Illusions of the Gold Standard: A Large-scale Analysis of Human Evaluation Protocols for Long-form Text Generation

Anonymous ACL submission

Abstract

Human evaluation plays a critical role in assessing the quality of generated text. However, the reliability and reproducibility of these evaluations depend on transparent and well-documented protocols—details that are frequently missing in current practice. In this work, we conduct a large-scale analysis of human evaluation protocols for evaluating long-form generation tasks in *CL conference publications from 2023–2025, including a full manual review of 356 papers and LLM-assisted analysis for another 1.8k+ papers. We define a set of 20 reportable criteria related to reproducibility of human evaluation studies, and apply these criteria to systematically examine reporting norms and practices within the community. We find widespread under-reporting of important aspects of human evaluation study design, leading to ambiguity about what was measured and how, who contributed judgments, and how judgments should be interpreted. Based on these findings, we outline actionable recommendations to support more transparent and reproducible reporting in future research.

1 Introduction

With the growing adoption of LLMs, long-form or open-ended generation now dominate NLP research.[1] Automated metrics work poorly in these settings and human evaluation of model performance is still considered the gold standard, especially in high-expertise domains such as healthcare (Fraile Navarro et al., 2025), science (Idahl and Ahmadi, 2025), law (Fei et al., 2025), policy (Rivera et al., 2024), and misinformation detection and mitigation (Mishra et al., 2024; Cho et al., 2024). Yet, as prior work has shown for specific domains and/or tasks (Awasthi et al., 2025), current human evaluation procedures lack proper standardization and operationalization, which can limit the validity of evaluation, the robustness of conclusions, comparability across studies, and reproducibility of evaluation findings (Elangovan et al., 2024; Fleisig et al., 2024).

Concurrently, as the (self-)evaluative capabilities of models improve (Madaan et al., 2023), human evaluation is increasingly supplemented and/or supplanted by LLM-judges (Bavaresco et al., 2025; Thakur et al., 2025; Posner and Saran, 2025). In these cases, human evaluations play an additional and vital meta-evaluative role in assessing the performance of LLM-judges. As the landscape evolves, human judgments remain a foundational part of our evaluation methodology, and should be held to similar rigor and standards as other research methods employed by our community.

With this motivation, we turn the research lens upon the *CL research community to understand reporting practices around human evaluation and to identify and critique shortcomings of current practice. Drawing from prior work on reproducibility and good study design and reporting (Munafò et al., 2017; Fleisig et al., 2024), we define 20 reportable criteria for human evaluation protocols, covering aspects of task definition, annotation operationalization, annotator information, and data analysis and interpretation. We focus on papers studying medium- and long-form natural language generation, as they have the highest burden for human evaluation, and lack clear and reproducible automated evaluation metrics. Through manual and LLM-assisted analysis of this generation literature, we examine reporting patterns for human evaluations as they have evolved over the last few years (2023–2025) according to our criteria.

Our findings highlight clear deficiencies in documentation, with systematic under-reporting of key aspects of human evaluation study design. For example, only around half of papers we analyze include guidelines for human evaluation tasks or provide justification for the dimensions that were eval-

[1] In our analysis, around half of *CL papers from 2025 study long-form generation tasks per our definition (Table 1).

uated. And perhaps unsurprisingly, good practices such as the use of power analysis for computing sample size or reporting statistical confidence are exceedingly rare. We also find that the proportion of papers using human evaluation for long-form generation evaluation is declining in recent years, while the proportion of papers using LLM-judge meta-evaluation appears to be increasing. Based on these results, we offer recommendations for how to improve reporting.

In sum, we contribute the following:

- We define a set of 20 core reportable criteria for human evaluation studies along the dimensions of task documentation, annotation design, and analysis and interpretation. In §3, we describe the formation of these criteria, our codebook for assessing them, and additional reportable elements associated with good study design;
- Using our criteria, we conduct a large-scale analysis of 9.1k+ papers published at *CL conferences from 2023–2025. Of over 1,800 papers that study long-form generation and include human evaluation, we manually annotate 356 papers in full and conduct LLM-assisted annotation of the remainder. §4 describes our methods for corpus construction and data sampling, and implementation of our human annotation protocol;
- Our analysis (§5) reveals pervasive under-reporting of important aspects of human evaluation, entrenched but poorly justified norms around evaluation design, and recent changes in evaluation and meta-evaluation practices. We summarize our recommendations in §6.

2 Related Work

Reproducibility in ML/NLP Across scientific fields, poor study design and reporting have contributed to a broader reproducibility crisis. These concerns extend to machine learning and NLP, where methodological complexity and differences across studies can make it difficult to replicate results (Howcroft et al., 2020; Belz et al., 2023; Thomson et al., 2024). In response, the community has introduced formal reproducibility frameworks (e.g., conference checklists (Dodge et al., 2019), evaluation sheets (Shimorina and Belz, 2022; Belz and Thomson, 2025), resource tracks, and structured documentation such as model and data cards (Mitchell et al., 2019; Rogers et al., 2021)) to standardize author disclosures about data, experimental design, and evaluation pipelines (Elangovan et al., 2024). Yet despite adoption, these frameworks have not fully addressed underlying problems. Checklists are completed by authors without external validation, and their accuracy or completeness is rarely audited during peer review; important aspects of study design and evaluation can and do go unreported. Our study complements prior work by not only contributing a framework for assessing human evaluation study reporting, but also conducting a large-scale analysis of current reporting practices and their implications for reproducibility.

Human Evaluation for Generation Tasks Human evaluation has generally been considered the "gold standard" evaluation for natural language generation (Celikyilmaz et al., 2020). Human evaluation methods usually focus on intrinsic evaluation that assesses the quality of LLM-generated text, via data collection approaches like pairwise comparison (i.e., annotators indicate the response they prefer, perhaps along a specific dimension) or scoring scales. In recent years, with emergent LLM-as-judge capabilities brought on by more powerful LLMs, human evaluation is also increasingly employed in a meta-evaluative capacity (Madaan et al., 2023; Bavaresco et al., 2025; Thakur et al., 2025; Posner and Saran, 2025). Given the importance of human evaluation, and critiques (Howcroft et al., 2020) and anecdotes of under-reporting, we focus our analysis on recent *CL literature and aim to understand current practices for use of human evaluation in generation tasks.

3 Reporting Criteria & Codebook

We define reportable criteria for human evaluation grounded in the reproducibility literature. These criteria capture author-reportable aspects of human evaluation design, data collection, and analysis—the core operational stages of the scientific method (Munafò et al., 2017). We develop these criteria and the associated codebook by mapping each stage to concrete reporting decisions made by authors, and further revise our codebook using insights from extant investigations of human evaluation pitfalls in NLP research (Fleisig et al., 2024).

Codebook development Two lead authors and two senior authors developed the criteria and codebook iteratively via a combination of inductive and deductive processes (Fereday and Muir-Cochrane, 2006). Over four iterations, two authors independently applied initial versions of the codebook to human evaluation protocols from a sample of five papers per iteration. Following each round of cod-

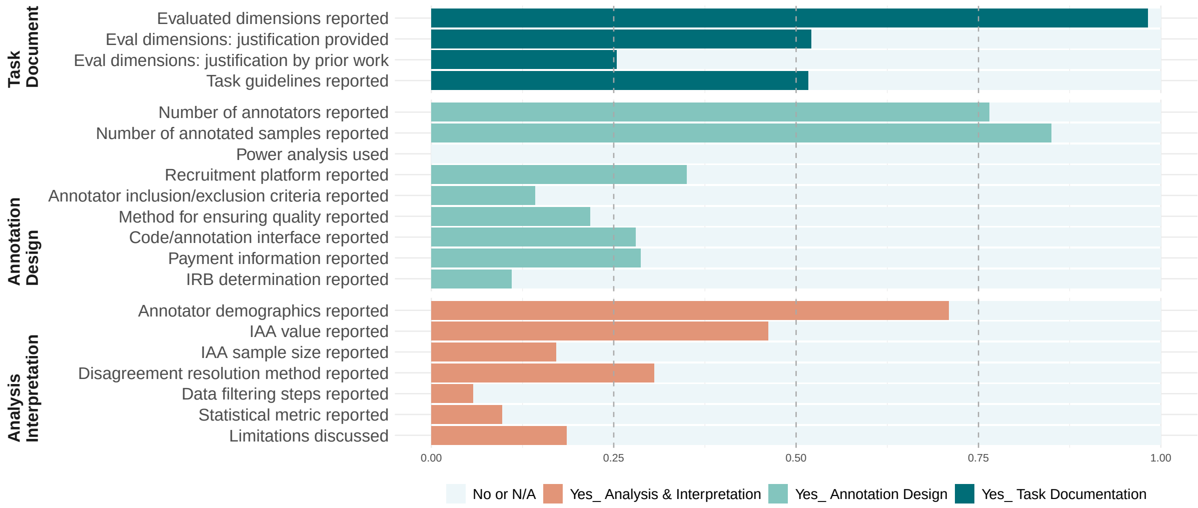


Figure 1: Average proportion of *CL papers reporting each of 20 core criteria related to the reproducibility of human evaluation protocols, estimated via bootstrapping. While most papers report evaluation dimensions, some annotator information, and annotation sample size, there is significant under-reporting of other aspects of evaluation study design. The bootstrapped standard deviations for all criteria fall in the range 0.01-0.03.

ing, the authors discussed and resolved disagreements and refined the codebook by either revising existing codes or introducing new inductive codes. At the end of this process, the study team reached consensus on the suitability of the final codebook.

The final codebook contains 37 questions, 20 of which form our core set of reportable criteria. The other 17 questions collect additional detailed information from each paper. The final set of codes and answer options is included in App. A.

