

Automatic Metrics in Natural Language Generation: A Survey of Current Evaluation Practices

Patřicia Schmidtová¹ ✉, Saad Mahamood², Simone Balloccu¹,
Ondřej Duřek¹, Albert Gatt³, Dimitra Gkatzia⁴,
David M. Howcroft⁴, Ondřej Plátek¹, and Adarsa Sivaprasad⁵

¹Charles University, Faculty of Mathematics and Physics, Prague, Czechia

²trivago N.V., Düsseldorf, Germany

³Utrecht University, Utrecht, Netherlands

⁴Edinburgh Napier University, Edinburgh, Scotland, United Kingdom

⁵University of Aberdeen, Aberdeen, Scotland, United Kingdom

✉ Corresponding author: schmidtova@ufal.mff.cuni.cz

Abstract

Automatic metrics are extensively used to evaluate natural language processing systems. However, there has been increasing focus on how they are used and reported by practitioners within the field. In this paper, we have conducted a survey on the use of automatic metrics, focusing particularly on natural language generation (NLG) tasks. We inspect which metrics are used as well as why they are chosen and how their use is reported. Our findings from this survey reveal significant shortcomings, including inappropriate metric usage, lack of implementation details and missing correlations with human judgements. We conclude with recommendations that we believe authors should follow to enable more rigour within the field.

1 Introduction

Evaluation practices in the field of Natural Language Processing (NLP) are increasingly coming under a microscope by researchers. There is now a significant body of contributions presenting experimental research, meta-analyses and/or best practice guidelines, on issues ranging from statistical significance testing (Dror and Reichart, 2018), to human evaluation methods (Howcroft et al., 2020a; van der Lee et al., 2021; Hämäläinen and Alnajjar, 2021; Shimorina and Belz, 2022a), error analysis (van Miltenburg et al., 2021a, 2023) and replicability of evaluations (Belz et al., 2021a, 2023a).

Automatic metrics and their usage for evaluation have also been under extensive examination by researchers. Similarity-based metrics are sometimes taken as proxies for human quality ratings, whereas findings suggest the two should not be conflated. This has led to concerns about metric validity (Belz and Gatt, 2008). For example, the validity of metrics such as BLEU (Papineni et al.,

2002) and ROUGE (Lin, 2004) has been put into question regarding their poor correlation with human judgements (Reiter and Belz, 2009; Novikova et al., 2017; Reiter, 2018). In addition, automatic metrics do not capture factuality or faithfulness issues in text (Gehrmann et al., 2023), such as incorrect names and numbers (Thomson and Reiter, 2020). Interpreting the meaning of scores generated by automatic metrics can also be challenging. For example, what researchers often report as a “BLEU score” actually consists of several metrics, depending on multiple parameters and varying across different implementations, which are not compatible with each other (Post, 2018). There are also questions on whether it is possible to encapsulate the performance of a given system with a single number or whether the use of a single metric to demonstrate improvements over prior systems provides sufficient dimensionality in reporting the performance characteristics of a given system (Gehrmann et al., 2023).

Given the well-documented shortcomings of automatic metrics, our goal in this paper is to survey the current state of play in metric-based evaluations of natural language generation (NLG). As with the above-mentioned studies focusing on other facets of evaluation, we aim to both understand how metrics are currently used in NLG, and to identify gaps and possible ways forward in an effort to improve the scientific quality of NLG research.

Specifically, we conduct an analysis of published work in the field, annotating which metrics are used, for what purposes, and how their usage is reported. In Section 2, we describe past survey efforts within the field of NLG to frame our contribution. In Section 3, we describe our paper selection procedure, the annotation procedure, the challenges we encountered, and the process and results of comput-

ing inter-annotator agreement between the annotators. The analysis and results from the annotation process are presented in Section 4, and we offer our insights into these results in Section 5. Finally, we wrap up with recommendations (Section 6) and concluding thoughts (Section 7) from the observations based on our results.

2 Evaluation Surveys in NLG

There have been several surveys inspecting the different aspects of evaluation practices within NLG over the last several years. Some surveys focused on quantifying the types of evaluations, the proportion of intrinsic and extrinsic evaluations over a defined period of time either for the field as a whole (Gkatzia and Mahamood, 2015), or for a specific domain such as question generation (Amidei et al., 2018). In addition, there has been an effort to understand the different types of metrics and evaluation approaches employed and to categorise the challenges faced by researchers (Celikyilmaz et al., 2020).

In addition to survey work covering shortcomings of automatic metrics (Gehrmann et al., 2023), a significant amount of work has focused on human evaluation practices within NLG. Past work has revealed a large variation in practices among researchers (van der Lee et al., 2019). This was followed up by an extensive survey which has shown that in addition to the large variety of practices, there are fundamental gaps in reported details by authors (Howcroft et al., 2020b). These issues have led to proposals for best practices for carrying out and reporting human evaluations in NLG (van der Lee et al., 2021; Shimorina and Belz, 2022b). However, the concern about human evaluation practices has also led researchers to consider whether human evaluations in NLG – and in NLP as a whole – are both reproducible and repeatable (Belz et al., 2023b) given the inconsistencies and gaps in reporting practices.

One area where reporting practices have received attention is the way in which errors made by NLG systems are documented. Van Miltenburg et al. (2021b) found that there is severe under-reporting of the different kinds of errors a given NLG system can make, which leaves the broader community “in the dark” due to this missing information. Beyond evaluations and reporting practices, there have been attempts to better understand the motivations of researchers and their reporting practices by directly

surveying them. Zhou et al. (2022) found that there is pressure towards a “kitchen sink” approach for evaluation. Even though researchers recognise the limitations of existing metrics, lack of clarity about their evaluation goals and quality criteria can lead to over-reporting potentially uninformative metrics (Zhou et al., 2022). Other work explored the barriers that researchers face to conducting error analyses (van Miltenburg et al., 2023): while respondents were generally positive about error analyses, there are multiple barriers such as page limits, lack of tools or resources, and a lack of time and/or money.

3 Survey Method

Although past surveys looked at the deficiencies of automatic metrics, none of them go beyond quantifying and aggregating their usage. This is necessary, considering that the use of automatic metrics increased by almost 25% in the 2016-2019 period, with some surveys reporting that almost half of the papers surveyed only use automatic metrics (van der Lee et al., 2021). To obtain a comprehensive and up-to-date view of current practices in automatic evaluation for NLG, we have focused on recently published articles in prominent, peer-reviewed venues.

Paper selection Our analysis is based on a snapshot of a total of 110 papers presented in 2023 in two relevant venues: the *International Conference on Natural Language Generation* (INLG) and the *Annual Meeting of the Association for Computational Linguistics* (ACL). All papers ($n = 36$, of which 26 are long papers) at the main conference track of INLG 2023 were included. For ACL, we used all the papers presented under the *Generation* track ($n = 74$, 63 are long papers). In addition to regular ACL papers, this included three papers originally accepted for publication in the *Transactions of the ACL* (TACL) journal and one NLG paper from the journal *Computational Linguistics*.

Annotation procedure Papers were randomly distributed among all the authors in a set of annotation batches, and independently annotated for the features summarised in Table 1. As the table indicates, the main purpose of the annotation was to identify which *automatic* evaluation metrics or *human* evaluation methods (if any) are reported in the paper and for which tasks. A full list of evaluation methods identified is provided in Appendix C. We

Feature	Description	IAA (J)	IAA (M)
Name	Which evaluation method was used? (Appendix C).	0.59	0.34
Newly introduced?	Was the metric newly introduced in this paper?	0.76	0.76
Task	Which task(s) this metric was used to evaluate (Appendix B)?	0.41	0.35
Human Correlations	Were automatic metric results directly related to human evaluation results? Was this correlation quantitative or qualitative?	0.47	0.47
Implementation	Were specific metric implementation details (e.g. links to the specific metric implementation, paper reference, etc.) provided or not?	0.44	0.45 ¹
Appendix	Was the metric only reported in the Appendix, rather than the main section of the paper?	0.61	0.60
Rationale	Did the authors explain the rationale for the metric?	N/A	N/A

Table 1: Features annotated for each paper. IAA (J/M): inter-annotator agreement between 6 authors for 4 papers on each criterion, using the Jaccard or MASI distance metrics.² Note that ‘Rationale’ is not included in the agreement computation since it was recorded in a free-text form to allow for more flexibility.

annotated the task type using definitions created by Howcroft et al. (2020b); annotators could also include other tasks not in this list if necessary (see Appendix B for details). Note that it is possible for papers to report different metrics for different evaluation experiments, depending on the (sub)task. Crucially, we also consider whether a metric is newly introduced in a paper or was previously published. In either case, we are interested in whether authors describe the rationale for their use of a metric. In case a paper included a human evaluation, we also annotate whether metric-based evaluations were quantitatively correlated with the outcomes of the human evaluation, or whether there was any qualitative discussion of the relationship between the two.

Iterative refinement and inter-annotator agreement Annotation proceeded in multiple rounds. During an initial round, we independently annotated a subset of papers and discussed the outcomes to fine-tune the annotation scheme. Subsequently, a random sample of 4 papers (2 from INLG; 2 from ACL) was selected and independently annotated by 6 of the authors. Inter-annotator agreement (IAA) for the features outlined in Table 1 was computed using both the Jaccard and MASI (Passonneau, 2006) distance metrics.² Following discussions, we addressed the disagreements by replacing the

originally annotated *link to metric* with *implementation details* and reporting *task* using a selection from a drop-down list following Howcroft et al. (2020b)’s definitions.

4 Analysis and Results

We present the results of our annotation of 110 papers in this section. Out of the 110 papers annotated, a total of 102 papers included an evaluation of a generation system. The excluded 8 papers did not propose any systems to be evaluated. For example, they either presented a corpus or methods to detect the decoding algorithm of a closed-source model. After the removal of these papers, a total of 69 ACL papers and 33 INLG papers were analysed.

