-specific human baselines, we risk standardizing benchmarks that may silently diverge from human perception on nuanced dimensions like *Task usefulness* (Figure 6).

NLG Evaluation Checklist To help researchers translate these challenges into practice, we propose a compact Evaluation Design Checklist and task-specific recommendations for DG, MT, TS, and QA (Appendix D).

1. Metric Selection

- ☐ **Are you evaluating open-ended generation?** If yes, acknowledge that optimizing for exact n-gram overlap with a single reference penalizes valid paraphrases. Shift legacy metrics to secondary baselines.
- ☐ **Have you included a semantic metric?** Ensure legacy metrics (BLEU, ROUGE) are supplemented, or replaced, by semantic metrics (e.g., BERTScore, COMET, BLEURT) that reward meaning over lexical matching.

2. Construct Validity

- ☐ **Is the target construct explicitly named?** Verify that every reported metric is explicitly tied to a specific quality criterion (e.g., "We use QAFactEval to measure factuality," not just "to measure quality").
- ☐ **Is the metric mathematically suited for the construct?** Ensure you are not using lexical overlap (e.g., ROUGE) to substantiate claims about high-level semantics (e.g., logical coherence, factual consistency).

3. LaaJ Deployment

- ☐ **Have you validated your LaaJ against human judgments?** Where a public meta-evaluation benchmark covers your target criterion, report your judge's correlation against it. Otherwise, annotate a subset per criterion and report appropriate correlation or agreement metrics.
- ☐ **Are human experts retained for sensitive or poorly correlated criteria?** Acknowledge that LaaJ reliability often collapses on strict linguistic nuances (e.g., *task usefulness*). For these criteria, rely on expert human evaluation as your primary signal.

7 Conclusion

We present a large-scale quantitative analysis of 3,334 NLG papers (2020–2025) across four NLP conferences, utilizing a human-validated, multi-LLM extraction pipeline. Our analysis indicates that NLG evaluation methodologies are failing to keep pace with the capabilities of modern generation models. We highlight three recurrent issues: the persistence of metric inertia, metric-criteria mapping problem, and a severe validation gap in LaaJ evaluation. Crucially, our human-annotated subset demonstrates that while LaaJ approximates aggregate quality, its alignment with humans degrades severely on fine-grained criteria. To ensure evaluation protocols mature alongside NLG capabilities, we contribute an Evaluation Design Checklist, urging researchers to explicitly map metrics and rigorously validate LaaJ on fine-grained dimensions.

Limitations

We list a few limitations that should further improve the paper in future research:

Multi-task papers: some papers deal with multiple tasks; however, our annotation guideline did not require users/LLMs to extract metadata per task (models, language, and evaluation metrics), thus, there can be some deviations in reported numbers.

Definition boundary of metrics and LaaJ: the definition of automatic metrics and LaaJ is not clearly separable. Some learned metrics that produce a score may have been considered as LaaJ by LLM-extraction; while performing annotation, humans do not always reach an agreement.

Task granularity: 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.

Extracted metadata evaluation: Our human evaluation of LLM-extraction on the metadata-level assigns a 1-3 likert scale (1=mismatch, 3=match); however, more fine-grained criteria could improve the evaluation, for example, by asking the type of errors (missing item, hallucinated item or paraphrased item). We leave this to future work for improving LLM information extraction.

Criteria mapping: when mapping the raw extracted criteria to the QCET taxonomy, the input to LLMs is only the criterion name with no context information; further information about the context can be extracted from papers (although some papers might not have such context, such as definition of their criteria) may help with better mapping.

Ethical Statement

The human annotators are NLP researchers who are either co-authors of this paper or students who are hired by their host institutions. AI (Claude) was used for coding (computing LR and agreements) and figure polishing, and in writing for grammar and sentence coherence improvement. All AI-assisted content were reviewed and verified by the authors.

Acknowledgments

We thank the anonymous reviewers at ACL Rolling Review for their valuable feedback. We also thank the additional human annotators (Awsam Agbaria, Alexandra Tsiakalou, Jiaao Li, Neda Foroutan) for their valuable time and effort, and our colleague Shushen Manakhimova for her constructive comments on the manuscript. This research is funded by the Berlin Institute for the Foundations of Learning and Data (BIFOLD) and the Federal Ministry of Education and Research (BMFTR) in the scope of the research project VeraXtract (ref. 16IS24066) and GenKI4Media (ref. 01MK25004E).

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. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12236–12264, Suzhou, China. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, 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.

Anya Belz, Simon Mille, Craig Thomson, and Rudali Huidrom. 2024. QCET: An interactive taxonomy of quality criteria for comparable and repeatable evaluation of NLP systems. In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 9–12, Tokyo, Japan. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300.

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, Zhihong 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 Srinath, 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.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. 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 Sellam. 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 LLM-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.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. MetricX-25 and GemSpanEval: Google Translate submissions to the WMT25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968, Suzhou, China. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, 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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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*.

Ruibo Liu, Jason Wei, and Soroush Vosoughi. 2021. Language model augmented relevance score. 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 6677–6690, Online. Association for Computational Linguistics.

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, Sebastien Bubeck, and 108 others. 2025. GPT-OSS-120B & GPT-OSS-120B model card. *Preprint*, arXiv:2508.10925.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, 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, Chengen 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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

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. Association for Computational Linguistics.