Reportable criteria The 20 core criteria are binary, e.g., a paper either reports the number of annotators (Yes), or no information is provided (No or N/A). We group them into the following categories for presentation (see Figure 1 for full list):

- Task documentation (4 criteria): Aspects related to what is evaluated and how it is described to annotators, e.g., what dimensions (preference, accuracy, factuality etc) are being evaluated, justification for these dimensions, and guidelines for how annotators should judge each dimension.
- Annotation design (9 criteria): Operational details of the annotation procedure, such as the annotation interface, sample size, and processes ensuring annotation quality; as well as the annotators, e.g., recruitment platform, inclusion/exclusion criteria, payment, number of annotators.
- Analysis & interpretation (7 criteria): Aspects related to how collected annotations are analyzed, interpreted, and presented. Results from human evaluation can vary due to (i) variance in annotator judgment and (ii) variance in methods used to analyze the collected data. This category therefore includes elements such as annotator demographics, agreement (i.e., interrater reliability), whether additional data processing steps are used prior to reporting results (disagreement resolution or data filtering), reporting of statistical metrics, and whether limitations are discussed.

Not all 20 criteria apply in every setting, e.g., if disagreements are resolved by discussion, then reporting IAA may be unnecessary. Nevertheless, we expect most studies involving human evaluation to report roughly 15-16 criteria from this list.

Other information Beyond the core reportable criteria, we also collect 17 additional pieces of information from each paper for further analysis. First, we collect information that helps obtain a more nuanced understanding of evaluation design, such as where human evaluation details are reported (in the main paper or appendix), and whether LLMs and humans are used to evaluate the same dimensions (to understand the increasing use of LLM-judges in evaluation). Second, we collect details about annotators to better characterize annotator populations, including whether they are students, experts, or authors, and the platforms from which they are recruited. Last, we track details of IAA metrics such as the specific metrics computed and any methods for resolving disagreements (e.g., majority vote or consensus processes). Refer to App. A for complete codebook details.

4 Dataset & Methods

Using our criteria and codebook, we conduct a large-scale manual and LLM-assisted analysis of *CL conference publications from 2023–2025. We focus on these venues because they constitute a coherent and influential publication ecosystem for computational linguistics and NLP while spanning a diverse range of research, offering a representative snapshot of prevailing evaluation practices. We analyze papers from the last three years to capture current practices during a period of rapid change: (i) the growth of LLMs enables new generation tasks; (ii) unlike tasks that can be assessed with automated metrics, medium- and long-form generation tasks lack reference answers and have higher evaluative burden; and (iii) the prevailing use of human evaluation for direct assessment and LLM-judge meta-evaluation raises questions about how human evaluation protocols should be designed.

Papers are included in our analysis if they meet our inclusion criteria: (i) studies a long-form generation task and (ii) employs human evaluation. We define *long-form generation* as free-form natural language generation, excluding tasks such as machine translation and code generation. We define *human evaluation* as the use of human annotators to examine and assess model outputs. A subset of 356 papers is manually annotated in full using our codebook from §3, while we conduct partial analysis on the remaining papers through LLM-assisted labeling. Below, we describe our procedures for corpus curation and sampling (§4.1), manual annotation (§4.2), and LLM-assisted labeling (§4.3).

4.1 Corpus Curation

We begin with 9,172 papers from major *CL conferences: ACL, EMNLP, and regional chapters NAACL, EACL, and AACL, published 2023–2025. We download conference proceeding paper PDFs from the ACL Anthology[2] and use GROBID[3] following Rohatgi (2022) to parse and extract clean textual content for downstream filtering and search.

Step 1: Keyword filters We first narrow the corpus to papers studying long-form text generation. We expand a set of seed keywords (summarize/-summarise, dialogue, long-form, etc.) with GPT-4 to cover related task variants (e.g., multi-turn dialogue, document-level generation). We apply case-insensitive matching with stemming across titles, abstracts, and the main text of each paper, retaining papers that match at least one expanded keyword. This produces a candidate set of 8,408 papers. We provide the full keyword filter set in App. B.

Step 2: LLM filters We apply a second-stage filter via majority vote among three LLMs: Gemini-2.5-Pro, Claude-3.7-Sonnet-20250219, and GPT-4o-mini-2025-04-16. Each model answers two binary screening questions: (i) whether the task in the paper is considered long-form natural language generation and (ii) whether the paper involves human evaluation. A paper is retained if at least two models answer "Yes" for both criteria. Following this step, 3,620 papers meet our inclusion criteria for long-form generation, and 1,891 papers further meet our criteria of using human evaluation. Prompts for filtering are in App. C.

To estimate the false negative rate of these LLM filters, we manually inspect a random sample of 50 papers that the LLM majority vote finds to be about long-form generation but rejects for not including human evaluation. Among these 50 papers, 3 are found to include human evaluation, corresponding to a FNR of around 6%.

Step 3: Sampling for manual annotation We sample 356 papers from the set of 1,891 for manual annotation. We sample only from conferences in 2024 and 2025 to focus our efforts on capturing recent practices. As such, we stratify our sample to include approx. 20% of papers from all 4 conferences in 2024 and NAACL and ACL in 2025.[4,5]

4.2 Manual Annotation Procedure

Two lead authors and three contributors coded the 356 sampled papers. The lead authors are experienced NLP and HCI researchers familiar with reading research papers, and the three contributors—one undergraduate and two masters students—have prior experience reading and writing NLP papers.

Annotation process To operationalize our annotation task, we provided each annotator with a codebook reference sheet with definitions for each code and how to select different answer options (App. D.1). The annotation task was conducted

[2] https://aclanthology.org/

[3] https://grobid.readthedocs.io/en/latest/

[4] EMNLP 2025 occurred after we completed manual annotations, though is included in automated analysis.

[5] Our final manual annotation set includes proportionally more papers from some conferences. We had planned to annotate more papers (before recalibrating due to the difficulty of the task) and our initial annotation assignments did not randomize conference order. We corrected for this midway through our annotation timeline to achieve suitable coverage over all included conferences.

Year	Conference	Counts (proportion)				
		Total	Step 1 (Keywords): Longform	Step 2A (LLM Filter): Longform	Step 2B (LLM Filter): Longform & Human Eval	Step 3: Sample for Manual Annotation
2023	EACL	281	225 (0.80)	65 (0.23)	39 (0.14)	-
	ACL	912	865 (0.95)	289 (0.32)	192 (0.21)	-
	AACL	73	64 (0.88)	22 (0.30)	19 (0.26)	-
	EMNLP	1048	919 (0.88)	355 (0.34)	196 (0.19)	-
2024	NAACL	489	452 (0.92)	153 (0.31)	105 (0.21)	20
	ACL	869	793 (0.91)	275 (0.32)	177 (0.20)	135
	EMNLP	1270	1156 (0.91)	381 (0.30)	256 (0.20)	81
	EACL	182	161 (0.89)	62 (0.34)	35 (0.19)	20
2025	NAACL	637	589 (0.92)	300 (0.47)	145 (0.23)	30
	ACL	1602	1517 (0.95)	806 (0.50)	351 (0.22)	70
	EMNLP	1809	1667 (0.92)	912 (0.50)	376 (0.21)	-
Total		9172	8408 (0.92)	3620 (0.39)	1891 (0.21)	356

Table 1: Progressive filtering of *CL papers for inclusion in analysis. Keyword filters (step 1) are followed by LLM filters for papers about long-form generation tasks (step 2A) **and** which contain human evaluation (step 2B). We then stratify sample over 6 conferences from 2024–2025 for manual annotation (step 3). Proportions represent per-row normalization. Conferences are ordered chronologically according to their official event dates.

using Google Sheets (see App. D.2 for annotation interface). Each batch of papers was assigned to an annotator in a new tab in their own spreadsheet. Within the interface, we grouped the 37 total questions answered for each paper together into subcategories likely to be documented in the same paper sections. We restricted all binary and multiple-choice questions to a predefined set of answer options (pre-configured in the spreadsheet); where appropriate, annotators may also select “Other” and provide a free-text explanation. In cases where multiple human evaluation protocols are described in the same paper, annotators were instructed to focus on the first protocol described in the paper.

Onboarding To ensure all contributors had a comprehensive understanding of the codebook and task, we began with a three-week training period consisting of: (i) an introductory group meeting to explain the codebook; (ii) an initial batch of 10 papers annotated independently by all contributors; (iii) comparing annotation results to the leads and one-on-one discussions to provide feedback and clarify disagreements (during these discussions, we also refined definitions for any ambiguous codes; and (iv) a second independent round of annotation of 5 papers to reassess performance.

All collaborators reached 73% agreement after iterations of feedback and were assigned their own non-overlapping batches of papers each week for annotation. Each annotator annotated between 50 and 138 papers over the study period. Furthermore, one of the lead authors conducted a random quality check on 105 (60%) papers of other annotators.

Inter-annotator agreement (IAA) We compute IAA for all annotators using 5 papers from the final onboarding annotation set. There are 155 questions per person (115 binary, 30 single-label multiple choice (MC), and 10 multi-label MC; open-ended questions not included in IAA computation). We report average pairwise agreement between each of the three contributing annotators and the consensus annotation provided by the two lead authors.

Binary questions (n=23) yield the most reliable judgments (percent agreement=81%; Cohen’s κ=0.54), while single-label MC questions (n=6) achieve fair agreement (percent agreement=59%; Cohen’s κ=0.24). Multi-label MC questions (n=2) also achieve fair agreement (percent agreement=30%; Cohen’s κ=0.3).

Data analysis & interpretation We construct additional binary dummy variables from questions that have numeric, MC, or qualitative answers. For example, from the numeric IAA value, we create a new binary variable `IAA_reported` that is coded “Yes” if a specific number is reported in the paper, and “No or N/A” otherwise. To estimate the true proportion of papers that report each criterion, we perform bootstrap resampling with replacement (n=500) to derive averages and standard errors.