A total of 59 papers (56.73%) of papers use human evaluations; in contrast, 98 papers (94.23%) used automatic metrics, a result similar to what was found by van der Lee et al. (2021), who reported 95% of papers using automatic metrics in both ACL tracks and INLG. There were only 53 papers (50.96%) containing both automatic and human evaluation.

Another aspect explored was whether authors provide any implementation details, such as link to the specific implementation used for the evaluation. We found that for 66.2% of INLG and 57.3% of ACL papers, these details were not mentioned either in the main body of the paper or within the appendices. Given the high percentage of papers not giving specific implementation details, this can make it difficult to conduct reproduction studies under the same conditions, especially, considering how challenging it is to reproduce the original scores of NLP evaluations (Belz et al., 2021b).

In the subsequent sections, we will explore in more detail how specific metrics are used (Sec-

¹Note that these correlation values relate to the initial annotation guidelines and the *link to metric* property, which directly compared implementation URLs and was not clear on the procedure if a paper used multiple implementations. The feature was then changed into the categorical *implementation details* with three options: no implementation details provided, implementation details provided, and multiple implementations used.

²We estimate agreement using the `AnnotationTask` class and `jaccard_distance` and `masi_distance` functions in the NLTK metrics library (Bird et al., 2009).

Metric Family Name	INLG	ACL	Total
BLEU	26	69	95
ROUGE	27	65	92
N-gram diversity	6	49	55
Style Classifier	5	37	42
BERTScore	8	32	40
Perplexity	3	29	32
METEOR	6	21	27
Semantic Similarity	9	12	21
Overlap	6	21	27
Factuality	5	13	18
Accuracy	8	8	16
Quality Estimation	7	7	14
Combination	0	14	14
BARTScore	2	10	12
NLI	44	8	12
F1	4	7	11
BLEURT	5	5	10
CIDEr	2	6	8
N-gram repetition	2	6	8
SARI	2	6	8
Sequence Length	3	5	8
MAUVE	0	8	8
Unieval	0	8	8
Distribution Comparison	0	7	7
NIST	0	7	7
MoverScore	1	5	6
PARENT	1	5	6
Recall	2	4	6
Edit Distance	1	5	6
Flesch Readability	1	3	4
Inference Speed	0	4	4
Precision	1	2	3
loss/error	0	3	3
chrF++	1	1	2

Table 2: Total automatic metric usage counts of each of the metric families for both INLG and ACL conferences.

Metric Task Name	INLG	ACL	Total
Overlap	71	201	272
Semantic Similarity	20	59	79
Match	15	61	76
Text Properties	12	63	75
Text Classifier	17	57	74
Factuality	49	21	70
Perplexity	3	37	40
Distance-based	1	15	16
Combination	0	14	14
Inference Speed	0	4	4

Table 3: Total usage counts of each of the high-level metric categories for both INLG and ACL conferences.

tion 4.1), what the relationship is between automatic and human evaluations (Section 4.2), how these relate to different NLG subtasks (Section 4.3), and whether the papers provide their code (Section 4.4), an important consideration given the concern about evaluation reproducibility.

4.1 Metric-Level Analysis

We identified 634 counts of automatic metric uses within these papers, with 283 different automatic metric names used by practitioners. To enable further analysis of these metrics and to derive useful insights into researcher practices, we manually grouped the metrics into 38 *metric families* that group together similar metrics. In particular, we

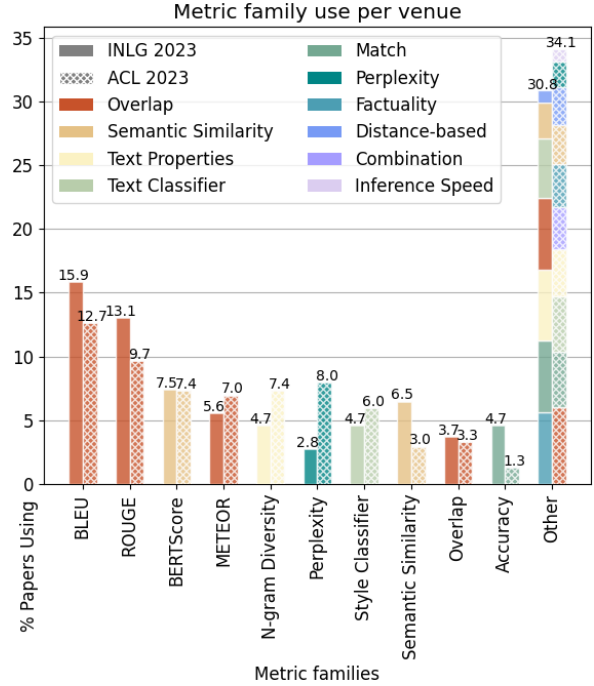


Figure 1: Usage percentages of top 10 metric families in INLG and ACL, with metric category color-coded.

aimed at the most informative grouping possible: We merged similar metrics which are individually relatively rare, while keeping frequently used metrics within their own family (e.g., BLEU). We further joined the metric families into 10 broad *metric categories* to enable a more high-level overview. Table 3 lists all metric categories with their usage counts across the surveyed papers. Table 2 shows the number of metric occurrences in papers across metric families, with colour codes corresponding to metric categories in Table 3. The overall usage of the most frequent metric families and the corresponding categories is further depicted in Figure 1. The full list of all identified metrics and their grouping can be found in Appendix C.

Frame of comparison: We further divide metrics into *reference-based* (use a human reference or pairwise output from another system), *source-based* (mostly checking for output fidelity/alignment with the input), *output only* (evaluating inherent text properties such as diversity), or *source and reference based*. We find that the dominant form is reference-based metrics: As show in Figure 2, this holds true in both INLG and ACL papers, with this metric type used more extensively in INLG compared to ACL. This suggests that researchers are primarily looking to evaluate the outputs of systems against reference corpora to get

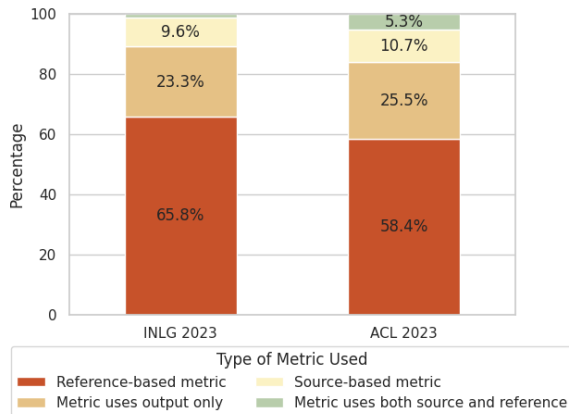


Figure 2: The percentage of automatic metric types used in both INLG and ACL conferences.

an estimation of performance. Some metrics, such as SelfBLEU, can be used in multiple different ways, which may inflate the usage estimates for reference-based metrics.

BLEU and ROUGE: Across both INLG and ACL papers, BLEU and ROUGE are the predominant metrics used for NLG automatic evaluations, as seen in Table 2. This is in line with previous qualitative observations (van der Lee et al., 2021; Gehrmann et al., 2023). Interestingly, as shown in Figure 1, the usage of BLEU and ROUGE is proportionately higher in INLG compared to the ACL Generation track. BLEU is the most popular metric in both INLG and ACL, despite the multiple concerns raised by researchers on its validity as an NLG metric (Reiter, 2018). Moreover, for 63.6% of papers using BLEU and 62.6% of those using ROUGE no implementation details were provided, despite the compatibility issues this can cause (Post, 2018; Grusky, 2023).

Trainable metrics (mostly from the Semantic Similarity, Text Classifier, and Factuality categories) only make up a minority, with 28.4% in INLG and 35.5% in ACL, respectively. This suggests that even though learning-based metrics such as BERTScore (Zhang et al., 2019), BLEURT (Selam et al., 2020), etc. are gaining traction, they are still not as popular as more basic approaches.

Metric Rationales: The vast majority of annotated metrics (486, 76.9%) did not include a rationale for the use of a metric. A total of 65 mentions of metrics in papers (10.3%) stated that they were following previous work. Authors rationalized five of the metrics by stating that they correlate with

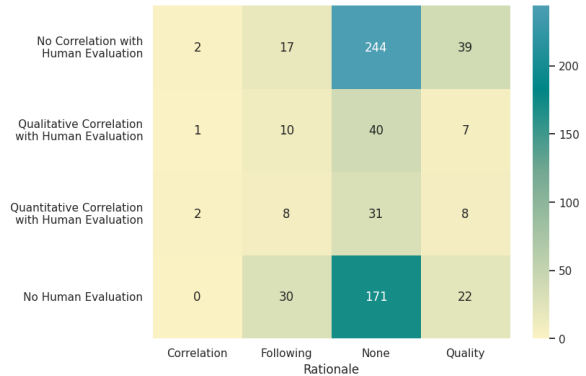


Figure 3: Co-occurrence of types of rationales with the authors correlating the metric results to human judgement.

human judgements, generally shown by previous work. Finally, for 76 metrics (12.0%), a rationale other than following previous work or correlating with human judgement was stated in the papers, e.g. that the given metric was included to measure fluency or diversity.

We also looked at the relationship between the type of rationale given for a metric and whether a correlation with human evaluation was discussed (Figure 3). It is very clear that for a vast majority of metrics no rationale is provided, irrespective of whether a human evaluation has been conducted or not.

New Metrics: We found that 76 new metrics were introduced, with eight of them named and proposed for future use:

- AlignScore (Zha et al., 2023)
- NegBleurt (Anschütz et al., 2023)
- NegMPNet (Anschütz et al., 2023)
- HAUSER (He et al., 2023a)
- WeCheck (Wu et al., 2023a)
- NEHR (Akani et al., 2023)
- LENS (Maddela et al., 2023)
- DecompEval (Ke et al., 2023)

All of these metrics are based on trainable components and mostly focus on factual correctness, going against the majority currently in use, but reflecting an emerging trend. It would be interesting to observe in the future whether these new metrics are adopted by the research community or not.

Appendix: We observed that, for a given paper, some metrics are only reported in the papers’ appendices. This was the case for nine metrics (4.8%) at INLG and 22 metrics (3.8%) at ACL.