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*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 [4] for the other LLMs inferences. We set temperature as 1 for all models. Total cost is around 50 dollars.

[4] https://novita.ai/

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) using SequenceMatcher[5] at threshold 0.9. 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 map the 7418 unique criterion strings (combining LLM and human extractions) into the QCET quality taxonomy (Belz et al., 2024), a three-level lattice (L1 frame-of-reference × L2 type × L3 aspect) with 111 leaf criteria organised under four frame-of-reference dimensions at L1: QO (Quality of outputs in their own right), QI (relative to input), QT (relative to target outputs), and QE (relative to an external frame of reference). Our pipeline runs in three steps. **(1) Raw criteria to QCET criteria classification.** An LLM classifies each raw variant against the 111 QCET leaves. The model returns either a QCET leaf with a confidence indicator (*strong* or *partial*), or *none* with a short justification. Criteria are processed in batches of 5; **(2) Clustering to find new leaves.** Criteria with *partial* or *none* match are embedded with `sentence-transformers/all-MiniLM-L6-v2` and clustered with HDBSCAN, producing 19 thematic clusters plus a noise one with 5,111 criteria. For each of the 19 clusters, an LLM call returns a verdict: keep (construct not captured by any QCET leaf), map (back to a existing leaf covers it), split (cluster mixes distinct criteria to be mapped separately), or drop (extraction noise, not a criterion). Four clusters become new QCET leaves placed at appropriate lattice positions: *Absence of Toxic / Harmful Content* (QOC-c-2), *Engagingness / Interestingness* (QEF-w-10), *Absence of Bias / Stereotypes* (QEC-c-3), and *Empathy / Emotional Appropriateness* (QEF-w-11). We additionally add two leaves based on high frequency in the noise cluster: *Creativity / Non-creativity* (QOF-w-6) and

[5] https://docs.python.org/3/library/difflib.html

Refusal Appropriateness (QEC-w-2). Together these six extensions take QCET from 111 to 117 leaves (Table 3). We also introduce two auxiliary categories: *Overall Quality / Preference* for holistic judgements that explicitly refuse decomposition into specific aspects, and *Other / Unclassifiable* for extraction noise (dropped from analysis). **(3) Re-classification.** Each criterion that were not a strong match in step 1 is mapped against the unified target set of 117 QCET leaves plus the two auxiliary categories.

Language We map the languages according to the list of ISO language codes[6] and keep the standard ISO language names; however, many multilingual 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 the years of release.

B.3 Example of Term Normalization Mapping

We list our term normalization results for the top 10 most frequent terms in Table 5-Table 10.

B.4 NLG Task Filtering

Table shows the filtered out tasks among the top-ranked ones (we rank them based on frequency, and keep the top-30 after filtering) and the reason for exclusion.

We further show the top-3 automatic metrics for the top excluded tasks in Table 12. The top metrics of these tasks confirm our exclusion reasoning: they are dominated by functional-correctness metrics, whereas our included NLG tasks are topped by text-quality metrics.

B.5 Human Annotation Details

Details of human-annotation studies referenced in §4 and 4.3 are presented below. The full annotation guidelines are in our project repository.

[6] https://en.wikipedia.org/wiki/List_of_ISO_639_language_codes

Annotator pool. All 11 annotators are NLP researchers (NLG-related expertise) recruited through the authors' institutions; participation was either by co-authorship or paid student work (see Ethical Statement).

Field-level extraction quality. For each of 60 papers (sampled from the 110), an annotator scored each LLM-extracted metadata field on a 1–3 scale: 1 = mismatch (extraction is wrong or substantially incomplete); 2 = partial match (some correct entries, some missing or off); 3 = full match (extraction matches the annotator's reading). Field-level mean scores are in Table 13. All 11 fields exceed mean 2.5; LaaJ models, languages, and LaaJ criteria score highest, while NLG models and output formats score lowest.

NLG-vs-non-NLG filter validation. We randomly sampled 120 papers from the LLM-filtered "non-NLG" set. Annotators followed the guideline "Does the paper address NLG tasks?" with positive examples (summarization, dialogue, MT, paraphrase, captioning) and negative examples (sentiment analysis, NER, extractive QA without generated text). 6 NLP-researcher annotators each labeled 40 papers, with each paper annotated twice. 117 of 120 papers (97.5%) had both annotators agreeing on "No"; the remaining 3 had one "Yes" and one "No".

Metric-criterion mapping To check whether papers actually map criteria that their metrics measure, we randomly selected 50 papers which have "yes" to Q2 (automatic metrics) from the 3,334 NLG papers. For each paper, we locate the passage where its metrics are introduced and extract for each metric the criterion it was mapped to. Each paper was annotated independently by two annotators (agreement 77.6%, Cohen's kappa 0.49).