4.3 LLM-assisted Annotation

To scale analysis to our whole corpus, we adopt LLMs to label papers we could not manually annotate. For each paper, we construct its input context from its abstract, introduction, and candidate human evaluation sections from its main paper and appendix identified using keywords (App. E.1).

We then prompt GPT-4o-mini-2025-04-16 with codebook questions to extract relevant information about the first human evaluation pipeline reported in each paper. The model is prompted with a chunk of questions from our codebook at a time. We conduct automatic type checking and apply numeric-or-NA constraints; if chunk-level or whole-paper validation fails, we re-run the full set of extraction prompts up to two more times.

For prompt refinement, we use a training set of 26 papers sampled from those manually annotated by the two lead authors. We test final prompt performance on an independent validation set of 125 papers evenly sampled from all five independent annotators (25 papers each). Model selection and prompting details can be found in App. E.

In §5, we only report LLM annotation results for questions where the LLM achieves a validation accuracy over 0.75. Validation accuracies for all binary and multiple-choice questions and final prompts are reported in App. E.3.

5 Results

Summary statistics Among 356 manually-annotated papers, 284 are confirmed by annotators to meet both inclusion criteria: (i) studying a long-form generation task and (ii) including a human evaluation pipeline; i.e., 72 papers did not meet either one or both of these criteria, and can be considered false positives from the LLM filters.

Reported tasks and dimensions For the evaluated dimensions reported in each paper, we apply *PorterStemmer* from the *nltk* library to normalize all dimension phrases. We also group each paper's generation task (as reported by authors) into categories for analysis using the manually-curated word-stem mappings in App. F. The most frequent task categories are *Dialogue and interactive systems* (n=40), *Summarization* (n=26), *Safety and jailbreaking* (n=26), *QA* (n=18), and *Story generation* (n=13). Evaluation dimensions show some consistency within task categories but are highly variable across studies; e.g., relevance is assessed across many task categories; coherence appears very often for story generation; correctness is most frequent in QA tasks. Figures and details in App. F.

Form of human annotation tasks and reporting Human judgments are most commonly collected via *binary* judgments (26%) and *pairwise comparison* (22%), followed by *likert scale* (22%), *numeric scale* (21%), *categorization* (13%), and *rank-based* (3%) tasks.[6] Of annotated papers, 47% include human evaluation details in both the main paper and appendix, while 33% report this information only in the main paper and 20% only in the appendix.

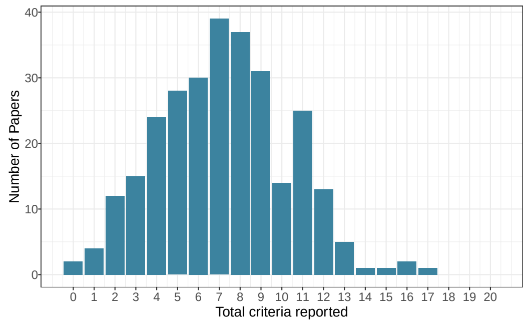


Figure 2: Distribution of total criteria reported; over half of papers report $\leq$7 of 20 reportable criteria.

5.1 Key observations

We report key findings below. Additional analysis can be found in App. I.

***CL papers pervasively under-report important aspects of human evaluation protocols.** As in Figure 1, while most papers report the evaluated dimensions (98%), number of annotators (77%), and number of annotated samples (85%), all other criteria are reported far less frequently. Key information such as justification for the chosen evaluation dimensions and how the evaluation is described to annotators (task guidelines) is only present in around 50% of papers. Important aspects of annotation design such as payment information (29%) and IRB determination (11%) are rarely reported. While a reasonable proportion of papers report IAA among annotators (46%), only 17% report the number of samples used to compute agreement metrics. Very few papers report statistical metrics (9%) and none use power analysis to derive sample sizes. Given the importance of empirical results in NLP research, the pervasive lack of statistical reporting is particularly troubling. Complete sample and bootstrap estimates can be found in App. G.

Most papers report fewer than 8 criteria. We find that the modal number of reported criteria is 7 out of a potential 20. More than half of papers we review report 7 or fewer reportable criteria (Median=7, SD=3)—note this is only whether an item is reported without any judgment on the content or sufficiency of what is reported. Very few papers

[6] These do not sum to 100% since each paper may include more than one form of annotation.

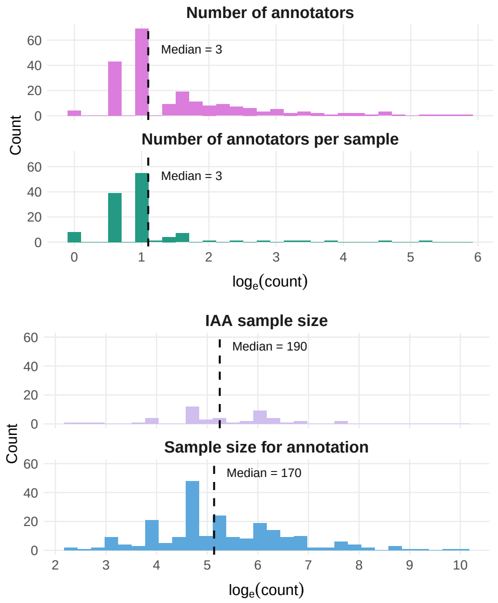


Figure 3: Distributions of annotator and sample counts

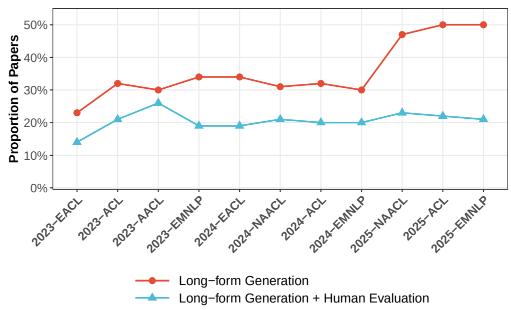


Figure 4: Temporal trends for 2023–2025 *CL conferences. While papers studying long-form generation have increased in the last year, the proportional use of human evaluation for these tasks has decreased.

(n=5) report more than 13 criteria, and no papers report all 20 reportable criteria.

Norms around sample size and annotator count are strong but not well justified. Annotation sample size and annotator count affect the inferences that can be drawn from a model evaluation study, yet this information is not consistently reported. We find that 15% of papers do not report sample size, and 23% do not report the number of annotators involved in human evaluation. Among papers that do report this information, we find no consistent practices for determining or justifying sample size (e.g., no papers use power analysis to determine sample size). Reported sample sizes vary widely, ranging from as few as 10 to a maximum of 23,040, with a median sample size of 170.

Among papers that report the number of annotators (Median=3, MAD=1.48[7]), three annotators is the most common configuration (32%), followed by two annotators (20%). To gauge task difficulty and variability, it is critical to report IAA as a measure of annotation consistency. However, our bootstrapped analysis estimates only 46% of papers report any measure of IAA; among papers with more than one annotator, only about half (51%) report IAA. Among papers that reported IAA, the median number of samples used to compute agreement is 190. Figure 3 shows log-scale distributions.

[7]MAD: median absolute deviation is a variability measure similar to standard deviation but less sensitive to outliers.

Annotator information is often missing or incomplete. Despite increasing evidence of the influence of annotator background on annotation outcomes (Sap et al., 2019; Ding et al., 2022; Al Kuwatly et al., 2020), *CL papers lack consistent reporting of annotator demographic information. We find that 29% of papers do not report any demographic information about annotators, and 65% do not report any information about recruitment platforms. Among papers that report some demographics, we find that 31% of these papers recruit students, 50% recruit domain experts (as described by authors), and 13% of papers recruit paper authors as annotators. Among other characteristics, education is most frequently provided (48%), followed by language (27%), gender (12%), and country of residence (9%).

Annotation quality control is rarely employed. We track whether researchers adopt any data filtering steps (i.e., attention checks, manipulation checks)—techniques to remove low-quality crowdsourced data or any other procedure to ensure annotation quality. We find that only 6% of papers include data filtering steps, and 22% of papers include procedures to ensure annotation quality. These procedures usually focus on ensuring annotation consistency, e.g., having a training or pre-assessment period for annotators, pilot studies to ensure annotation clarity, or methods to improve the reliability of annotation (such as only keeping samples with full agreement among annotators, or having multiple stages of quality checks). Around 30% of papers report steps for resolving disagreements among annotations; majority vote (38%) is the most common, followed by *averaging* (24%) and *having a consensus process* (17%). Only a small proportion (19%) of papers discuss the limitations of their human evaluation pipeline.

5.2 Temporal trends

The proportion of long-form generation papers with human evaluation is declining. While the number of studies of long-form generation tasks is increasing, the proportion using human evaluation is declining. As Figure 4 and Table 1 show, while the proportion of papers studying long-form generation tasks has increased from around 30% to 50% in 2025, the proportion of overall papers that include human evaluation has been stable around 20%. We also observe that half of papers in our manually annotated set adopt LLM-judges for evaluation, and 20% of these papers use LLM-judges for evaluation without human evaluation.

Use of human evaluation for meta-evaluation of LLM-judges is on the rise. We observe the increasing use of human evaluation to assess the performance of LLM-judges (more in App. H). Compared with EMNLP'23, the proportion of papers using LLM-judges tripled for EMNLP'25 from 4% to 12%. Meta-evaluation of LLM-judges also necessitates a reliable human evaluation pipeline. However, we did not find improvements in reporting among studies that use human evaluation for meta-evaluation in our manually annotated data.