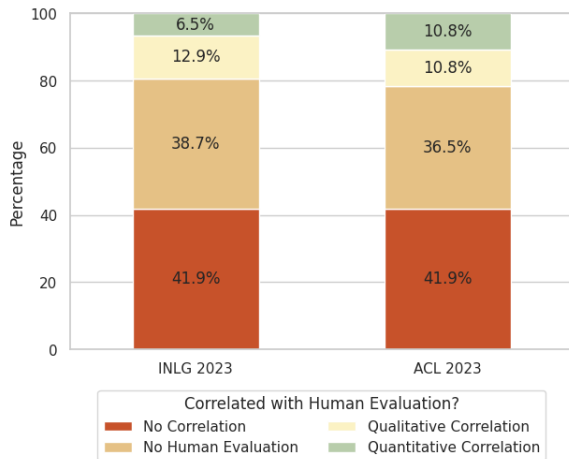


Figure 4: The percentage of papers that state a form of correlation between their automatic and human evaluation results.

4.2 Automatic vs. Human Evaluations

We conducted an additional analysis to better understand whether researchers treat their automatic and human evaluations as separate entities, or seek a more unified interpretation of results from the two, by looking for correlations between them. We annotated papers with four approaches to their human evaluations:

- *Quantitative Correlation* - Cases where the authors check if their automatic metric result(s) quantitatively correlate with evaluation results from their own or previous work.
- *Qualitative Correlation* - When authors only draw qualitative conclusions on the relation between their automatic and human evaluation results, without statistical analysis to back this claim.
- *No Correlation* - No stated correlation either quantitatively or qualitatively can be found in the paper.
- *No Human Evaluation* - No evaluation involving human participants was performed by the researchers.

Interestingly, papers from the ACL generation track and INLG are very similar in terms of correlating with human evaluations, as shown in Figure 4. Papers predominantly either did not perform a human evaluation or if they did, they did not check for a correlation between their automatic and human evaluation results. Authors who provided either a qualitative or quantitative analysis between their automatic and human evaluation results are very much in the minority.

Task Name	INLG	ACL	Total
Summarisation (text-to-text)	6	17	23
Feature-Controlled Generation	5	13	18
Dialogue Turn Generation	3	10	13
Data-to-text Generation	5	8	12
Machine Translation	0	10	10
Question Generation	1	9	10
Paraphrasing/Lossless Simplification	1	9	10
Question Answering	0	8	8
End-to-End Text Generation	1	7	8
Story Generation	3	3	6
LM Sampling	2	3	5
Referring Expression Generation	2	0	2
Content Selection/Determination	1	2	3
Surface Realisation (SLR to Text)	0	2	2
Song Lyric Generation	0	2	2
Compression/Lossy Simplification	0	2	2
Commonsense Reasoning	0	2	2
Aggregation	0	1	1

Table 4: List of NLG task types, with counts of relevant papers from the annotated sets. Task definitions are based on those used by Howcroft et al. (2020b).

A possible reason for the low level of reported correlations between automatic and human evaluations could be the known issues between lexical overlap evaluation metrics and specific NLG sub-tasks, such as referring expression generation (Belz and Gatt, 2008). An alternative possibility is that while automatic metrics may give an approximate estimate of language quality, they do not measure content quality (Reiter and Belz, 2009) and therefore researchers are looking to measure different aspects with their automatic and human evaluations.

4.3 Task Representation

Table 4 shows the counts for each of the task types, with the majority of papers focusing on text-to-text summarisation. We analysed the relationship between the paper task and metric usage, shown in Figure 5. Overlap metrics dominate most tasks, especially question generation (75%) and data-to-text generation (61.6%). Interestingly, feature-controlled generation seems to be the only task that sees some of the lowest usage of Overlap metrics (17.8%); moreover, in comparison to other tasks it is the only one where other metrics are dominant.

4.4 Paper Resources Findings

Our last area of analysis was the completeness of paper code resources. Given the importance of complete code and resources for the reproduction and repeatability of experiment results, we manually checked papers to see not only if they provided

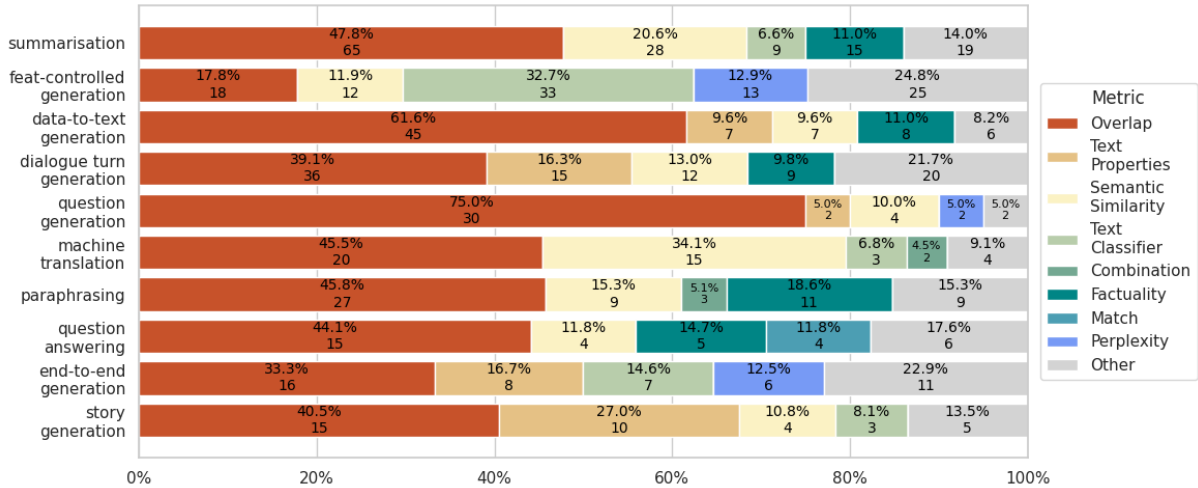


Figure 5: Distributions of different metric families used to evaluate a given task across ACL and INLG (with percentages of metric usages for the given task on top and absolute counts below).

a link to an implementation, but also if the given link contained any code or data.

Annotation approach: We classified papers into three groups: *delivered* if the code was present, *no code* if not and the paper did not promise any code, and finally *missing*, which applied to papers that linked to code repositories, but these were either dead, empty or contained only abstracts or titles or promises of a future release. For papers that delivered code, we also annotated the following aspects (see appendix D.1 for more details):

- Granularity of installation instructions: *None, Basic, Detailed*
- Clarity of experiments structure in the code, whether experiments described in the papers are “discoverable”: *None, Some, Many*
- Level of documentation detail, such as if hyperparameters are described and how experiments can be executed: *None, Basic, Detailed*

Code availability: In terms of available code, 75% of INLG and 70.2% of ACL papers published their code. 18.2% INLG papers and 11.9% ACL papers published no code. This is similar to the results of Mieskes et al. (2019), who found no code in 14% cases and no experimental resources in 11.1% cases. A larger proportion of ACL papers (17.9%) promised to deliver code but did not, compared to 6.8% for INLG. We examined the papers annotated as *missing* to further understand if there was a difference between authors who come from industry as compared to academia. Papers were classified as being “industry” papers if a majority of author affiliations are not from an academic in-

stitution. We found that the majority of *missing* papers have either complete or partial industry authorship ($n = 13$), compared to purely academia papers ($n = 5$). Whilst the numbers detected are too small to draw definite conclusions, we hypothesise that additional business constraints increase the likelihood of not releasing the code even if promised by the authors.

Examining code releases: For papers that had published code, we considered the level of detail of the installation instructions provided. For 52.7% of ACL papers and 50% of INLG papers, no installation instructions were provided. For the remaining papers, 13.9% and 10.8% of INLG and ACL papers respectively provided basic installation instructions. This leaves a minority of 36.1% and 36.5% of INLG and ACL papers with detailed installation instructions.

A similar story holds for how discoverable experiments are within papers that have published code. In only a minority of papers (27.8% for INLG and 37.8% for ACL), half or more of their experiments could be directly linked to scripts provided within the code.

In terms of code documentation, an alarming 44.4% of INLG and 43.2% of ACL paper resources provide no instructions whatsoever.

Metrics and Paper Resources: We also explored the relationship between inclusion of metric implementation details in a paper and the availability of paper resources. Figure 6 shows a visualisation of this analysis. The main point that stands out is that for metrics with no implementation details,

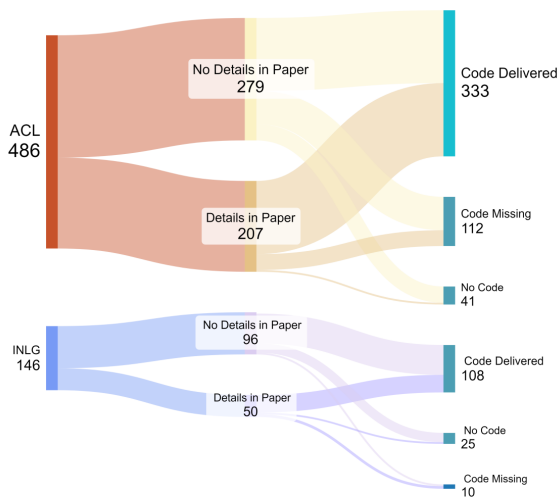


Figure 6: The proportion of metrics across ACL and INLG and the availability of paper resources.

there is a larger proportion of papers with missing code. This seems to hold true for both ACL and INLG metrics.

5 Discussion

Our survey reveals both positive and negative aspects of current trends in NLG evaluation. Undoubtedly positive is the fact that the vast majority of researchers do make their code and resources available after publication, despite no obligation to do so. Additionally, it is encouraging to see that types of metrics used differ given the task, suggesting an effort to use metrics which are relevant to the research goals. Overlap metrics are mostly complemented by metrics from other categories (cf. Figure 9 in the Appendix).