Results from manual annotation reveal that out of 183 metrics from the 50 papers, only 28-36% of the metrics (per annotator; 21% by both annotators' consensus) mentioned which criterion it measures; only 6% of papers stated metric-criterion mapping for every metric used (by consensus; 10-12% per annotator). 22 of the 31 metrics with a consensus-stated criterion occur only once in our sample, while the most frequently used metrics are rarely mapped (BLEU 12%, ROUGE 11%, BERTScore and METEOR 0%). Among the metrics that mentioned the criteria they measure, most are designed for the specific criterion (e.g.: Distinct for Diver-

New leaf id	Name	Origin	Variants	Paper-occ.	Top raw variants (count)
QOC-c-2	Absence of Toxic / Harmful Content	cluster	95	409	harmfulness (57); safety (53); Safety (48)
QEF-w-10	Engagingness / Interestingness	cluster	48	248	engagingness (36); Engagement (28); interestingness (27)
QEC-c-3	Absence of Bias / Stereotypes	cluster	109	172	fairness (9); Bias (7); bias (7)
QEF-w-11	Empathy / Emotional Appropriateness	cluster	34	108	Empathy (46); empathy (18); Emotional Support (4)
QOF-w-6	Creativity / Non-creativity	frequency	26	157	creativity (44); Creativity (39); Novelty (18)
QEC-w-2	Refusal Appropriateness	frequency	45	63	Refusal (8); refusal (6); Abstention (2)

Table 3: The 6 extension leaves added to QCET (111 $\rightarrow$ 117). **Origin** indicates whether the leaf was discovered by HDBSCAN clustering of stage-1 classification residuals (*cluster*) or by frequency analysis of remaining unclustered residual variants (*frequency*). **Variants** is the number of distinct raw criterion strings mapped to this leaf; **Paper-occ.** sums the LaaJ and human paper-occurrences across those variants.

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 4: Top-10 normalized tasks and their top variants

sity, Knowledge-F1 for informativeness). Other frequently used metrics have no unique construct: when BLEU, ROUGE or Perplexity are mapped, the stated criteria differ from paper to paper (BLEU: coherence / content retention / relevance; ROUGE: relevance / text reuse / grammar-and-redundancy; Perplexity: fluency / coherence / style). This confirms that general-purpose metrics are widely used as quality estimators without agreed and validated constructs.

LaaJ-against-Human extraction. To evaluate the supplementary LLM extractions for LaaJ-human validation, we sampled 90 papers (72 predicted "`Yes`" and 18 "`No`"). Nine NLP researchers annotated them, each reviewing 20 papers to ensure two annotations per paper. For the binary classification (presence of validation), inter-annotator agreement was 75.6% (Cohen's $\kappa = 0.45$). After merging the human annotations by taking the union of extracted rows per paper, the human–LLM agreement reached 88.9% ($\kappa = 0.67$). However, for the per-paper count of distinct extracted rows (metric, criterion, judge-model), the exact match rate was only 21.1% (Pearson $r = 0.32$). Restricting to papers with validation content (LLM: 72; human-union: 70) and deduplicating within (paper, metric, criterion, value), humans extracted an average of 13.86 distinct rows per paper compared to the LLM's 4.31. For the rows the LLM does extract, value fidelity is high: 91% of LLM-extracted values lie within ± 0.05 of a human-annotated value for the same (paper, criterion, metric) cell. The bias is concentrated on aggregate versus fine-grained criteria: for *Overall Quality / Preference*, LLM and human extract the same number of rows (31

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); ϵ-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 5: Top-10 normalized automatic metrics and their top variants

QCET Target (id)	Cnt	Top Variants (count)
Correctness of Outputs (QOC-w-1)	480	correctness (173); accuracy (111); Correctness (83)
Factual Truth (QEC-c-1)	465	Factuality (36); factuality (32); factual consistency (25)
Relevance to Input (QIG-c-3)	438	relevance (187); Relevance (100); adequacy (7)
Coherence (QOG-c-3)	279	coherence (93); Coherence (67); logical coherence (8)
Consistency with Input (QIC-c-3)	260	faithfulness (43); Faithfulness (23); instruction-following (11)
Absence of Toxic Content (QOC-c-2)	241	harmfulness (42); harmlessness (35); safety (31)
Usefulness (QEG-w-3)	236	helpfulness (137); Helpfulness (77); usefulness (9)
Overall Quality / Pref. (AUX-OQ)	223	overall quality (33); quality (23); Quality (20)
Absence of Omissions (QIC-c-1)	196	completeness (38); Completeness (36); comprehensiveness (12)
Internal Consistency of Outputs (QOG-c-4)	187	consistency (56); Consistency (37); logical consistency (9)

Table 6: Top-10 LaaJ criteria after QCET-based normalization, with paper count and the most frequent raw variants that mapped to each target.

QCET Target (id)	Cnt	Top Variants (count)
Fluency (QOG-w-3)	806	fluency (477); Fluency (269); Language Fluency (6)
Relevance to Input (QIG-c-3)	805	relevance (280); Relevance (155); adequacy (64)
Factual Truth (QEC-c-1)	778	factuality (58); Factuality (42); factual consistency (36)
Correctness of Outputs (QOC-w-1)	694	correctness (211); accuracy (139); Correctness (87)
Coherence (QOG-c-3)	630	coherence (248); Coherence (163); semantic coherence (8)
Overall Quality / Pref. (AUX-OQ)	499	overall quality (108); preference (60); quality (46)
Consistency with Input (QIC-c-3)	476	faithfulness (101); Faithfulness (42); meaning preservation (28)
Absence of Omissions (QIC-c-1)	408	completeness (65); Completeness (45); comprehensiveness (18)
Informativeness (QOG-c-2)	391	informativeness (183); Informativeness (97); Informative (5)
Internal Consistency of Outputs (QOG-c-4)	333	consistency (90); Consistency (71); logical consistency (11)

Table 7: Top-10 human evaluation criteria after QCET-based normalization, with paper count and the most frequent raw variants that mapped to each target.