6 Discussion & Recommendations

Our analysis reveals clear gaps between the central role human evaluations play in NLP research, and the current rigor (or lack thereof) in reporting practices. Across *CL papers studying long-form generation from the last three years, we observe under-reporting of important criteria, high variability in human evaluation study design, and a recent and rapid shift in the way human judgments are used, especially in a meta-evaluative capacity for LLM judges. We discuss implications of these findings and outline recommendations for the community and for future work.

R1: Report core reportable criteria for reproducibility. While reporting details such as recruitment information, IAA, and task guidelines can take up important space, we argue that it is possible to report such crucial details succinctly. For example, for the study in this paper:

> We analyze reporting practices for human evaluation in *CL papers using a codebook of 37 question, including 20 core reportable criteria associated with reproducible science. The codebook was iteratively developed based on the reproducible science framework, several rounds of pilot testing, iterative feedback from the research team, and prior work (cite). Using the codebook, we manually annotate 356 papers studying long-form generation with human evaluation drawn from the *CL corpus 2023–2025. Five annotators (2 PhD, 2 Masters, and 1 undergraduate student, all with experience reading and writing NLP papers), who are also authors of this paper, underwent a multi-week calibration process. All annotators met and exceeded an IAA threshold on a held-out set of 5 papers (155 questions per annotator), achieving 73% agreement on binary questions (κ=0.54) and moderate agreement (κ=0.25-0.3) on multiple choice questions. For analysis, we report descriptive statistics with bootstrapped confidence intervals for reporting frequencies. Task guidelines and screenshots of our annotation interface are provided in the Appendix.

We recommend this (based on our 20 criteria) as a minimum template for human evaluation reporting.

R2: Avoid bespoke evaluation design when possible Papers evaluating the same task type (e.g., summarization, QA), can differ widely in evaluated dimensions (App. Figure 10), how those dimensions are operationalized, annotation formats, and analysis methods. While some variation is expected, justification is rarely provided (25% of papers justify dimensions using prior work). This degree of heterogeneity makes it difficult to compare results across studies or determine whether studies are measuring similar underlying constructs. We recommend that researchers deliberately adopt dimensions, scales, and evaluation protocols from prior work while accounting for differences in study objectives. If new evaluation facets are introduced, rationales and full operationalization details should be clearly described.

R3: Hold human evaluation to a higher standard Temporal analysis highlights a shift in human evaluation practices, especially their increasing use for meta-evaluation of LLM judges. If human evaluation is poorly documented or inconsistent across studies, it cannot serve as a reliable gold standard for assessing LLM judges. Weak human evaluation protocols can introduce error into downstream systems, as shaky foundations lead to structural failure. Rather than reducing the importance of human evaluation, its growing meta-evaluative role demands increased rigor, transparency, and higher standards of documentation.

Conclusion

Human evaluation remains a cornerstone of NLP research, especially for long-form and open-ended generation tasks. The community could substantively improve its reporting practices with only modest changes in authoring norms as we have suggested above. We hope that the criteria list and recommendations presented here can serve as a practical reference point for the future evolution of human evaluation and documentation practices.

Limitations

Our meta-analysis is limited to papers in the past three years (2023–2025) and *CL conferences. This leaves open questions about the reporting practices of NLP research papers in other conferences and journals, or adjacent research communities. We encourage future work to build on our motivation of reproducible science and broaden the examination of evaluation and reporting practices in our research communities.

In addition, we acknowledge that what is considered "reportable" can vary significantly depending on what task is being performed by models and assessed, and the role of the evaluation itself. Our work does not aim to critique any individual study for its design choices, but is geared towards understanding norms and patterns in the community as a whole and offering recommendations for how to improve documentation practices where there are clear gaps. It may be useful in future work to consider needs specific to certain NLP tasks or user groups, perhaps through more granular or adaptive criteria lists.

Ethical considerations

We identify no immediate ethical concerns with our research study or conclusions. This study analyzes publicly available academic papers and does not involve collecting new data from human participants. All human annotation is conducted by the authors and trained collaborators on published materials, without collecting personal or sensitive information. As such, this work does not require institutional review board (IRB) review.

We acknowledge that human judgments may reflect annotator perspectives and subjective biases. To mitigate this, we employ a structured codebook, annotator training, and compute inter-annotator agreement. Our use of LLMs is limited to supporting large-scale analysis, and we recognize broader ethical concerns surrounding LLM-based evaluation, including bias propagation and over-reliance on automated judgment. Accordingly, we report human-annotated results as our primary findings, and use LLM annotations for supplementary evidence, applying a validation accuracy threshold of 0.75 to ensure reliability, as described previously.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Raghav Awasthi, Atharva Bhattad, Sai Prasad Ramachandran, Shreya Mishra, Ashish K Khanna, Jacek B Cywinski, Kamal Maheshwari, Dwarikanath Mahapatra, Izabella DiRosa, Anabelle Cohen, et al. 2025. Human evaluation of large language models in healthcare: gaps, challenges, and the need for standardization. *npj Health Systems*, 2(1):40.

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.

Anja Belz and Craig Thomson. 2025. Heds 3.0: The human evaluation data sheet version 3.0. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 60–81.

Anja Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, et al. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in nlp. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrara, et al. 2024. Can language model moderators improve the health of online discourse? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7478–7496.

Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of annotator demographics on sentiment dataset labeling. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. 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 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1160.

Zhiwei Fei, Songyang Zhang, Xiaoyu Shen, Dawei Zhu, Xiao Wang, Jidong Ge, and Vincent Ng. 2025. Internlm-law: An open-sourced chinese legal large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9376–9392.

Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5(1):80–92.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

David Fraile Navarro, Enrico Coiera, Thomas W Hambly, Zoe Triplett, Nahyan Asif, Anindya Susanto, Anamika Chowdhury, Amaya Azcoaga Lorenzo, Mark Dras, and Shlomo Berkovsky. 2025. Expert evaluation of large language models for clinical dialogue summarization. *Scientific reports*, 15(1):1195.

David M Howcroft, Anja Belz, Miruna 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.

Maximilian Idahl and Zahra Ahmadi. 2025. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. 2017. A manifesto for reproducible science. *Nature human behaviour*, 1(1):0021.

Eric A Posner and Shivam Saran. 2025. Judge ai: Assessing large language models in judicial decision-making. *University of Chicago Coase-Sandor Institute for Law & Economics Research Paper*, (2503).

Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. 2024. Escalation risks from language models in military and diplomatic decision-making. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 836–898.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shaurya Rohatgi. 2022. Acl anthology corpus with full text. Github.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Anastasia Shimorina and Anja Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in nlp. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 404–430.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common flaws in running human evaluation experiments in nlp. *Computational Linguistics*, 50(2):795–805.

A Final codebook

We include the complete criteria list and final codebook in Table 2. Specifically, we group questions into three categories: (i) task documentation, (ii) annotation design, and (iii) analysis & interpretation. Starred items are included in our core criteria list of 20 reportable items. We include the exact questions that annotators answer when annotating each paper.

B Corpus construction: keyword filters

In Table 3, we provide the complete set of keywords used to identify papers studying long-form text generation. Keywords are matched in a case-insensitive manner with stemming against titles, abstracts, and main text extracted using GROBID. Papers matching at least one keyword are retained for subsequent LLM-based filtering.

C Corpus construction: LLM filters

The full prompt text used for LLM-based second-stage corpus filtering is reproduced in Figure 5.

D Details for manual annotation

D.1 Annotation codebook reference

Instructions for annotation task and item answers are reproduced in Table 5. For criteria that are difficult to assess, we clarify each answer option to maximize annotation consistency.

D.2 Annotation interface

In Figure 6, we include a partial screenshot of the annotation interface we develop in Google Sheets. Answer options are restricted to valid types.

E Details for LLM-assisted annotation

E.1 Keywords for section selection

Table 4 provides the complete list of keywords and phrases used to identify human evaluation sections in each paper. Keywords are matched in a case-insensitive manner with stemming and are used to select candidate sections, which are then passed to the LLM-based annotation prompt (see Figure 7).

E.2 LLM selection & validation

We conduct a pilot study to select an appropriate large language model for automatic annotation of human evaluation details. We compared three state-of-the-art models: Gemini-2.5-Pro, Claude-3.7-Sonnet-20250219, and GPT-4o-mini-2025-04-16 with identical prompts and input contexts. Performance is validated on the manual annotations of a held-out set consisting of 26 papers. For each model, we assess annotation quality using question-level validation accuracy, measuring consistency with human-annotated ground truth across the full set of codebook questions.

Among the models we test, GPT-4o-mini-2025-04-16 achieves the highest overall accuracy. Based on this empirical comparison, we select GPT-4o-mini-2025-04-16 as the annotation model for the remainder of our corpus.

E.3 Prompts for LLM-assisted annotation

To control context length and improve reliability, we split codebook questions into five semantically coherent chunks for prompting. Questions from each chunk are answered in separate API calls, with the model instructed to return a flat JSON object with answers.

Prompts used for LLM-assisted annotation are reproduced in three figures: (i) the task introduction in Figure 7, (ii) the chunked question structure in Figure 8, and (iii) the full list of annotation questions in Figure 9. We validate LLM-based annotation using GPT-4o on a held-out set of 125 manually-annotated papers (25 from each of the five annotators), or 3,875 annotations in total (31 binary or multiple choice questions for each of the 125 papers). Table 6 reports the percentage agreement between GPT-4o and human annotations.

F Task-level analysis

Task-category stem-keyword mapping Table 7 shows the stem keyword–to–task-category mapping used to assign each paper to a primary NLP task for task-level analysis.

Author-reported tasks and evaluation dimensions Figure 10 presents the distribution of the 15 most frequently evaluated stemmed dimensions across six major NLP task groups. Overall, relevance, coherence, fluency, and correctness dominate the evaluation dimensions across tasks. These dimensions assess whether the generated content

Table 2: Full list of questions annotated for each paper. ★ indicates those corresponding to core reportable criteria.