On the other hand, the predominance of Overlap metric types is concerning given their well-known caveats, such as their inability to measure faithfulness and poor correlation with human judgements (Reiter, 2018). This is also compounded by the tendency to not state the rationale for the use of a metric. Without any rationale of why a given metric or set of metrics are being used, there is uncertainty on what researchers are looking to measure and whether they chose the right metrics. Our survey also reveals an over-reliance on reference-based metrics. This might be a hold-over from when generation tasks were more highly constrained and focused on more closed-domain problems such as weather forecast generation, with a defined set of

reference “gold-standard” corpora. However, most generation problems are increasingly open-ended and require accepting a wider range of outputs that are not possible to cover in a given reference set. Therefore, it is possible that an attitudinal or structural change is needed within the research community to ask deeper questions on the use of inappropriate metrics.

Another observation is the relationship between automatic and human evaluations. Out of the metrics that had no rationale provided, around half performed human evaluations, yet did not investigate any link between the automatic and human evaluation results. This suggests that the majority of researchers treat their evaluations as separate entities. However, given the overall lack of rationales provided for the use of automatic metrics, we cannot be certain that authors are looking to measure different aspects with their automatic and human evaluations or whether the evaluations are in fact intended to be complementary. Ultimately, this creates uncertainty for researchers reading papers and makes the reproduction of evaluations challenging.

6 Recommendations

6.1 Evaluation Quality

Rationalize your selection of metrics Authors should consider the appropriateness of the metrics they are using and whether adding more automatic metrics will in fact yield interesting insights. In particular, we advise authors to state clearly what they expect to evaluate with each given metric so that there is clarity for those trying to interpret reported results. In our investigation, we found that less than 13% of metric occurrences are supported by a rationale other than following previous work. Rationales are also important due to the number of metrics used – 283 unique metrics were used at the surveyed venues last year. We cannot reasonably expect readers to be familiar with all of them, which strengthens the need for justification.

Do not copy-paste widely used metrics We found that around 10% of metric usages (and an unknown portion of the 77% with no rationale provided) are justified on the grounds that they follow evaluations done in previous work. Authors should question whether these metrics truly measure the intended qualities in the evaluation, and if they do, the authors should share their reasoning in the paper. However, if the metrics fail to show a correlation with human judgment or a specific quality,

we strongly advise authors to omit them, or at least relegate them into the appendix to clearly show their decreased priority.

Comment on metric combinations Given that automatic metrics frequently have blind spots, we also recommend commenting on the chosen combination of metrics: how do the metrics complement each other to provide a more objective evaluation of a system?

Respect the intended use of metrics Generally, when a new metric is proposed, its authors demonstrate its suitability for a given setting or task. However, we frequently see metrics used for purposes that they were not intended for. In such a case, the authors should justify their use of the metric from first principles or empirically.

Discuss (dis)agreements between human and automatic evaluation For both automatic and human evaluations, it is important to state the similarities or differences between their measurements. Where there are overlaps in what is being measured, authors should consider commenting on whether they see correlations between the reported results or not.

6.2 Reproducibility

Share evaluation details When using a library implementation of an automatic metric, the authors should first and foremost disclose which library was used – this happened for only 34.2% of the metrics used at INLG and 42.6% at ACL. Furthermore, it is also desirable to share in the appendix the parameters used to obtain the results. Such parameters can include the version of the library, the tokenizer, the preprocessing methods, and so on. Even better, some libraries, such as SacreBLEU (Post, 2018) include easily shareable version strings with the encoding of these parameters.

Share data samples The lack of error analyses conducted within the NLG research community is a known problem (van Miltenburg et al., 2023), given the lack of comprehensiveness of both automatic and human evaluations. If possible, authors should consider sharing example outputs with metric results and adding human annotations (if a human annotation has been performed).

Additionally, we encourage the authors to release the full datasets with the evaluated system outputs. As a result, the future authors will have

the possibility of using other, possibly new metrics to compare to their new systems.

Release code The final set of recommendations relate to provision of experimental code and resources. While code is often provided now, practices still vary considerably. Improvements include not just releasing the code for the evaluations conducted, but also giving appropriate installation instructions and describing how the code relates to results in the paper. The inclusion of generated outputs enables evaluation reproductions and allows future evaluations with newer or alternative metrics. Finally, a structural improvement that the research community could consider is to make code and resources a requirement, subject to validation, with the camera-ready version of an accepted paper.

7 Conclusion

We have presented our analyses and a new dataset of 102 papers annotated with nine attributes to ascertain the different metrics, used currently by authors in NLG across publications in 2023 in both INLG and ACL venues. The process of creating and validating the annotation schema, the analyses that we have conducted, and the results we have obtained are described in this paper.

From the results that we have obtained, we have shown that there are outstanding issues related to the type and number of metrics used, the lack of comparison and linkage between automatic and human evaluation results, and missing justifications for the selection of metrics.

We have proposed several recommendations in the hope to offer possible solutions to these structural problems. However, while many papers have or will make recommendations on improving evaluation practices, it is only when these solutions are adopted that we as a research field can make progress on these issues.

Our main conclusion is, that as a field, we need to provide more information on the usage of automatic metrics and the motivations behind their usage. Only by doing this can we start to bring more clarity to how evaluations are being conducted and help to alleviate adjacent challenges such as the reproduction and repeatability of evaluations.

Limitations

While this work provides a snapshot of automatic evaluation practices in NLG during 2023, quantita-

tively capturing long-term trends in these practices was out of the scope of this work.

Ethics Statement

The focus of this work is to gain better insights into automatic evaluation practices. The annotations made in this paper were made by the authors and therefore we did not recruit any external annotators nor process any personal data.

Supplementary Materials Availability Statement

The code for the analysis, the selection of papers, and the annotations presented in this paper are made available from our GitHub project repository at https://github.com/patuchen/nlg_metric_usage.

Acknowledgements

This work was co-funded by the European Union (ERC, NG-NLG, 101039303) and Charles University projects GAUK 40222 and SVV 260 698.

References

- Eunice Akani, Benoit Favre, Frederic Bechet, and Romain Gemignani. 2023. [Reducing named entity hallucination risk to ensure faithful summary generation](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 437–442, Prague, Czechia. Association for Computational Linguistics.
- Mina Almasi and Anton Schiønning. 2023. [Fine-tuning GPT-3 for synthetic Danish news generation](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 54–68, Prague, Czechia. Association for Computational Linguistics.
- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. [Evaluation methodologies in automatic question generation 2013-2018](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Miriam Anshütz, Diego Miguel Lozano, and Georg Groh. 2023. [This is not correct! negation-aware evaluation of language generation systems](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 163–175, Prague, Czechia. Association for Computational Linguistics.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Anja Belz and Albert Gatt. 2008. [Intrinsic vs. extrinsic evaluation measures for referring expression generation](#). In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200, Columbus, Ohio. Association for Computational Linguistics.
- Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021a. [A systematic review of reproducibility research in natural language processing](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.
- Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021b. [A systematic review of reproducibility research in natural language processing](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.
- Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Kraemer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023a. [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, Dubrovnik, Croatia. Association for Computational Linguistics.
- Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023b. [Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.
- Prabín Bhandari and Hannah Brennan. 2023. [Trustworthiness of children stories generated by large language models](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 352–361, Prague, Czechia. Association for Computational Linguistics.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly, Beijing.

- Andrew P. Bradley. 1997. [The use of the area under the roc curve in the evaluation of machine learning algorithms](#). *Pattern Recognition*, 30(7):1145–1159.
- Michele Cafagna, Kees van Deemter, and Albert Gatt. 2023. [HL dataset: Visually-grounded description of scenes, actions and rationales](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 293–312, Prague, Czechia. Association for Computational Linguistics.
- Nitay Calderon, Subhabrata Mukherjee, Roi Reichart, and Amir Kantor. 2023. [A systematic study of knowledge distillation for natural language generation with pseudo-target training](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14632–14659, Toronto, Canada. Association for Computational Linguistics.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.
- Haw-Shiuan Chang, Zonghai Yao, Alolika Gon, Hong Yu, and Andrew McCallum. 2023. [Revisiting the architectures like pointer networks to efficiently improve the next word distribution, summarization factuality, and beyond](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12707–12730, Toronto, Canada. Association for Computational Linguistics.
- John Chung, Ece Kamar, and Saleema Amershi. 2023. [Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.
- Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. [Context-aware document simplification](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.
- Bhuvan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. [Handling divergent reference texts when evaluating table-to-text generation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.
- George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research, HLT '02*, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Rotem Dror and Roi Reichart. 2018. The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, pages 1383–1392.
- Venkatesh E, Kaushal Maurya, Deepak Kumar, and Maunendra Sankar Desarkar. 2023. [DivHSK: Diverse headline generation using self-attention based keyword selection](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1879–1891, Toronto, Canada. Association for Computational Linguistics.
- Yuxi Feng, Xiaoyuan Yi, Xiting Wang, Laks Lakshmanan, V.S., and Xing Xie. 2023. [DuNST: Dual noisy self training for semi-supervised controllable text generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8760–8785, Toronto, Canada. Association for Computational Linguistics.
- Rudolph Flesch. 1948. [A new readability yardstick](#). *Journal of Applied Psychology*, 32(3):p221 – 233.
- Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. [RARR: Researching and revising what language models say, using language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.
- Nicolas Garneau and Luc Lamontagne. 2023. [Guided beam search to improve generalization in low-resource data-to-text generation](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 1–14, Prague, Czechia. Association for Computational Linguistics.
- Sebastian Gehrmann, Elizabeth Clark, and Thibault Selam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.
- Dimitra Gkatzia and Saad Mahamood. 2015. [A snapshot of NLG evaluation practices 2005 - 2014](#). In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60, Brighton, UK. Association for Computational Linguistics.
- Max Grusky. 2023. [Rogue scores](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934, Toronto, Canada. Association for Computational Linguistics.
- Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. 2023. [Controllable text generation via probability density estimation in the latent space](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12590–12616, Toronto, Canada. Association for Computational Linguistics.