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 8: Top-10 normalized languages 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); HUMANEVAL (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 9: 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 10: Top-10 normalized NLG models and their top variants

each), but for every other QCET criterion the LLM under-extracts by 3–10×, with several criteria (*Usefulness for Task*, *Control over Style*, *Absence of Toxic/Harmful Content*) extracted by humans but missed entirely by the LLM. As a share of total extracted rows, 26.3% of LLM-extracted rows are tagged *Overall Quality / Preference* versus 5.9% for humans — a $4.5\times$ over-representation. Manual inspection of the LLM prompts suggests this pattern is caused by the PDF-to-text conversion step in our pipeline rather than the extractor LLM itself: tabular content, where per-criterion correlations are typically reported, is largely lost during PDF parsing, while the text mentions of aggregate scores survive.

B.6 Association Measures and LR Formulations

Likelihood Ratio (LR) We list each LR variant below. For all three, *A* and *B* are sets of papers defined by the presence of a specific term.

Metric (or criteria)–task LR: *A* = papers with a specific task, *B* = papers with a specific metric (note that $B \not\subseteq A$):

$$LR_{\text{metric|task}}(A \to B) = \frac{P(B|A)}{P(B|\neg A)} \quad (1)$$

Metric–criteria LR: *A* = papers with a specific metric, *B* = papers with a specific criterion:

$$LR_{\text{criteria|metric}}(A \to B) = \frac{P(B|A)}{P(B|\neg A)} \quad (2)$$

Excluded task	Reason
Code Generation	Not natural language output; verified by execution, not text quality
Mathematical Reasoning	Evaluated by answer correctness against gold answer, not linguistic criteria
Math Problem Solving	Same as Mathematical Reasoning
Math Reasoning	Same as Mathematical Reasoning
Text-To-SQL Generation	Query output; verified by execution accuracy
Semantic Parsing	Meaning representation; evaluated by exact/graph match
Instruction Following	Measures constraint compliance (task success), not text quality
Language Modeling	Evaluated intrinsically (perplexity)
Natural Language Inference	Classification
Text Classification	Classification
Multiple Choice QA	Discriminative selection scored by accuracy
Sentiment Analysis	Classification
Named Entity Recognition	Sequence labeling scored by span F1; no new text generated
Automatic Speech Recognition	Transcription of given content; not open-ended generation
Text-To-Speech Generation	Audio output; evaluated on acoustic criteria, not text

Table 11: Excluded Tasks and Reasons

Top excluded tasks	Top-3 evaluation metrics
Code Generation (401)	Accuracy (33%), pass@k (30%), Exact Match (15%)
Instruction Following (131)	Accuracy (41%), Win Rate (28%), ROUGE (16%)
Math Reasoning, all variants (154)	Accuracy (73%), pass@k (24%), Exact Match (14%)
Text-to-SQL Generation (41)	Execution Accuracy (46%), Exact Match Accuracy (32%), Exact Match (24%)
Semantic Parsing (55)	Exact Match (36%), Accuracy (29%), Exact Match Accuracy (24%)
Language Modeling (102)	Perplexity (78%), Accuracy (36%), BLEU (24%)

Table 12: Top-3 metrics for newly added tasks

Field	Mean (1–3)
Q1: tasks	2.62
Q1: datasets	2.71
Q1: languages	2.88
Q1: NLG models	2.52
Q1: output formats	2.52
Q2: automatic metrics	2.69
Q3: LaaJ models	2.91
Q3: LaaJ methods	2.79
Q3: LaaJ criteria	2.88
Q4: human-eval methods	2.79
Q4: human-eval criteria	2.79

Table 13: LLM-extraction field-level Likert scores (60 papers; 1=mismatch, 2=partial, 3=full match). Means above 2.5 across all fields indicate that LLM extraction substantially aligns with human annotators.

Dunning G^2 test For a 2×2 contingency table with cells $(k_{11}, k_{12}, k_{21}, k_{22})$ (where k_{11} = papers with both *A* and *B*, k_{12} = papers with *A* but not *B*, k_{21} = papers with *B* but not *A*, k_{22} = papers with neither), the G^2 statistic is:

$$G^2 = 2\sum_{i,j} O_{ij} \ln \frac{O_{ij}}{E_{ij}}, \quad E_{ij} = \frac{r_i \cdot c_j}{N} \tag{3}$$

where r_i and c_j are row and column marginals and N is the total number of papers. $G^2 \sim \chi^2_1$ under H_0 (independence), giving a p-value directly. We apply Benjamini–Hochberg FDR correction (Benjamini and Hochberg, 1995) to all p-values within each analysis family and keep pairs with adjusted $p \leq 0.05$.