Category & Element Name	Question for Annotation
Category: Task Documentation (5)	
★ Evaluated dimensions	What dimensions are annotators asked to evaluate regarding the models' output?
★ Eval dimensions: justification provided	Is there justification provided for the selected dimensions or the human evaluation pipeline?
★ Eval dimensions: justification by prior work	If yes to the previous question, is prior work cited?
★ Task guidelines reported	Are task introductions/guidelines included?
Code/annotation interface reported	Is code for or image of the annotation interface shared?
Category: Annotation Design (18)	
Main Task	What is the main task the paper focuses on (e.g. summarization, dialogue)?
	↪ If other for the last entry, provide a detailed description of the form of the task.
Domain	What is the domain that the main tasks related to (e.g.medicine, programming)?
	↪ If other is the response for the previous entry, describe the domain.
Longform generation	Free-form Generation Evaluation?
Human evaluation	Human evaluation?
More than one evaluation task	Is there more than one human evaluation pipeline?
Form of annotation task	What is the form of annotation task for human evaluation?
	↪ If other for the last entry, provide a detailed description of the form of the task.
Claims task is novel	Do the authors claim the long-form generation task is newly introduced (novel)?
Sections w/ human eval details	Which section(s) include details about the human evaluation?
	↪ What is the location of the main design details of human evaluation (i.e.necessary information for reproducing the evaluation)?
Only human eval used	Is human evaluation the only evaluation method being used for assessing model performance?
LLMs used for eval	If no to the previous question, are LLMs being used to evaluate model outputs?
LLMs and humans eval same dimensions	If yes to the previous question, are human and LLMs evaluating the same dimensions of model outputs?
★ Sample size for annotation	Total sample being annotated
★ Power analysis used	Is sample size determined by power analysis?
★ Recruitment platform	Recruitment platform (NA if not reported)
★ Annotator inclusion/exclusion criteria reported	If recruitment platform is not NA, any restrictions on participation for annotation (Yes/No)
★ Payment information reported	Is payment to annotator reported?
★ IRB determination	Is IRB/ethics review used?
★ Method for ensuring quality reported	Is there any procedure used to ensure human annotation quality (e.g., training period of annotators)?
	↪ If yes, please copy paste the exact text from the paper
Analysis & Interpretation (14)	
★ Annotator demographics	What demographic information of participants (if any) is reported?
Annotators are students	Are the human annotators students?
Annotators are authors	Are authors also annotators?
Annotators are experts	Are human annotators referred to as experts or have domain expertise?
	↪ What is the description of the annotators' expertise (copy and paste content from paper)?
★ Number of annotators	Number of annotators
★ Number of annotators per sample	Number of annotators for each annotated item
★ IAA value reported	Is interrater agreement reported?
★ IAA sample size	Number of samples used to compute IAA
IAA metrics	What metrics are reported for interrater agreement?
	↪ If other is selected for the previous question, write down the metric here.
★ Data filtering steps reported	Are any filtering steps applied after human annotations are collected? (e.g., outlier removal, attention checks, manipulation checks)
Strength of agreement	How strong is the agreement (report agreement quality based on kappa interpretation)?
	↪ Comments on agreement description
★ Disagreement resolution method	How is disagreement being treated?
★ Statistical metric reported	Are any of the following metrics reported for the human evaluation data: standard error/deviation, confidence interval?
★ Limitations discussed	Are there any limitations noted in regards to the human evaluation pipeline?
	↪ Is yes to the previous column, record what authors mentioned regarding the limitation

is semantically and factually appropriate and related to the task (e.g., relevance, correctness), and also assess the surface-level linguistic quality (e.g., fluency, coherence).

Different task groups exhibit different evaluation priorities. In *Dialogue and Interactive Systems*, relevance, coherence, and fluency remain the primary dimensions, accompanied by closer attention to accuracy, while in *Safety and Jailbreak* tasks prioritize relevance, quality, and safety-related dimensions such as correctness and consistency. For *Summarization*, informativeness and faithfulness receive higher emphasis, indicating the importance of content coverage and information consistency. In *Question Answering*, correctness and relevance dominate. *Story generation* places very strong emphasis on coherence. These variations on the evaluated dimensions highlight how evaluation criteria are systematically adapted to the functional goals of different free-form generation tasks.

G Bootstrapped Estimates of Proportion

We included bootstrapped estimates for reporting frequency for each of our core criteria in Table 8 along with raw proportions from our manually-annotated sample.

H Temporal trends

We provide analysis of temporal trends in reporting based on LLM annotations across *CL papers (2023-2025) with human evaluation and long-form

Table 3: Open-ended natural language generation keyword set used in **Step 1: Keyword filters**.

Task	Keywords Used for Filtering
General Long-form Keywords	long form, long-form, Summarisation/ Summarization
Summarisation / Summarization	Extractive Summarisation/ Summarization, Abstractive Summarisation/ Summarization, Multimodal Summarisation/ Summarization, Multilingual Summarisation/ Summarization, Conversational Summarisation/ Summarization, Query(-)focused Summarisation/ Summarization, Multi-document Summarisation/ Summarization, Multidocument Summarisation/ Summarization, Long(-)form Summarisation/ Summarization, Few(-)shot Summarisation/ Summarization, Document Summarisation/ Summarization, Text Summarisation/ Summarization, Opinion Summarisation/ Summarization, Review Summarisation/ Summarization, Legal Document Summarization, Scientific Paper Summarisation/ Summarization, News Summarisation/ Summarization, Explanatory Summarisation/ Summarization
Narrative & Story Generation	Narrative Generation, Story Generation
Question Answering	Long-Form Question Answering, Long Form Question Answering, Open-Domain Question Answering, Open Domain Question Answering, Explanatory Question Answering, Document-based Question Answering, Document Question Answering, Long-Form QA, Long Form QA, Open-Domain QA, Open Domain Question Answering, Explanatory QA, Document-based QA, Document QA
Conversational Systems	Reading Comprehension, Dialogue, Dialog, Conversation, Conversational AI, Dialogue Management, Conversational Agent, Chatbot, Conversational Interface, Dialogue System, Chat-oriented Dialogue System, Chat oriented Dialogue System, Open-domain Conversational System, Open domain Conversational System, Closed-domain Conversational System, Closed domain Conversational System
Report & Writing Generation	Report Generation, Essay Generation, Script Writing, Book Writing, Content Creation, Extended Abstract Generation, Technical Documentation Generation, Healthcare Documentation, Collaborative writing, open-ended generation
Editing & Research	deep research, text simplification, paraphrasing, document editing

Table 4: Keyword list used to identify and extract human evaluation sections from papers.

Category	Human Evaluation Section Selection Keywords
Human Evaluation Indicators	human evaluation, manual evaluation, expert evaluation, human judg, human assess, expert assess, human preference, expert preference, user study, human study, participant, annotator, rater, subject, evaluator, human subject, human judgment,interface, screenshot
Evaluation Setup & Protocol	Likert, pairwise, A/B, MOS, rating, assessment, preference, inter-annotator, Cohen's kappa, Krippendorff
Recruitment Platforms & Payment	AMT, Mechanical Turk, mturk, Prolific, crowdsourcing, paid, volunteer, Upwork,IRB, consent, compensation

generation (N=1891). As shown in Figure 11, we observe a similar frequency of reporting criteria of evaluation protocols. However, we also find an increasing adoption of LLM-judges for long-form generation tasks.

I Additional analysis

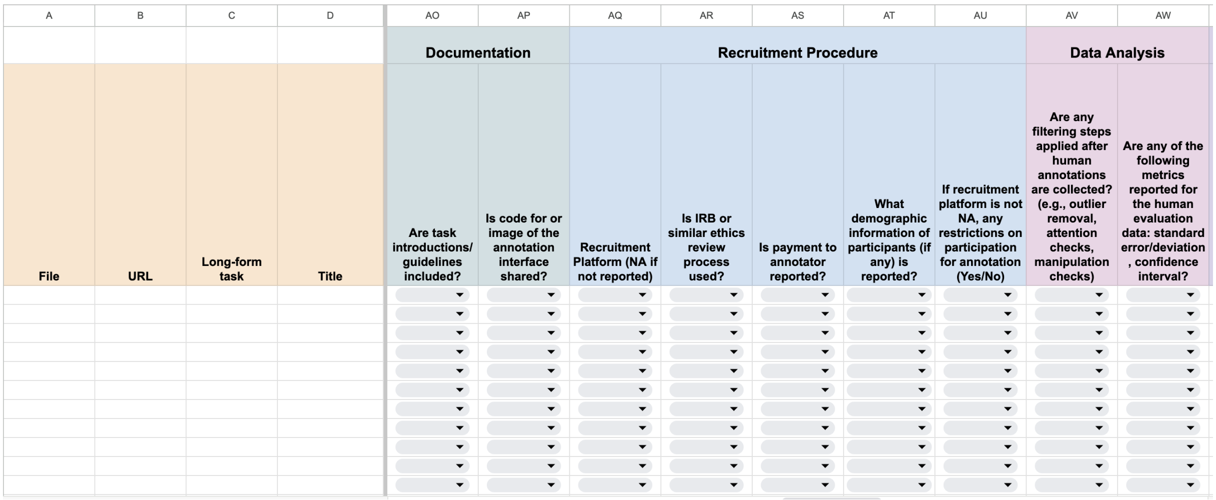
Frequency of documenting reportable criteria varies by the most frequent main tasks of the models. We provide additional analysis of the breakdown of reportable criteria across different common tasks. As shown in Figure 12, evaluated dimensions, number of annotators, and number of annotated samples are often reported among papers that focus on common tasks. However, reporting for other details related to annotation design and analysis remains infrequent across tasks.