- Jingxuan Han, Quan Wang, Licheng Zhang, Weidong Chen, Yan Song, and Zhendong Mao. 2023a. [Text style transfer with contrastive transfer pattern mining](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7914–7927, Toronto, Canada. Association for Computational Linguistics.
- Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 2023b. [SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11575–11596, Toronto, Canada. Association for Computational Linguistics.
- Qianyu He, Yikai Zhang, Jiaqing Liang, Yuncheng Huang, Yanghua Xiao, and Yunwen Chen. 2023a. [HAUSER: Towards holistic and automatic evaluation of simile generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12557–12572, Toronto, Canada. Association for Computational Linguistics.
- Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023b. [On the blind spots of model-based evaluation metrics for text generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.
- Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2023c. [DiffusionBERT: Improving generative masked language models with diffusion models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4521–4534, Toronto, Canada. Association for Computational Linguistics.
- Eran Hirsch, Valentina Pyatkin, Ruben Wolhandler, Avi Caciularu, Asi Shefer, and Ido Dagan. 2023. [Revisiting sentence union generation as a testbed for text consolidation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7038–7058, Toronto, Canada. Association for Computational Linguistics.
- 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. 2020a. [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.
- 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. 2020b. [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.
- Chieh-Yang Huang, Ting-Yao Hsu, Ryan Rossi, Ani Nenkova, Sungchul Kim, Gromit Yeuk-Yin Chan, Eunye Koh, C Lee Giles, and Ting-Hao Huang. 2023a. [Summaries as captions: Generating figure captions for scientific documents with automated text summarization](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 80–92, Prague, Czechia. Association for Computational Linguistics.
- Fei Huang, Pei Ke, and Minlie Huang. 2023b. [Directed acyclic transformer pre-training for high-quality non-autoregressive text generation](#). *Transactions of the Association for Computational Linguistics*, 11:941–959.
- Xuancheng Huang, Zijun Liu, Peng Li, Tao Li, Maosong Sun, and Yang Liu. 2023c. [An extensible plug-and-play method for multi-aspect controllable text generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15233–15256, Toronto, Canada. Association for Computational Linguistics.
- Dae Yon Hwang, Yaroslav Nechaev, Cyprien de Lichy, and Renxian Zhang. 2023. [GAN-LM: Generative adversarial network using language models for downstream applications](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 69–79, Prague, Czechia. Association for Computational Linguistics.
- Mika Härmäläinen and Khalid Alnajjar. 2021. [Human Evaluation of Creative NLG Systems: An Interdisciplinary Survey on Recent Papers](#). In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 84–95, Online. Association for Computational Linguistics.
- Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. [Preventing generation of verbatim memorization in language models gives a false sense of privacy](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia. Association for Computational Linguistics.
- Qi Jia, Yizhu Liu, Haifeng Tang, and Kenny Zhu. 2023a. [In-sample curriculum learning by sequence completion for natural language generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11937–11950, Toronto, Canada. Association for Computational Linguistics.

- Qi Jia, Haifeng Tang, and Kenny Zhu. 2023b. [Reducing sensitivity on speaker names for text generation from dialogues](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2058–2073, Toronto, Canada. Association for Computational Linguistics.
- Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. [LLM-blender: Ensembling large language models with pairwise ranking and generative fusion](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.
- Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. 2023. [Multi-source semantic graph-based multimodal sarcasm explanation generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11349–11361, Toronto, Canada. Association for Computational Linguistics.
- Yining Juan, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. [Generating multiple questions from presentation transcripts: A pilot study on earnings conference calls](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 449–454, Prague, Czechia. Association for Computational Linguistics.
- Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. [NU-BIA: NeUral based interchangeability assessor for text generation](#). In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.
- Pei Ke, Fei Huang, Fei Mi, Yasheng Wang, Qun Liu, Xi-aoyan Zhu, and Minlie Huang. 2023. [DecompEval: Evaluating generated texts as unsupervised decomposed question answering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9676–9691, Toronto, Canada. Association for Computational Linguistics.
- Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xi-aoyan Zhu, and Minlie Huang. 2022. [CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319, Dublin, Ireland. Association for Computational Linguistics.
- Minbeom Kim, Hwanhee Lee, Kang Min Yoo, Joon-suk Park, Hwaran Lee, and Kyomin Jung. 2023. [Critic-guided decoding for controlled text generation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4598–4612, Toronto, Canada. Association for Computational Linguistics.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. [Evaluating the factual consistency of abstractive text summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.
- Vaibhav Kumar, Hana Koorehdavoudi, Masud Mosh-taghi, Amita Misra, Ankit Chadha, and Emilio Ferrara. 2023. [Controlled text generation with hidden representation transformations](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9440–9455, Toronto, Canada. Association for Computational Linguistics.
- Alexander Hanbo Li, Mingyue Shang, Evangelia Spiliopoulou, Jie Ma, Patrick Ng, Zhiguo Wang, Bonan Min, William Yang Wang, Kathleen McKeown, Vittorio Castelli, Dan Roth, and Bing Xiang. 2023a. [Few-shot data-to-text generation via unified representation and multi-source learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16171–16189, Toronto, Canada. Association for Computational Linguistics.
- Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2023b. [Language modeling with latent situations](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12556–12571, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Liang Li, Ruiying Geng, Chengyang Fang, Bing Li, Can Ma, Rongyu Cao, Binhua Li, Fei Huang, and Yongbin Li. 2023c. [CATS: A pragmatic Chinese answer-to-sequence dataset with large scale and high quality](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2983–3000, Toronto, Canada. Association for Computational Linguistics.
- Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023d. [Contrastive decoding: Open-ended text generation as optimization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.
- Yafu Li, Leyang Cui, Jianhao Yan, Yongjing Yin, Wei Bi, Shuming Shi, and Yue Zhang. 2023e. [Explicit syntactic guidance for neural text generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14095–14112, Toronto, Canada. Association for Computational Linguistics.

- Xiaobo Liang, Zecheng Tang, Juntao Li, and Min Zhang. 2023. [Open-ended long text generation via masked language modeling](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 223–241, Toronto, Canada. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Xin Liu, Muhammad Khalifa, and Lu Wang. 2023a. [BOLT: Fast energy-based controlled text generation with tunable biases](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 186–200, Toronto, Canada. Association for Computational Linguistics.
- Ye Liu, Stefan Ultes, Wolfgang Minker, and Wolfgang Maier. 2023b. [System-initiated transitions from chat to task-oriented dialogues with transition info extractor and transition sentence generator](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 279–292, Prague, Czechia. Association for Computational Linguistics.
- Tyler Loakman, Chen Tang, and Chenghua Lin. 2023. [TwistList: Resources and baselines for tongue twister generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–589, Toronto, Canada. Association for Computational Linguistics.
- Congda Ma, Tianyu Zhao, Makoto Shing, Kei Sawada, and Manabu Okumura. 2023. [Focused prefix tuning for controllable text generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1116–1127, Toronto, Canada. Association for Computational Linguistics.
- Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. [LENS: A learnable evaluation metric for text simplification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.
- Laura Mascarell, Ribin Chalumattu, and Julien Heitmann. 2023. [Entropy-based sampling for abstractive multi-document summarization in low-resource settings](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 123–133, Prague, Czechia. Association for Computational Linguistics.
- R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. [How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN](#). *Transactions of the Association for Computational Linguistics*, 11:652–670.
- Clara Meister, Tiago Pimentel, Luca Malagutti, Ethan Wilcox, and Ryan Cotterell. 2023a. [On the efficacy of sampling adapters](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1455, Toronto, Canada. Association for Computational Linguistics.
- Clara Meister, Wojciech Stokowiec, Tiago Pimentel, Lei Yu, Laura Rimell, and Adhiguna Kuncoro. 2023b. [A natural bias for language generation models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 243–255, Toronto, Canada. Association for Computational Linguistics.
- Yarik Menchaca Resendiz and Roman Klinger. 2023. [Affective natural language generation of event descriptions through fine-grained appraisal conditions](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 375–387, Prague, Czechia. Association for Computational Linguistics.
- Margot Mieskes, Karën Fort, Aurélie Névéol, Cyril Grouin, and Kevin B Cohen. 2019. NLP community perspectives on replicability. In *Recent Advances in Natural Language Processing*.
- Sourabrata Mukherjee and Ondrej Dusek. 2023. [Leveraging low-resource parallel data for text style transfer](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 388–395, Prague, Czechia. Association for Computational Linguistics.
- Sharan Narasimhan, Pooja H, Suvodip Dey, and Maunendra Sankar Desarkar. 2023. [On text style transfer via style-aware masked language models](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 362–374, Prague, Czechia. Association for Computational Linguistics.
- Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. [Conditional generation with a question-answering blueprint](#). *Transactions of the Association for Computational Linguistics*, 11:974–996.
- Piotr Nawrot, Jan Chorowski, Adrian Lancucki, and Edoardo Maria Ponti. 2023. [Efficient transformers with dynamic token pooling](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6403–6417, Toronto, Canada. Association for Computational Linguistics.
- Iftitahu Nimah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. [NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational*