Illustrative example of LR and G^2 computation Table 14 illustrates LR and G^2 on two real pairs from our data. In both cases, papers are split by whether they contain term *A* and/or term *B*, forming a 2×2 contingency table:

	B	$\neg B$
A	k_{11}	k_{12}
$\neg A$	k_{21}	k_{22}

Example A is a *task–metric* pair (A = task, B = metric) drawn from all $N = 3{,}334$ papers. Example B is a *metric–criterion* pair (A = metric, B = criterion) drawn from the $N = 1{,}711$ papers that report human evaluation. Despite near-identical LR (≈ 9.2–9.4), the two examples reach opposite conclusions because k_{11} differs by a factor of 45.

Table 14: LR and G^2 on two real pairs from our data with nearly the same LR but opposite significance outcomes. Example A (task–metric): *A* = question answering, *B* = exact match; full set ($N = 3{,}334$). Example B (metric–criterion): *A* = coverage, *B* = relative factual accuracy (QCET QEC-c-2); human-evaluation subset ($N = 1{,}711$). The large difference in k_{11} (136 vs. 3) drives the difference in E_{11} and hence G^2, producing opposite conclusions after BH-FDR correction.

	Example A *(QA → exact match)*	**Example B** *(coverage → rel. factual accuracy)*
N (papers)	3,334	1,711
k_{11} (*A* and *B*)	136	3
k_{12} (*A*, not *B*)	326	19
k_{21} (*B*, not *A*)	90	25
k_{22} (neither)	2,782	1,664
$P(B\|A) = k_{11}/n_+$	$136/462 = 0.294$	$3/22 = 0.136$
$P(B\|\neg A) = k_{21}/n_-$	$90/2872 = 0.031$	$25/1689 = 0.015$
$\mathrm{LR} = P(B\|A)/P(B\|\neg A)$	**9.4**	**9.2**
$E_{11} = n_+(k_{11} + k_{21})/N$	31.3	0.36
G^2	292.4	8.0
Raw p-value	<0.001	0.0046
BH-FDR-adj. p	<0.001	0.175
Significant (adj. $p \leq 0.05$)?	**Yes**	**No**

C Additional Results

Key	Description
Q1: NLG task presence (`answer_1`)	
`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.
Q2: Automatic evaluation metrics (`answer_2`)	
`answer`	"Yes" if automatic metrics are used.
`quote`	Excerpt mentioning automatic evaluation.
`metrics`	List of metrics (e.g., BLEU, BERTScore).
Q3: LaaJ (`answer_3`)	
`answer`	"Yes" if an LLM is used *after* 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).
Q4: Human evaluation (`answer_4`)	
`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 15: Schema of the JSON object returned by the LLM for each paper.

Model	Q1	Q2	Q3	Q4
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 α	0.7101	0.6879	0.8048	0.8124

Table 16: Pairwise agreement (%) between each LLM and LLM-harmonized results (row 1-3), and Krippendorff's α among the three LLMs.

Model	Q1	Q2	Q3	Q4
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 17: *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	227	0
ACL-2021	571	243	42.6	124	582	76	751	277	0	231	1
ACL-2022	603	275	45.6	196	686	88	842	365	5	284	5
ACL-2023	910	463	50.9	319	1202	162	1211	582	28	409	22
ACL-2024	864	591	68.4	506	1546	190	1392	841	375	586	97
ACL-2025	1600	1096	68.5	832	2596	233	2221	1667	918	1050	309
EMNLP-2020	751	291	38.7	159	578	86	910	323	0	219	1
EMNLP-2021	847	326	38.5	151	674	131	1004	410	1	270	4
EMNLP-2022	828	381	46.0	247	950	193	1062	436	14	358	23
EMNLP-2023	1047	553	52.8	395	1401	231	1297	727	148	477	37
EMNLP-2024	1268	823	64.9	634	1985	339	1735	1122	506	706	160
EMNLP-2025	1809	1374	76.0	994	3073	225	2438	2027	1064	1123	341
INLG-2020	46	44	95.7	31	96	11	137	78	3	56	1
INLG-2021	45	44	97.8	29	89	28	116	72	0	89	0
INLG-2022	25	23	92.0	24	57	12	89	44	0	47	0
INLG-2023	36	36	100.0	27	94	23	107	77	9	65	2
INLG-2024	53	51	96.2	43	100	28	212	103	32	86	7
INLG-2025	50	43	86.0	36	121	15	146	101	58	75	41
NAACL-2021	477	177	37.1	105	394	55	554	279	1	171	2
NAACL-2022	442	184	41.6	125	462	62	569	262	0	191	0
NAACL-2024	486	288	59.3	253	850	234	740	445	144	313	42
NAACL-2025	635	416	65.5	331	1131	153	1023	770	400	468	114
Total	**14171**	**8019**	**56.6**	**3308**	**11289**	**801**	**12995**	**6805**	**2752**	**4953**	**783**

Table 18: 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.