Frequency of disagreement resolution approaches In Figure 13, we include the distribution breakdown of disagreement resolution approaches across the sample of papers we annotated.

```
Prompt for Three-Question LLM Labeling (Q1–Q3)

You are helping to fill out a structured research codebook for NLP papers that conduct human evaluation. Respond only using the available options or clearly specified formats.
– BEGIN PAPER TEXT –
<Full paper text>
– END PAPER TEXT –

Answer Instructions: - Output MUST be valid JSON (use double quotes, no comments).
- Use keys: "Q1", "Q2", "Q3", and their corresponding "-reason".
- If something is not reported, set the value as "No or N/A". - For multiple-choice questions, only choose from the listed options.


JSON example format:
{
"Q1": "<"Yes" or "No or N/A">",
"Q1-reason": "<Why do you believe human participants were or were not involved?>",

"Q2": "«"Yes" or "No or N/A"»",
"Q2-reason": "<Describe the model's output and explain why it is or is not considered free-form language generation>",

"Q3": "<Answer varies depending on Q1 and Q2>",
"Q3-reason": "<Explain how you arrived at this answer based on the earlier steps>"
}

Q1: Human Evaluation Involvement Was human judgment involved in evaluating model-generated outputs? Answer "Yes" if any form of human rating, annotation, or qualitative evaluation is present. Otherwise answer "No or N/A". Provide reasoning.
Q2: Free-form Natural Language Generation: What is the model trying to generate?

If the model is generating free-form natural language (e.g., summaries, captions, dialogues), answer "Yes". If the task is extractive, structured, or deterministic (e.g., code generation, translation), answer "No or N/A".
Describe the nature of the output and explain your reasoning.

Q3: Evaluation Details Based on Prior Answers

Now answer based on Q1 and Q2:

If Q1 is Yes and Q2 is True:
What exactly did human participants evaluate (e.g., summaries, explanations)? Be specific.

If Q1 is No and Q2 is True:
Was automatic evaluation used? If "Yes", was an LLM used in that evaluation process?

If Q2 is False:
Skip Q3 and simply write "Q3": "No or N/A" and explain in the reason why it's not applicable.

Respond in exact json format.
```

Figure 5: Prompt used for LLM-based filtering to identify papers studying long-form generation tasks and which employ human evaluation. Papers satisfying both conditions are included for manual annotation (through stratified sampling) and LLM-assisted annotation.

Distribution of IAA We provide a detailed breakdown of the strength of interrater agreement reported by our sample. Around 55% papers did not report the strength of agreement among our annotated sample. Among the papers that reported this information, we find that around 35% of annotated papers reached moderate agreement and above (see Figure 14).



A	B	C	D	AO	AP	AQ	AR	AS	AT	AU	AV	AW
				Documentation		Recruitment Procedure					Data Analysis	
File	URL	Long-form task	Title	Are task introductions/ guidelines included?	Is code for or image of the annotation interface shared?	Recruitment Platform (NA if not reported)	Is IRB or similar ethics review process used?	Is payment to annotator reported?	What demographic information of participants (if any) is reported?	If recruitment platform is not NA, any restrictions on participation for annotation (Yes/No)	Are any filtering steps applied after human annotations are collected? (e.g., outlier removal, attention checks, manipulation checks)	Are any of the following metrics reported for the human evaluation data: standard error/deviation , confidence interval?

Figure 6: Partial screenshot of our annotation interface in Google Sheets showing questions pertaining to documentation, recruitment procedure, and data analysis. These are 9 out of the total 37 questions that annotators answer for each paper they annotate.

Prompt Intro

You are an expert NLP paper auditor.
You will read three PAPER CONTENT blocks that precede this instruction in the prompt:

- ABSTRACT
- INTRODUCTION
- HUMAN-EVAL FILTERED (sections labeled MAIN/APPENDIX)

Answer only using evidence from those blocks. Do not infer beyond them. Be deterministic and conservative: if you are unsure, answer "NA" or "No or N/A".

Strict output rules (per chunk)

- Answer the questions based on the **first human evaluation pipeline** mentioned in the paper.
- **Schema lock (per chunk):** output only the keys for this chunk and their corresponding reason fields (e.g., `Q8`, `Q8_reason`). Do not include any other `Q*` keys.
- Use labels verbatim where specified (e.g., "Yes", "No or N/A", "main", "appendix", "both", "neither").
- For numeric answers, return only numerals (e.g., "3").
- For list items, return a single comma-separated string.
- For each answer, also output the corresponding `Q*_reason` with a short quote or paraphrase (maximum 40 words, include section/page pointers if available).
- Output exactly one flat JSON object (no prose, no code fences).

Figure 7: Prompt for LLM-assisted annotation: input prompt structure for each LLM call.

Chunk Structure (Q1–Q46)

Chunk 1 (Q1–Q7): Overview & Task Setup
Chunk 2 (Q8–Q18): Human Evaluation – Overview & Design
Chunk 3 (Q19–Q33): Human Evaluation – Task 1 Details (Annotators, Samples, IAA)
Chunk 4 (Q34–Q40): Documentation & Recruitment
Chunk 5 (Q41–Q46): Data Analysis, Quality, and Limitations

Figure 8: Prompt for LLM-assisted annotation: chunk structure for codebook questions.

Table 5: Codebook Reference Sheet: these clarifications of codes and answer options are provided to annotators.

Annotation field	Options	Clarification
Are task introductions or guidelines for human evaluation included?	Yes	Paper describes task introduction and instructions for annotators.
	No or N/A	
What is the **domain** that the main task is related to (e.g., medicine, programming)?	General	Select *General* if no specific domain is related.
	Medicine	
	Legal	
	Coding/Programming	
	Journalism	
	Other	
Long-form Generation Evaluation?	Yes	Free-form natural language (e.g., summaries, captions, dialogues); answer "Yes".
	No or N/A	If the task is extractive, structured, or deterministic (e.g., code generation, translation), answer "No or N/A".
Human evaluation?	Yes	If the study involves human participants for evaluation of model-generated outputs (e.g., including benchmark papers where humans assess LLM-generated outputs to curate a benchmark; **exclude** benchmark papers if humans are only used to provide data).
	No or N/A	
Is there more than one human evaluation pipeline?	Yes	Yes if there are human evaluations used for separate tasks or procedures in the study.
	No or N/A	
Which **section(s)** include details about the human evaluation? [comma-separated list]	Open-ended	Copy and paste the section name(s) which involve details of the human evaluation.
What is the **location** of the main design details of human evaluation (i.e., necessary information for reproducing the evaluation)?	main	Select this option if the main details about the human evaluation pipeline (e.g., recruitment, task description, samples) are included in the main paper.
	appendix	Select this option if the main details about the human evaluation pipeline (e.g., recruitment, task description, samples) are included in the appendix.
	both	Select this option if the main details about the human evaluation pipeline (e.g., recruitment, task description, samples) are included in both the main paper and the appendix.
	neither	Select this option if any information related to human evaluation is not found anywhere.
Is human evaluation the only evaluation method being used for assessing model performance?	Yes	This means there are no automatic metrics and no LLMs used to evaluate model performance. **Only human participants** are used to evaluate model outputs.
	No or N/A	
Total sample being annotated		Count unique examples presented for human annotation.
Is **interrater agreement** reported?	Yes	If the study mentions interrater agreement among annotators.
	No or N/A	
Is the number of samples used to compute IAA reported?	Yes	Yes only if the authors describe exact sample counts or state all annotators annotated all samples.
	No or N/A	Often not explicitly mentioned.
How strong is the agreement?	[Select Options]	Interpret numeric values using standard kappa guidelines if no interpretation is provided.
	NA	if no agreement is reported.
How is disagreement treated?	Majority vote	
	Average	
	Pick one	
	Consensus process is applied	A group decision-making process aiming for broad agreement.
	Other	