- Linguistics (Volume 1: Long Papers)*, pages 1240–1266, Toronto, Canada. Association for Computational Linguistics.
- Jekaterina Novikova, Ondřej Dušek, Amanda Caracas Curry, and Verena Rieser. 2017. **Why we need new evaluation metrics for NLG**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Rebecca J. Passonneau. 2006. Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 831–836, Genoa, Italy. European Language Resources Association (ELRA).
- Jonathan Pei, Kevin Yang, and Dan Klein. 2023. **PREADD: Prefix-adaptive decoding for controlled text generation**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10018–10037, Toronto, Canada. Association for Computational Linguistics.
- Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *NeurIPS*.
- Maja Popović. 2015. **chrF: character n-gram F-score for automatic MT evaluation**. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. **A call for clarity in reporting BLEU scores**. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Tao Qian, Fan Lou, Jiatong Shi, Yuning Wu, Shuai Guo, Xiang Yin, and Qin Jin. 2023. **UniLG: A unified structure-aware framework for lyrics generation**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 983–1001, Toronto, Canada. Association for Computational Linguistics.
- Ehud Reiter. 2018. **A structured review of the validity of BLEU**. *Computational Linguistics*, 44(3):393–401.
- Ehud Reiter and Anja Belz. 2009. **An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems**. *Computational Linguistics*, 35(4):529–558.
- Fahime Same, Guanyi Chen, and Kees van Deemter. 2023. **Models of reference production: How do they withstand the test of time?** In *Proceedings of the 16th International Natural Language Generation Conference*, pages 93–105, Prague, Czechia. Association for Computational Linguistics.
- Yuichi Sasazawa, Terufumi Morishita, Hiroaki Ozaki, Osamu Imaichi, and Yasuhiro Sogawa. 2023. **Controlling keywords and their positions in text generation**. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 407–413, Prague, Czechia. Association for Computational Linguistics.
- Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. **Do massively pretrained language models make better storytellers?** In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China. 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*. Association for Computational Linguistics.
- Xu Sheng, Fumiyo Fukumoto, Jiye Li, Go Kentaro, and Yoshimi Suzuki. 2023. **Learning disentangled meaning and style representations for positive text reframing**. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 424–430, Prague, Czechia. Association for Computational Linguistics.
- Anastasia Shimorina and Anya Belz. 2022a. **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, Dublin, Ireland. Association for Computational Linguistics.
- Anastasia Shimorina and Anya Belz. 2022b. **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, Dublin, Ireland. Association for Computational Linguistics.
- Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. **Distilling reasoning capabilities into smaller language models**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.
- Judith Sieker, Oliver Bott, Torgrim Solstad, and Sina Zarrieß. 2023. **Beyond the bias: Unveiling the quality of implicit causality prompt continuations in language models**. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 206–220, Prague, Czechia. Association for Computational Linguistics.

- Gabriella Skitalinskaya, Maximilian Spliethöver, and Henning Wachsmuth. 2023. [Claim optimization in computational argumentation](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 134–152, Prague, Czechia. Association for Computational Linguistics.
- Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. [Teaching the pre-trained model to generate simple texts for text simplification](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9345–9355, Toronto, Canada. Association for Computational Linguistics.
- Shiv Surya, Yohan Jo, Arijit Biswas, and Alexandros Potamianos. 2023. [A zero-shot approach for multi-user task-oriented dialog generation](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 196–205, Prague, Czechia. Association for Computational Linguistics.
- Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023a. [Enhancing dialogue generation via dynamic graph knowledge aggregation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4604–4616, Toronto, Canada. Association for Computational Linguistics.
- Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. [MVP: Multi-task supervised pre-training for natural language generation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8758–8794, Toronto, Canada. Association for Computational Linguistics.
- Craig Thomson, Clement Rebuffel, Ehud Reiter, Laure Soulier, Somayajulu Sripada, and Patrick Gallinari. 2023. [Enhancing factualness and controllability of data-to-text generation via data views and constraints](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 221–236, Prague, Czechia. Association for Computational Linguistics.
- Craig Thomson and Ehud Reiter. 2020. [A gold standard methodology for evaluating accuracy in data-to-text systems](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.
- Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Yiwen Chen, Tagyoung Chung, Jing Huang, and Nanyun Peng. 2023. [Unsupervised melody-to-lyrics generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9235–9254, Toronto, Canada. Association for Computational Linguistics.
- Jan Trienes, Paul Youssef, Jörg Schlötterer, and Christin Seifert. 2023. [Guidance in radiology report summarization: An empirical evaluation and error analysis](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 176–195, Prague, Czechia. Association for Computational Linguistics.
- Chris van der Lee, Thiago Castro Ferreira, Chris Emmery, Travis J. Wiltshire, and Emiel Kraemer. 2023. [Neural data-to-text generation based on small datasets: Comparing the added value of two semi-supervised learning approaches on top of a large language model](#). *Computational Linguistics*, pages 555–611.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Kraemer. 2021. [Human evaluation of automatically generated text: Current trends and best practice guidelines](#). *Computer Speech & Language*, 67:101151.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Kraemer. 2021. [Human evaluation of automatically generated text: Current trends and best practice guidelines](#). *Computer Speech & Language*, 67:101–151.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Kraemer. 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.
- Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021a. [Underreporting of errors in NLG output, and what to do about it](#). In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021b. [Underreporting of errors in NLG output, and what to do about it](#). In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Stephanie Schoch, Craig Thomson, and Luou Wen. 2023. [Barriers and enabling factors for error analysis in nlg research](#). *Northern European Journal of Language Technology*, 9(1).
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. [Cider: Consensus-based image description evaluation](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Sujian Li, and Yajuan Lyu. 2023a. [WeCheck: Strong factual consistency checker via weakly supervised learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 307–321, Toronto, Canada. Association for Computational Linguistics.
- Yiquan Wu, Weiming Lu, Yating Zhang, Adam Jatowt, Jun Feng, Changlong Sun, Fei Wu, and Kun Kuang. 2023b. [Focus-aware response generation in inquiry conversation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12585–12599, Toronto, Canada. Association for Computational Linguistics.
- Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. [The next chapter: A study of large language models in storytelling](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics.
- Jiacheng Xu, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. 2023a. [Best-k search algorithm for neural text generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12385–12401, Toronto, Canada. Association for Computational Linguistics.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. [Optimizing statistical machine translation for text simplification](#). *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Yi Xu, Shuqian Sheng, Jiexing Qi, Luoyi Fu, Zhouhan Lin, Xinbing Wang, and Chenghu Zhou. 2023b. [Unsupervised graph-text mutual conversion with a unified pretrained language model](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5130–5144, Toronto, Canada. Association for Computational Linguistics.
- Dingyi Yang and Qin Jin. 2023. [Attractive storyteller: Stylized visual storytelling with unpaired text](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11053–11066, Toronto, Canada. Association for Computational Linguistics.
- Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Xiangpeng Wei, Zhengyuan Liu, and Jun Xie. 2023a. [Fantastic expressions and where to find them: Chinese simile generation with multiple constraints](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 468–486, Toronto, Canada. Association for Computational Linguistics.
- Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2023b. [Tailor: A soft-prompt-based approach to attribute-based controlled text generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–427, Toronto, Canada. Association for Computational Linguistics.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [BartScore: Evaluating generated text as text generation](#). *Advances in Neural Information Processing Systems*, 34:27263–27277.
- Rohola Zandie, Diwanshu Shekhar, and Mohammad Mahoor. 2023. [COGEN: Abductive commonsense language generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 295–302, Toronto, Canada. Association for Computational Linguistics.
- Weihao Zeng, Lulu Zhao, Keqing He, Ruotong Geng, Jingang Wang, Wei Wu, and Weiran Xu. 2023. [Seen to unseen: Exploring compositional generalization of multi-attribute controllable dialogue generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14179–14196, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. [Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications](#). In

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 314–324, Seattle, United States. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texus: A benchmarking platform for text generation models. *SIGIR*.

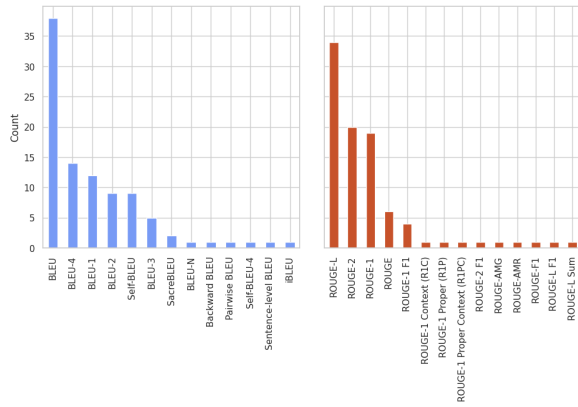


Figure 7: BLEU and ROUGE variant counts across INLG and ACL papers

Appendices

A Additional Results

A.1 BLEU and ROUGE Variants

Figure 7 shows the distribution of the different variants of BLEU and ROUGE respectively used by researchers across both INLG and ACL papers.

A.2 Evaluation Rationales

Figure 8 provides a granular view of the number of metrics per paper against the rationale type given. We can see that correlation with human judgment is only used as a rationale when there are less metrics (2-4). Furthermore, if authors use 9 or more metrics, they rarely provide some insight into why the metrics were chosen.

A.3 Metric Category Co-occurrences

Figure 9 supports the finding that Overlap metrics are generally used with another type of metric.

B List of NLG Tasks

The following is the list of NLG (sub-)tasks commonly mentioned in the annotated papers. Annotators were also able to note tasks not in this list.

- aggregation
- compression / lossy simplification
- content ordering/structuring
- content selection/determination
- data-to-text generation
- deep generation (DLR to text)
- dialogue turn generation
- end-to-end text generation
- feature-controlled generation
- lexicalisation
- machine translation

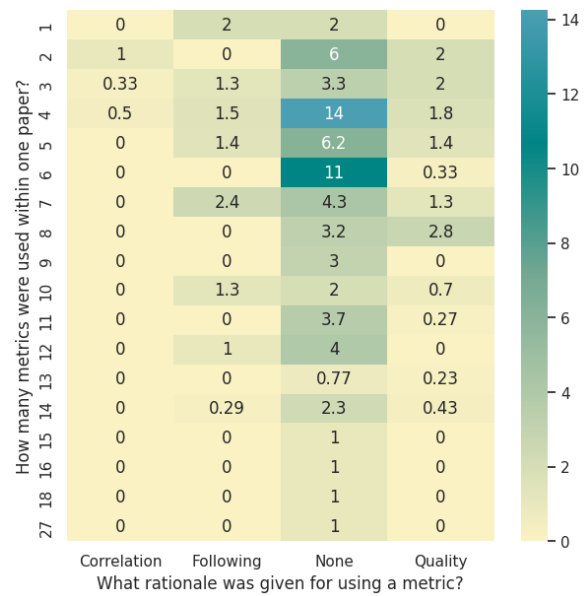


Figure 8: Number of metrics per paper against the rationale type given. If a paper provided more than one type of rationale, its contribution was proportionally divided into more categories.