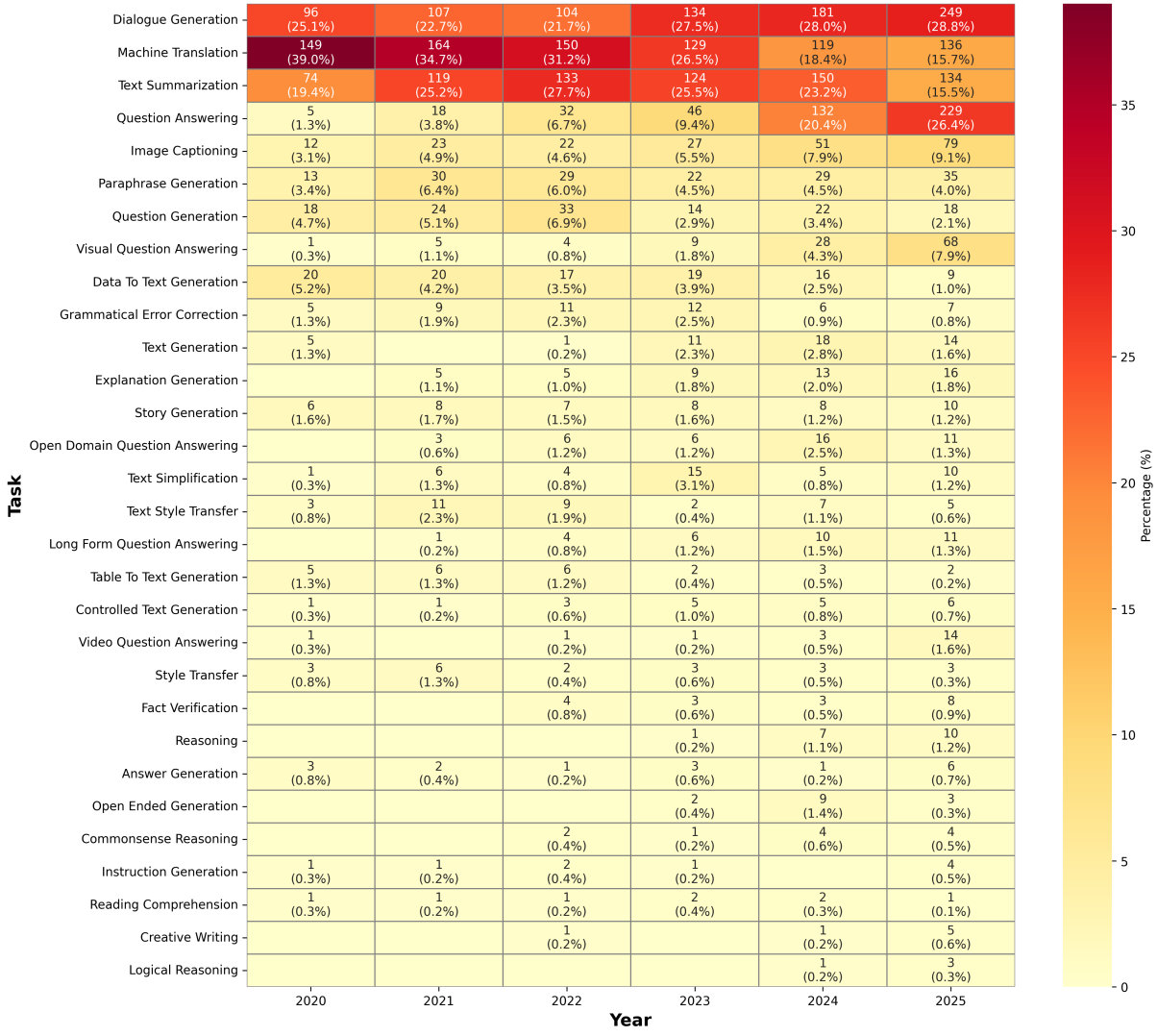

Figure 7: Heatmap of task-year distribution of the top-30 tasks. Both the total number and percentage per year are shown. The percentage of MT is decreasing while QA-related tasks increase significantly (QA, VQA).