Full Question List

Q1: Free text (central empirical task, e.g., summarization, dialogue, QA, information extraction, data-to-text, evaluation/benchmarking, classification).
Q2: ACL tracks. ALWAYS answer "NA".
Q3: What is the domain of the main task? options: General, Medicine, Legal, Coding/Programming, Finance. If none is specific, answer "General".
Q4: "Yes" if free-form natural language generation; otherwise "No or N/A".
Q5: "Yes" if humans evaluate model-generated outputs (exclude benchmarks where humans only supply dataset labels); otherwise "No or N/A".
Q6: "Yes" if there is more than one human evaluation pipeline included in the paper; otherwise "No or N/A".
Q7: "Yes" if authors claim they proposed a novel NLP task; otherwise "No or N/A".
Q8: Comma-separated section numbers & names that include human-eval details.
Q9: What is the location of the main design details of human evaluation (i.e. necessary information for reproducing the evaluation); options: "main" / "appendix" / "both" / "neither".
Q10: "Yes" if human evaluation is the ONLY evaluation used to assess model performance; otherwise "No or N/A".
Q11: "Yes" if LLMs are used to evaluate model-generated outputs; otherwise "No or N/A".
Q12: "Yes" if humans and LLMs evaluate the SAME dimensions; otherwise "No or N/A".
Q13: Free text (e.g., coherence, human-like, appropriateness) as a comma-separated list.
Q14: "Yes" if justification for human evaluation dimensions selection is provided; otherwise "No or N/A".
Q15: "Yes" if prior work is cited for justification of human evaluation dimensions; otherwise "No or N/A".
Q16: "Yes" if prior work is cited for the pipeline used in human evaluation; otherwise "No or N/A".
Q17: What is the form of annotation task for human evaluation? Choose one or more: binary, user studies, numeric scale, pair-wise comparison, likert scale, rank-based, categorization, other.
Decide using these rules (don't rely on numbers alone):
• likert scale: Discrete ordinal options with verbal anchors (e.g., strongly disagree. . . strongly agree; poor/fair/good/very good/excellent; very bad. . . very good). Numbers (1–5/7) may appear but anchors define the scale. Keywords: “Likert(-type)”, agree/disagree, poor/good/excellent, very/slightly.
• numeric scale: Pure numeric ratings without Likert-style anchors, often MOS/continuous (e.g., MOS 0–100, “give a score from 1–10” with no named categories). Keywords: “MOS”, “Mean Opinion Score”, “0–100”, “points” with no anchors.
• pair-wise comparison: A vs B preference.
• rank-based: Order multiple systems/items (best→worst, top-k).
• binary: Yes/No, Correct/Incorrect, Accept/Reject.
• categorization: Choose a category label (e.g., error type A/B/C).
• user studies: Interactive/usability tasks with the system (e.g., SUS/UX), not isolated output judgements.
Priority / defaults:
1) If any verbal anchors are present (even alongside numbers)
2) If explicitly MOS or purely numeric with no anchors: numeric scale.
3) If both terms appear, prefer likert scale due to anchors.
4) If ambiguous “rate 1–5 quality” and anchors are implied or unclear: choose likert scale.
Output for Q17 must be a comma-separated subset of: binary, user studies, numeric scale, pair-wise comparison, likert scale, rank-based, categorization, other.
Q18: If "other" in Q17, describe (free text); else "NA".
Q19: Number of annotators. options: Numeric or "No or N/A".
Q20: "Yes" if annotators are referred to as experts or have domain expertise; otherwise "No or N/A".
Q21: Copy/paste expertise description.
Q22: "Yes" if annotators are students; otherwise "No or N/A".
Q23: "Yes" if authors are annotators; otherwise "No or N/A".
Q24: Total unique examples annotated (numeric) or "No or N/A".
Q25: Annotators per item (numeric) or "No or N/A".
Q26: "Yes" if power analysis determines sample size; otherwise "No or N/A".
Q27: "Yes" if IAA is reported; otherwise "No or N/A".
Q28: "Yes" if #samples for IAA is reported; otherwise "No or N/A".
Q29: Numeric #samples for IAA (if reported) or "No or N/A".
Q30: What metrics are reported for interrater agreement? options: Cohen’s kappa, Fleiss’ kappa, Krippendorff’s alpha, Percent agreement, Pearson, Kendall tau, Intraclass Correlation Coefficient, Other, NA.
Q31: How strong is the agreement (report the average agreement level based on kappa interpretation) options: <0 No Agreement, 0–0.20 Slight, 0.21–0.40 Fair, 0.41–0.60 Moderate, 0.61–0.80 Substantial, 0.81–1.00 Almost perfect, or "NA" if not reported.
Q32: Free text or "NA".
Q33: How is disagreement being treated? options: "Majority vote" / "Average" / "Pick One" / "Consensus process is applied" / "Other" / "NA".
Q34: "Yes" if full text of task introductions/guidelines of human evaluation are included; otherwise "No or N/A".
Q35: "Yes" if code or image of the interface is shared; otherwise "No or N/A".
Q36: Recruitment Platform (NA if not reported): Volunteers / Upwork / Prolific / AmazonTurk / Other / NA.
Q37: "Yes" if IRB/ethics review is used; otherwise "No or N/A".
Q38: "Yes" if payment to annotators is reported; otherwise "No or N/A".
Q39: What demographic information of participants (if any) is reported? options: education, age, gender, language, Residence country/Location, other, No or N/A.
Q40: "Yes" if participation restrictions exist when recruiting annotators via platform; otherwise "NA".
Q41: "Yes" if post-annotation filtering (outliers, attention/manipulation checks); otherwise "NA".
Q42: "Yes" if SE/SD/CI reported; otherwise "No or N/A".
Q43: "Yes" if procedures ensure annotation quality (e.g., training); otherwise "No or N/A".
Q44: If Q43 is "Yes", exact text (free text); else "No or N/A".
Q45: "Yes" if limitations of human-eval pipeline are noted; otherwise "No or N/A".
Q46: If Q45 is "Yes", copy/paste limitation text (free text); otherwise "No or N/A".

Figure 9: Prompt for LLM-assisted annotation: full question schema used for LLM annotation.

Table 6: GPT-4o validation accuracy on held-out set of 125 papers. Only questions with validation accuracy greater than 0.75 (shown in **bold**) meet our criteria for presenting results.

Question	Type	Validation Acc.
Category: Task Documentation		
Eval dimensions: justification provided	Binary	0.52
Eval dimensions: justification by prior work	Binary	**0.81**
Task guidelines reported	Binary	0.50
Code/annotation interface reported	Binary	0.71
Category: Annotation Design		
Domain	Multiple choice	0.72
Longform generation	Binary	0.66
Human evaluation	Binary	**0.78**
More than one evaluation task	Binary	0.68
Form of annotation task	Multiple choice	0.34
Claims task is novel	Binary	0.54
Sections w/ human eval details	Multiple choice	0.45
Only human eval used	Binary	**0.91**
LLMs used for eval	Binary	0.72
LLMs and humans eval same dimensions	Binary	**0.76**
Power analysis used	Binary	**1.00**
Recruitment platform	Multiple choice	0.71
Annotator inclusion/exclusion criteria reported	Binary	**0.83**
Payment information reported	Binary	**0.78**
IRB determination	Binary	**0.88**
Method for ensuring quality reported	Binary	0.58
Category: Analysis & Interpretation		
Annotator demographics	Multi-label	0.58
Annotators are students	Binary	**0.84**
Annotators are authors	Binary	**0.86**
Annotators are experts	Binary	0.74
IAA value reported	Binary	**0.78**
IAA metrics	Multi-label	0.54
Data filtering steps reported	Binary	**0.90**
Strength of agreement	Multiple choice	0.58
Disagreement resolution method	Multiple choice	0.74
Statistical metric reported (SE/SD/CI)	Binary	**0.94**
Limitations discussed	Binary	0.58

Table 7: Keyword stem-to-category mapping used to assign papers to primary NLP tasks for visualization.

Category	Stem Keywords Used for Mapping
Dialogue & Interactive Systems	`dialog`, `dialogu`, `convers`, `interact`, `empathi`
Summarization	`summar`, `summari`, `summarizast`
Question Answering	`question`, `qa`, `answer`
Safety & Jailbreak	`safeti`, `align`, `harm`, `jailbreak`, `hallucin`, `toxic`, `hate`, `privaci`, `inappropri`
Reasoning & Planning	`reason`, `plan`, `logic`, `multihop`, `deduct`, `induct`, `counterfactu`, `think`
Instruction & Prompting	`instruct`, `prompt`
Story Generation	`stori`, `narr`, `novel`, `drama`
Style Transfer	`style`, `simplif`
Misinformation Detection	`fake`, `misinform`, `fallaci`
Caption Generation	`caption`, `script`
Personalized Generation	`persona`, `person`, `role`
Information Retrieval	`extract`, `inform`, `retriev`

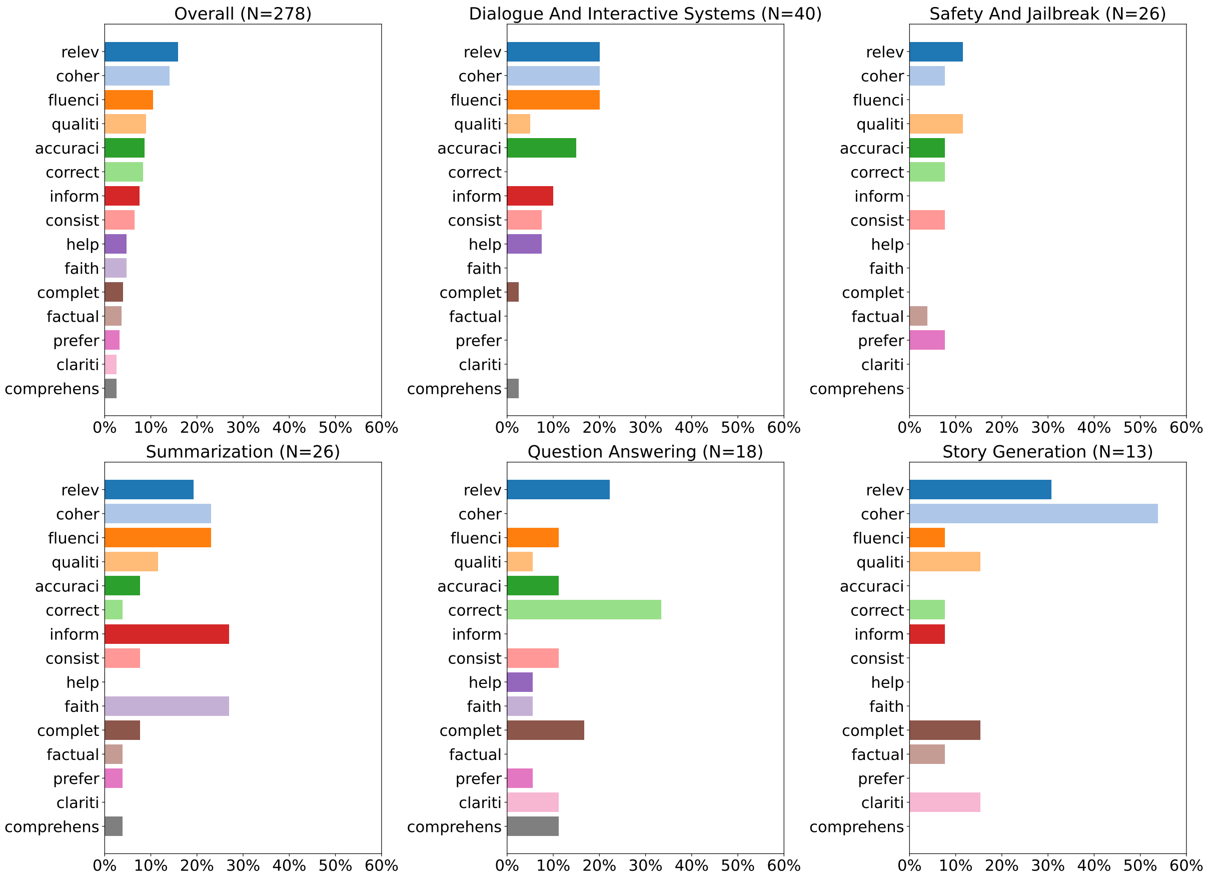


Figure 10: Distribution of stemmed evaluation dimensions across all papers in the manually annotated set (N=278 as 6 out of 284 papers did not report evaluation dimensions), and for the five most frequently occurring NLP task groups.