- paraphrasing / lossless simplification
- question answering
- question generation
- referring expression generation
- summarisation (text-to-text)
- surface realisation (SLR to text)

The following tasks were added during the annotation:

- story generation
- language model sampling
- song lyric generation
- commonsense reasoning

C Evaluation Metrics Used in the Annotated Papers

In this section, we present all of the metrics we encountered during our annotation process. We assigned a family (fine-grained) and a category (high-level) to each metric to increase the clarity of presented results. In some cases, e.g. for ‘Combination’, family and type are identical. Similarly, if a metric is prevalent, it can be in its own singleton family.

C.1 Combination

Multiple metrics in a simple (e.g. mean) or trained combination.

- AUC (Bradley, 1997)
- Average (Gu et al., 2023)

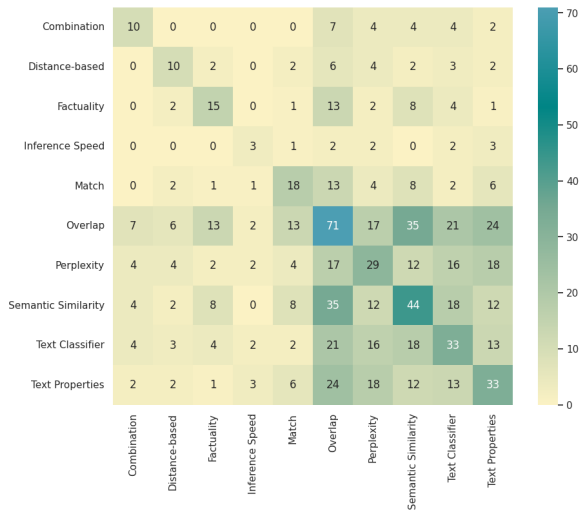


Figure 9: Co-occurrence of metric categories within papers.

- Average of ROUGE-1, ROUGE-2, and ROUGE-L (Calderon et al., 2023)
- BLEU area under curve (Meister et al., 2023b)
- G-score (Han et al., 2023a)
- GeomMean(.) (Yang and Jin, 2023)
- GeomMean(Acc,Sim,Fl) (Jia et al., 2023a)
- Harmonic Mean of Pairwise BLEU and BLEU (E et al., 2023)
- HAUSER Quality (He et al., 2023a)
- J(Acc,Sim,Fl) (Jia et al., 2023a)

C.2 Distance-based

Metrics that measure the distance between two distributions or sequences.

C.2.1 Distribution Comparison

Metrics that measure the distance between two distributions.

- Forward KL divergence of learned distribution (Meister et al., 2023a)
- Jensen-Shannon divergence of learned distribution (Meister et al., 2023a)
- Reverse cross-entropy of learned distribution (Meister et al., 2023a)
- Reverse KL divergence of learned distribution (Meister et al., 2023a)
- Total variation distance of learned distribution (Meister et al., 2023a)
- Weighted macro-F1 (Meister et al., 2023a)
- Zipf’s Coefficient (Han et al., 2023b)

C.2.2 Edit Distance

Metrics that measure the edit distance between two sequences.

- D_{lex} (Li et al., 2023e)
- D_{syn} (Li et al., 2023e)
- Edit Distance (Ippolito et al., 2023)
- $Pres_{COMB}$ (Gao et al., 2023)
- TER (Li et al., 2023a; Zandie et al., 2023)

C.2.3 Loss/Error

Metrics that measure the loss or error between the generated output and a gold reference.

- Agreement - the number of questions generated by GPT-2 (#Q) matches the number of GPT-3 annotated questions for a given problem (Shridhar et al., 2023)
- Bias (Pei et al., 2023)
- Cropped sentence ratio (Tian et al., 2023)

C.3 Factuality (Category)

Metrics that either directly or indirectly aim to measure factuality.

C.3.1 Factuality (Family)

Metrics that either directly aim to measure factuality.

- AlignScore (Zha et al., 2023)
- CheXpert factuality (Trienes et al., 2023)
- Content Selection (Thomson et al., 2023)
- DecompEval (Ke et al., 2023)
- FactCC (Kryscinski et al., 2020)
- FEQA (Zha et al., 2023)
- NEHR (Akani et al., 2023)
- NER Overlap (Zha et al., 2023)
- Q^2 (Wu et al., 2023a)
- QAFactEval (Zha et al., 2023; Wu et al., 2023a)
- QuestEval (Zha et al., 2023; Wu et al., 2023a)
- Relation Generation (Thomson et al., 2023)
- WeCheck (Wu et al., 2023a)

C.3.2 NLI

Classifiers into three classes: logical entailment, contradiction, and neutrality.

- ANLI (Wu et al., 2023a; Narayan et al., 2023)
- $Attr_{AUTO}$ (Gao et al., 2023)
- DeBERTaxxlargev2 (Hirsch et al., 2023)
- NLI (Garneau and Lamontagne, 2023; Li et al., 2023a)
- NLI-warmup (Wu et al., 2023a)
- NUBIA Agreement (Kane et al., 2020)
- NUBIA Contradiction (Kane et al., 2020)
- NUBIA Neutrality (Kane et al., 2020)
- P-NLI (Zeng et al., 2023)
- SUMMAC (Wu et al., 2023a)

C.4 Inference Speed

Metrics that measure the inference speed of a model.

- Inference Time (Kumar et al., 2023)
- Latency (Huang et al., 2023b)
- Speed (token per s) (Liu et al., 2023a)
- Throughput (Huang et al., 2023b)

C.5 Match

Metrics that measure the match between a generated output and a gold label.

C.5.1 Accuracy

Metrics that measure accuracy.

- Accuracy
- Accuracy of comparator (Yang et al., 2023a)
- Accuracy of keyword inclusion (Sasazawa et al., 2023)
- Accuracy of keyword inclusion at a specified position (Sasazawa et al., 2023)
- Accuracy of vehicle (Yang et al., 2023a)
- Completion Sensitivity Score (Sieker et al., 2023)
- Domain Accuracy (Liu et al., 2023b)
- Domain Slot Value Accuracy (Liu et al., 2023b)
- Exact Match (Tang et al., 2023b)
- Exact Match Accuracy (Skitalinskaya et al., 2023)
- Inform (Tang et al., 2023b)
- Proportion of sentences with comparator words (Yang et al., 2023a)
- Stress-duration alignment (Tian et al., 2023)
- Success (Tang et al., 2023b)
- Transition Accuracy (Liu et al., 2023b)

C.5.2 F1

Metrics that measure F1.

- F1
- F1 (Lexical Simplification) (Sun et al., 2023)
- F1-score (appraisal) (Menchaca Resendiz and Klinger, 2023)
- Format F1 (Qian et al., 2023)
- Knowledge-F1 (Huang et al., 2023b)
- macro-F1 (Same et al., 2023; Feng et al., 2023)
- micro-F1 (Xu et al., 2023b)
- QA-F1 (informativeness/grounding) (Narayan et al., 2023)
- weighted macro-F1 (Same et al., 2023)

C.5.3 Precision

Metrics that measure precision.

- Knowledge-Precision (Huang et al., 2023b)
- Precision (Same et al., 2023)
- Precision (Lexical Simplification) (Sun et al., 2023)

C.5.4 Recall

Metrics that measure recall.

- Knowledge-Recall (Huang et al., 2023b)
- Local Recall (van der Lee et al., 2023)
- Recall (Same et al., 2023; Li et al., 2023b)
- Recall (Lexical Simplification) (Sun et al., 2023)
- Recall@N (Hwang et al., 2023)

C.6 Overlap (Category)

Metrics that measure the overlap between two sequences.

C.6.1 BLEU

Multiple variants of the BLEU score (Papineni et al., 2002).

- Backward BLEU (Xie et al., 2023)
- BLEU (Papineni et al., 2002)
- BLEU-1 (Papineni et al., 2002)
- BLEU-2 (Papineni et al., 2002)
- BLEU-3 (Papineni et al., 2002)
- BLEU-4 (Papineni et al., 2002)
- BLEU-N (Wu et al., 2023b)
- iBLEU (Li et al., 2023e)
- Pairwise BLEU (E et al., 2023)
- SacreBLEU (Post, 2018)
- Self-BLEU (between source and target) (Zhu et al., 2018)
- Self-BLEU (between more system-generated outputs) (Zhu et al., 2018)
- Self-BLEU-4 (He et al., 2023c)
- Sentence-level BLEU (Tian et al., 2023)

C.6.2 chrF++

This family consists solely of the chrF++ metric (Popović, 2015).

C.6.3 CIDEr

This family consists solely of the CIDEr metric (Vedantam et al., 2015).

C.6.4 METEOR

This family consists solely of the METEOR metric (Banerjee and Lavie, 2005).

C.6.5 NIST

Multiple variants of the NIST metric (Doddington, 2002).

- NIST (Doddington, 2002)
- NIST-1 (Tang et al., 2023a)
- NIST-2 (Tang et al., 2023a)
- NIST-3 (Tang et al., 2023a)
- NIST-4 (Tang et al., 2023a)

C.6.6 Overlap (family)

Metrics that measure the overlap between two sequences.

- Add (Sun et al., 2023)
- Copy Success Rate (word) (Huang et al., 2023c)
- Coverage (van der Lee et al., 2023; Li et al., 2023c)
- Coverage (of keywords) (Liu et al., 2023a)
- D-add (Sun et al., 2023)
- D-delete (Sun et al., 2023)
- Delete (Sun et al., 2023)
- Dkeep (Sun et al., 2023)
- Extractive fragment density (ρ) (Mascarell et al., 2023)
- HAUSER Creativity (He et al., 2023a)
- Keep (Sun et al., 2023)
- MS-Jaccard (Xie et al., 2023)
- Phonetic Overlap (Loakman et al., 2023)
- Proper Noun Ratio (P Ratio) (Chang et al., 2023)
- Salient word coverage (Tian et al., 2023)
- Slot Coverage (Surya et al., 2023)
- SMART (Cripwell et al., 2023)
- Weisfeiler Lehman graph hash (Bhandari and Brennan, 2023)

C.6.7 PARENT

Multiple scores produced by the PARENT metric (Dhingra et al., 2019).