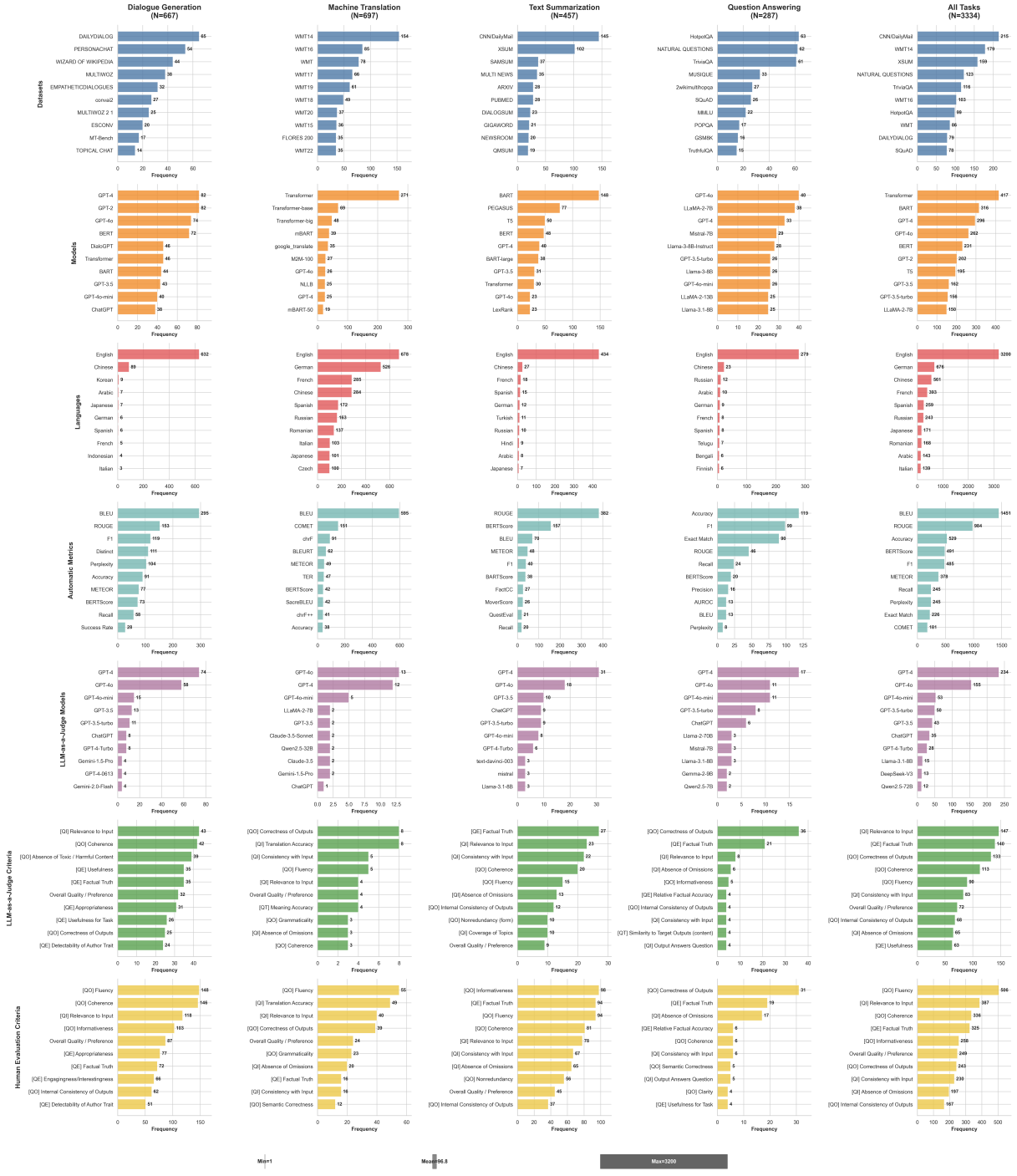

Figure 8: Distribution of metadata by tasks, columns include top-four and all-tasks, and rows are the metadata extracted (counts are after term normalization).

D Task-Specific Evaluation Guidelines

We distill our findings (§5) into concrete, criterion-level evaluation recommendations for each of the four major NLG tasks (Table 19). For each task–criterion pair, we list metrics whose strong task or criterion association in our data, supported by external validation studies, makes them well-suited for that purpose. We also list metrics to *avoid* as primary evidence for a given criterion—those that co-occur frequently but exhibit low LR with the target criterion, indicating indiscriminate application rather than deliberate, validated selection. These recommendations are grounded in the trends documented in Figures 3–6 and are intended as a starting point; readers should consult the cited references for task-specific validation evidence before adopting any metric.

Minimal Evaluation Recipes per Task. Building on Table 19, we distill a minimal three-to-four-step protocol for each task that researchers can adopt or adapt:

- **MT:** (1) Report COMET (Rei et al., 2020) (or the emerging MetricX series (Juraska et al., 2025)) as the primary adequacy metric; (2) retain BLEU only for comparability with prior work (Mathur et al., 2020), not as a standalone quality signal; (3) if LaaJ is deployed for MT-specific criteria (e.g., terminology, style), validate it against expert human MT-error annotations on the same test set.
- **DG:** (1) Report Distinct-n (Li et al., 2016) for diversity; (2) deploy LaaJ with a multi-criterion rubric covering engagement (appropriateness, naturalness) and safety (harmlessness, helpfulness); (3) reserve human evaluation for the safety dimension at minimum, since it is the criterion where LaaJ uniquely surfaces categories that traditional human evaluation rarely covers (§5).
- **TS** (1) Report ROUGE as a coverage baseline; (2) select at least one criterion-specific metric matched to your primary quality concern: SummaC (Laban et al., 2022) or UniEval (Zhong et al., 2022) for consistency, QAFactEval (Fabbri et al., 2022) or MiniCheck (Tang et al., 2024) for factuality, MoverScore (Zhao et al., 2019) or BARTScore (Yuan et al., 2021) for coherence — noting that adoption of these criterion-specific metrics remains a minority of TS papers in our corpus, so reporting them alongside ROUGE (rather than in place of it) is the realistic transition path; (3) if factuality is critical, pair automatic metrics with LaaJ factuality assessment on a human-validated subset.
- **QA:** (1) Report EM and F_1 only for comparability with prior work, not as primary signals for open-ended QA; (2) for open-ended or free-form answers, add a semantic evaluation layer — LaaJ with a correctness rubric is the most widely adopted option in our corpus, with emerging alternatives such as AlignScore (Zha et al., 2023); (3) if factual accuracy matters, supplement with an emerging fact-verification metric such as MiniCheck (Tang et al., 2024); (4) document the expected limitations of each metric (e.g., "EM ignores paraphrase; BERTScore captures semantic overlap but not factual correctness").

These recipes are intentionally minimal: they define a starting set rather than a comprehensive evaluation suite, and each step can be validated against the cited literature before deployment.