Table 8: Bootstrapped estimates (N=500) of proportion of *CL papers that report each of the 20 core criteria, along with the sample proportion (as measured over the manually annotated set).

Question	Sample Proportion	Bootstrapped Proportion	Bootstrapped Standard Error
Category: Task Documentation			
Evaluated dimensions reported	0.9823	0.9823	0.0003
Eval dimensions: justification by prior work	0.2553	0.2542	0.0011
Eval dimensions: justification provided	0.5213	0.5209	0.0013
Task guidelines reported	0.5177	0.5169	0.0013
Category: Annotation Design			
Number of annotators reported	0.7660	0.7651	0.0011
Number of annotated samples reported	0.8511	0.8498	0.0009
Power analysis used	0.0000	N/A	N/A
Recruitment platform reported	0.3511	0.3503	0.0013
Annotator inclusion/exclusion criteria reported	0.1418	0.1420	0.0009
Method for ensuring quality reported	0.2163	0.2179	0.0011
Code/annotation interface reported	0.2801	0.2802	0.0012
Payment information reported	0.2872	0.2872	0.0011
IRB determination reported	0.1099	0.1099	0.0008
Category: Analysis & Interpretation			
Annotator demographics reported	0.7092	0.7094	0.0012
IAA value reported	0.4610	0.4616	0.0013
IAA sample size reported	0.1702	0.1710	0.0010
Disagreement resolution method reported	0.3050	0.3057	0.0012
Data filtering steps reported	0.0567	0.0573	0.0006
Limitations discussed	0.1844	0.1858	0.0010
Statistical metric reported	0.0957	0.0973	0.0007

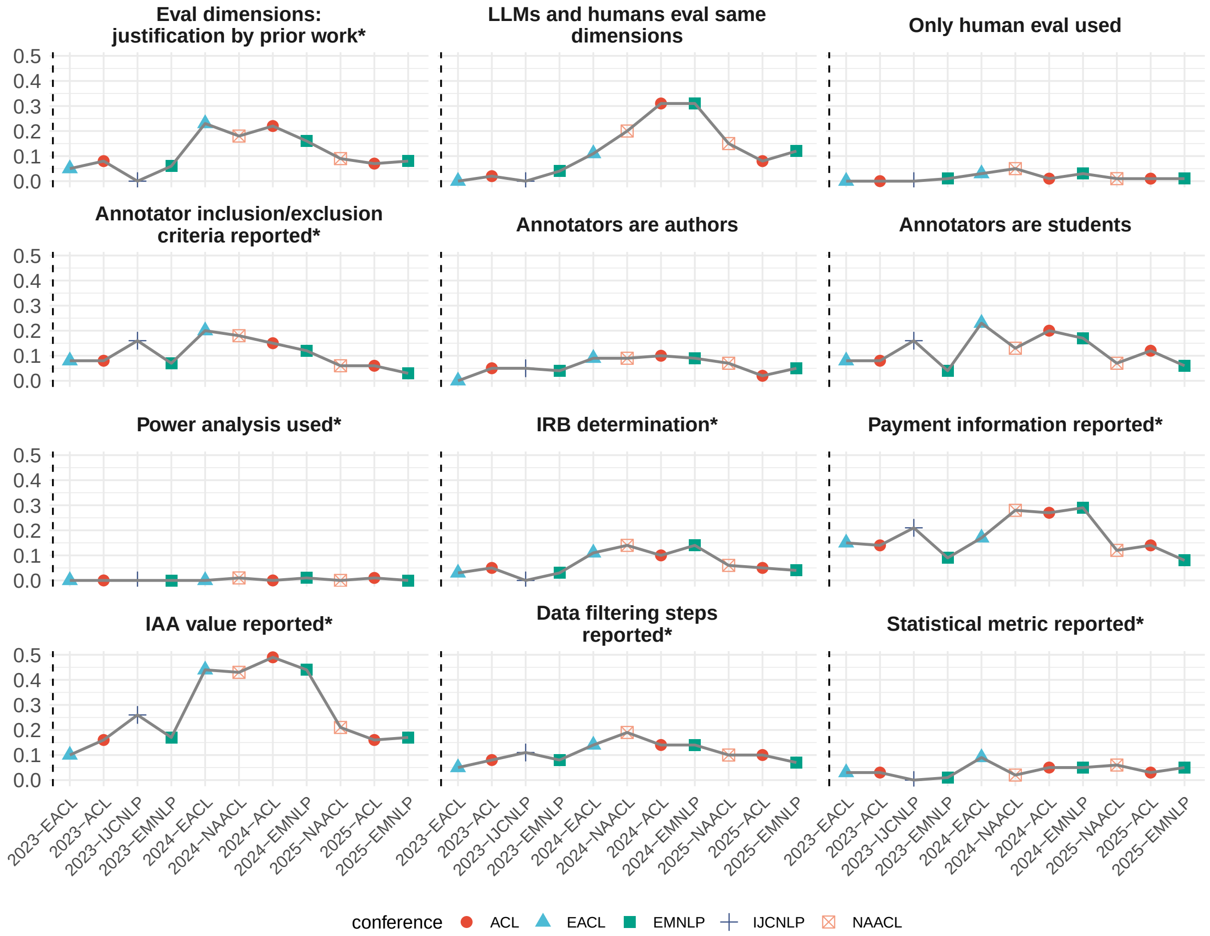


Figure 11: Temporal trends in reporting: across all *CL papers (2023-2025) with human evaluation and long-form generation (N=1,891), frequency of reporting criteria of evaluation protocols remains similar. Notably, we find that the use of LLM-judges is on the rise. Criteria marked with * are among the 20 core reportable criteria.

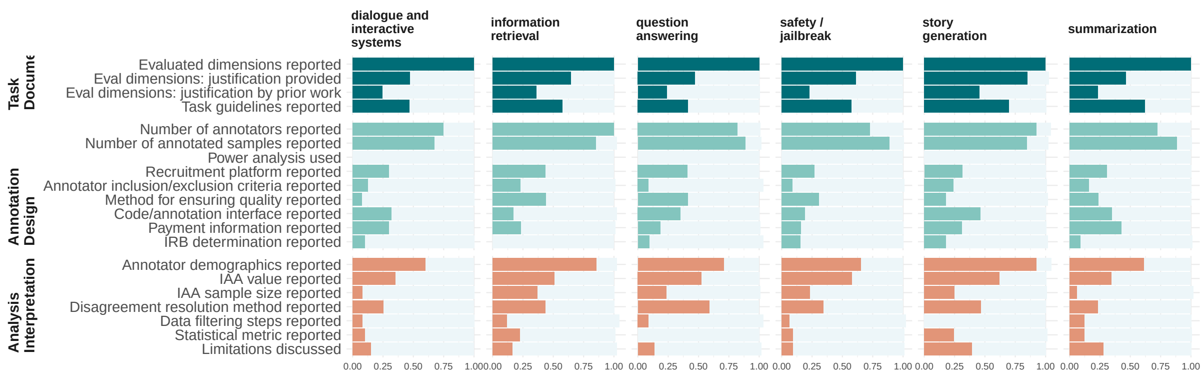


Figure 12: Frequency of Reporting Criteria for Common NLP Tasks

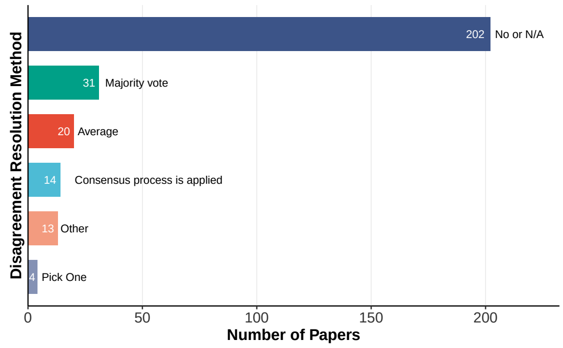


Figure 13: Frequency of disagreement resolution method reported in manually-annotated sample: Most papers tend not to report how they address disagreement among annotators (n=202). Among the ones that report this criteria, majority vote (n=31) is the most common approach for addressing disagreement among annotators, followed by averaging (n=20), consensus process (n=14), other (n=13), or picking one annotation (n=4).

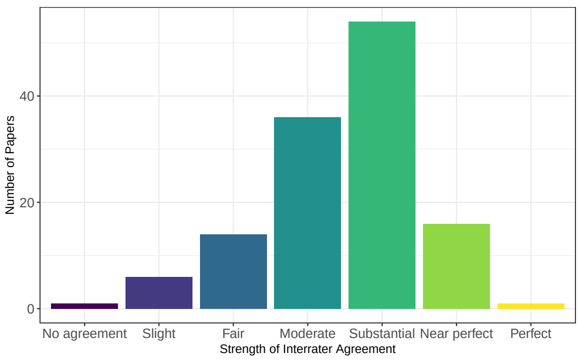


Figure 14: Distribution of IAA strength reported in manually-annotated sample. If an IAA metric value is reported (e.g., Cohen's kappa), we classify the metric value into strength of agreement based on how the metric is usually interpreted.