How to read this table. Table 19 should be read row-wise: for each task, we identify the key quality criteria that our data show are both (a) frequently evaluated and (b) poorly served by generic metrics. For **Machine Translation**, BLEU appears in 81.5% of MT papers overall and declines from 88.6% (2020) to 61.8% (2025), while COMET grows to 39.0% adoption in 2025; xCOMET grew from 0 papers in 2023 to 18 in 2025, and MetricX-25 from 0 to 12. We therefore recommend COMET and MetricX-25 as primary adequacy metrics and suggest retaining BLEU only for historical comparison (Mathur et al., 2020). For **Dialogue Generation**, LaaJ uniquely surfaces safety criteria (*Harmlessness*/*Safety*/*Non-Toxicity*) that traditional human evaluation rarely covers (§5); 47 papers in our corpus deploy LaaJ with explicit safety criteria, predominantly using GPT-4/GPT-4o as the judge. For **Text Summarization**, UniEval (23 papers) and SummaC (28 papers) are the most commonly adopted specialized factuality metrics in our corpus, with QAFactEval (12 papers) and MiniCheck (7 papers, concentrated in 2024–2025) as growing alternatives. For **Question Answering**, where Exact Match still anchors 27.1% of 2025 papers despite the free-form nature of modern LLM-generated answers, 137 of 462 QA papers in our top-30 corpus use LaaJ-based evaluation, and AlignScore appears

Task	Criterion	Recommended Metric(s)	Avoid as Primary	LaaJ viable?	Key References
Machine Translation	Adequacy / Translation Quality	COMET (Rei et al., 2020), MetricX-25 (Juraska et al., 2025) (COMET 39.0% of 2025 MT; BLEU still 61.8%)	BLEU (task-LR$=2.63$, low criterion-LR)	✓ (system-level; segment-level weaker (Lavie et al., 2025))	(Lavie et al., 2025; Mathur et al., 2020)
Dialogue Generation	Diversity	Distinct-n, Self-BLEU (Li et al., 2016)	BLEU (task-LR≈ 1 for DG)	✗ (auto metrics sufficient)	(Li et al., 2016)
	Engagement / Appropriate-ness	LaaJ + human (multi-item Likert)	ROUGE (n-gram overlap)	✓ (rubric)	(van der Lee et al., 2019; Bavaresco et al., 2025)
	Safety / Harmlessness	LaaJ rubric + human red-teaming (47 DG papers use LaaJ with safety criteria)	—	✓ (validate vs. humans; judge: GPT-4/GPT-4o)	(Bavaresco et al., 2025; Wang et al., 2024)
Text Summarization	Consistency / Faithfulness (vs. input)	SummaC (Laban et al., 2022) (28), UniEval (Zhong et al., 2022) (23)	ROUGE (low criterion-LR with consistency)	✓ (validation evidence thin, §5.3)	(Laban et al., 2022; Zhong et al., 2022)
	Factual Accuracy	QAFactEval (Fabbri et al., 2022) (12), MiniCheck (Tang et al., 2024) (7, mostly 2024–2025)	ROUGE / BERTScore (similarity $\neq$ factuality)	✓ (validation required, §5.3)	(Fabbri et al., 2022; Tang et al., 2024)
	Coherence / Redundancy	MoverScore (Zhao et al., 2019), BARTScore (Yuan et al., 2021)	BLEU (task-LR$=0.54$ in TS)	✗ (well-validated)	(Yuan et al., 2021; Zhao et al., 2019)
	Coverage / In-formativeness	LaaJ rubric (coverage scoring), ROUGE (as secondary)	—	✓ (rubric)	(Belz et al., 2025)
Question Answering	Semantic Correctness	LaaJ (pairwise / rubric), AlignScore (Zha et al., 2023)	EM / F_1 (ignore paraphrase; EM still 27.1% of 2025 QA)	✓ (primary for open-ended QA; 137/462 QA papers adopt LaaJ)	(Bavaresco et al., 2025; Zha et al., 2023)
	Factual Accuracy	MiniCheck (Tang et al., 2024), AlignScore (Zha et al., 2023)	BERTScore (semantic sim. $\neq$ factuality)	✓ (validation required, §5.3)	(Tang et al., 2024; Zha et al., 2023)

Table 19: Task-specific evaluation guidance derived from our empirical analysis (§5). "Avoid as Primary" lists metrics that co-occur frequently but exhibit *low* LR with the target criterion, indicating inertial use rather than validated selection. ✓ and ✗ indicate whether an LLM-as-a-judge evaluator is recommended as a viable complement. All recommendations should be validated on the specific task and language before deployment.

in 19 papers across the corpus (14 in TS, 6 in QA). We stress that LaaJ recommendations are conditional on human validation: our data show fewer than 8% of NLG papers validate LaaJ against human judgements, and where validation does exist, per-criterion alignment is highly variable (§5.3).

E Prompts for annotation

E.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
↪ ([]).

* For **answer_1.tasks** choose one or more from: {"Text Summarization","Dialogue
↪ Generation","Paraphrase Generation", "Machine Translation","Image Captioning","Code Generation"}. If
↪ 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.
```

E.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
```

E.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_annotator_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.
```

F Human Annotation Guideline

To demonstrate, we show our human annotation guideline of the 110 manually annotated extractions below. For the guidelines of other types of annotations, please refer to our project repository.

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" $\rightarrow$ 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