- PARENT(Dhingra et al., 2019)
- PARENT-T-F1 (Huang et al., 2023b)
- PARENT-T-Precision (Huang et al., 2023b)
- PARENT-T-Recall (Huang et al., 2023b)

C.6.8 ROUGE

Multiple variants of the ROUGE score (Lin, 2004).

- ROUGE (Lin, 2004)
- ROUGE-1 (Lin, 2004)
- ROUGE-1 Context (RIC) (Chang et al., 2023)
- ROUGE-1 F1 (Chang et al., 2023)
- ROUGE-1 Proper (R1P) (Chang et al., 2023)

- ROUGE-1 Proper Context (R1PC) (Chang et al., 2023)
- ROUGE-2 (Lin, 2004)
- ROUGE-2 F1 (Jia et al., 2023b)
- ROUGE-AMG (Juan et al., 2023)
- ROUGE-AMR (Juan et al., 2023)
- ROUGE-F1 (Huang et al., 2023a)
- ROUGE-L (Lin, 2004)
- ROUGE-L F1 (Jia et al., 2023b)
- ROUGE-L Sum (Narayan et al., 2023)

C.6.9 SARI

Two scores produced by the SARI metric (Xu et al., 2016).

- DSARI (Xu et al., 2016)
- SARI (Xu et al., 2016)

C.7 Perplexity (Category)

Metrics that directly or indirectly measure perplexity.

C.7.1 MAUVE

MAUVE metric (Pillutla et al., 2021) with various underlying language models.

- MAUVE (Pillutla et al., 2021)
- MAUVE (ELECTRA-large) (He et al., 2023b)
- MAUVE (GPT2-large) (He et al., 2023b)
- MAUVE (RoBERTa-large) (He et al., 2023b)

C.7.2 Perplexity (family)

Metrics that directly measure perplexity.

- Bits per character (BPC) (Nawrot et al., 2023)
- Fluency (Pei et al., 2023)
- Fluency (Perplexity) (Yang and Jin, 2023)
- GPT-PPL (He et al., 2023b)
- MLM-PPL (He et al., 2023b)
- Model PPL (Feng et al., 2023)
- Output PPL (Feng et al., 2023)
- Perplexity
- Perplexity (Liang et al., 2023; Tang et al., 2023b)
- Perplexity (Chinese GPT-2) (Yang et al., 2023a)
- Perplexity (GPT-2) (Tian et al., 2023)

C.8 Semantic Similarity (Category)

Metrics that measure semantic similarity.

C.8.1 BARTScore

Multiple scores produced by the BARTScore metric (Yuan et al., 2021).

- BARTScore(Yuan et al., 2021)
- BARTScore faithfulness (He et al., 2023b)

- BARTScore fscore (He et al., 2023b)
- BARTScore precision (He et al., 2023b)
- BARTScore recall (He et al., 2023b)

C.8.2 BERTScore

Multiple scores produced by the BERTScore metric (Zhang et al., 2019).

- BERTScore (Zhang et al., 2019)
- BERTScore F1 (Zhang et al., 2019)
- BERTScore Precision (Zhang et al., 2019)
- BERTScore Recall (Zhang et al., 2019)

C.8.3 MoverScore

This family consists solely of the MoverScore metric (Zhao et al., 2019).

C.8.4 Semantic Similarity (family)

Metrics that directly measure semantic similarity.

- Coherence (Li et al., 2023b,d)
- Cosine Similarity (Chung et al., 2023)
- Embedding Similarity (Mukherjee and Dusek, 2023)
- MPNet Cosine Similarity (Anschütz et al., 2023)
- NegMPNet Cosine Similarity (Anschütz et al., 2023)
- NUBIA Semantic Similarity (Kane et al., 2020)
- P-SIM (Zeng et al., 2023)
- RANK (Garneau and Lamontagne, 2023)
- Relevance (Pei et al., 2023)
- Semantic Similarity (Jia et al., 2023a)
- Sentence-BERT (Surya et al., 2023)
- Sentence-BERT Cosine Similarity (Jing et al., 2023)
- SimCSE (Zha et al., 2023)
- Spearman Rank Correlation (Hwang et al., 2023)
- SR (Semantic Repetition) (Liang et al., 2023)
- Topic modelling (Bhandari and Brennan, 2023)

C.9 Text Classifiers

Type of metrics that classify various properties of the generated text.

C.9.1 BLEURT

Metrics based on BLEURT (Sellam et al., 2020).

- BLEURT (Sellam et al., 2020)
- NegBLEURT (Anschütz et al., 2023)
- Purity Score (Cafagna et al., 2023)

C.9.2 Quality Estimation

Quality estimation metrics for referenceless evaluation. Also includes a small set of classifiers trained to distinguish human-written from machine-generated texts.

- BERT Classification F1 (Almasi and Schiønning, 2023)
- BERT Classification Precision (Almasi and Schiønning, 2023)
- BERT Classification Recall (Almasi and Schiønning, 2023)
- BLANC (Zha et al., 2023)
- COMET-QE (He et al., 2023b)
- CTC (Nimah et al., 2023)
- CTRL Eval (Ke et al., 2022)
- GPTRank (Jiang et al., 2023)
- LR Classification F1 (Almasi and Schiønning, 2023)
- LR Classification Precision (Almasi and Schiønning, 2023)
- LR Classification Recall (Almasi and Schiønning, 2023)
- Naturalness (Narasimhan et al., 2023)
- PRISM-QE (He et al., 2023b)
- USR (Ke et al., 2023)

C.9.3 Style Classifiers

Classifiers that were trained to classify style, sentiment, or topic.

- Accuracy (Sentiment) (Huang et al., 2023c)
- Accuracy (Tense) (Huang et al., 2023c)
- Accuracy (Topic) (Huang et al., 2023c)
- Act - Classification accuracy (A-ACC) Roberta (Zeng et al., 2023)
- Act - Multiple Attribute Evaluation (A-MAE) (Zeng et al., 2023)
- Bias (absolute value of relevance - 50) (Ma et al., 2023)
- C-Ext (Han et al., 2023b)
- Content Ordering (Thomson et al., 2023)
- Correctness (Yang et al., 2023b)
- custom trained relevance classifier (Ma et al., 2023)
- Detoxify (Bhandari and Brennan, 2023)
- Emotion - Classification accuracy (E-ACC) Roberta (Zeng et al., 2023)
- Emotion - Multiple Attribute Evaluation (E-MAE) (Zeng et al., 2023)
- Fluency (Jia et al., 2023a)
- Grammaticality (Kim et al., 2023; Xu et al., 2023a; Yang et al., 2023b)
- Integrity (Qian et al., 2023)

- Intented Sentiment (external classifier) (Liu et al., 2023a)
- Intented Sentiment (internal classifier)(Liu et al., 2023a)
- Label Accuracy (Chung et al., 2023)
- LENS (Maddela et al., 2023)
- Negative Sentiment (Kumar et al., 2023)
- P-Multiple Attribute Evaluation (P-MAE) (Zeng et al., 2023)
- Positiveness (Kim et al., 2023)
- RoBERTa fine-tuned for sentiment (Ma et al., 2023)
- Sentiment (Gu et al., 2023)
- Sentiment Accuracy (Han et al., 2023a; Mukherjee and Dusek, 2023)
- Simile confidence (Yang et al., 2023a)
- Simplicity (Kumar et al., 2023)
- Structure F1 (Qian et al., 2023)
- Style Accuracy (Yang and Jin, 2023; Jia et al., 2023a)
- Style Transfer Accuracy (Narasimhan et al., 2023)
- Success (Pei et al., 2023)
- Topic (Gu et al., 2023)
- Toxicity (Pei et al., 2023; Kim et al., 2023)
- Toxicity (Kumar et al., 2023)
- Δ TextBlob (Sheng et al., 2023)

C.9.4 Unieval

Various scores produced by the Unieval metric (Zhong et al., 2022).

- UniEval (Zhong et al., 2022)
- Unieval - coherence (He et al., 2023b)
- Unieval - consistency (He et al., 2023b)
- Unieval - fluency (He et al., 2023b)
- Unieval - overall (He et al., 2023b)
- Unieval - relevance (He et al., 2023b)
- UniEval (Dial) (Ke et al., 2023)
- UniEval (Summ) (Ke et al., 2023)

C.10 Text Properties

Type of metrics that measure various text properties.

C.10.1 Flesch Readability

Flesch Readability scores.

- Flesch Reading Ease Score (Bhandari and Brennan, 2023)
- Flesch-Kincaid grade level (FKGL) (Flesch, 1948)

C.10.2 N-gram Diversity

N-gram diversity metrics.

- Averaged Distinctiveness (Huang et al., 2023c)
- Bigram TTR (van der Lee et al., 2023)
- Dist-n (Feng et al., 2023)
- Distinct-1 (Li et al., 2016)
- Distinct-2 (Li et al., 2016)
- Distinct-3 (See et al., 2019)
- Distinct-3 (proportion) (Liu et al., 2023a)
- Distinct-4 (Tang et al., 2023b)
- Distinct-n (Surya et al., 2023)
- Distinctness (Gu et al., 2023)
- Div-4 (He et al., 2023c)
- Diversity (Li et al., 2023b,d)
- Diversity (of questions) (Juan et al., 2023)
- Diversity score (Cafagna et al., 2023)
- Diversity-1 (Xu et al., 2023a)
- Diversity-2 (Xu et al., 2023a)
- Diversity-3 (Xu et al., 2023a)
- Ent-4 (Tang et al., 2023a)
- Initial Phonetic Overlap (Loakman et al., 2023)
- Mean segmented type-token ratio (van der Lee et al., 2023)
- n-gram novelty (n from 1-10) (McCoy et al., 2023)
- Novelty (van der Lee et al., 2023)
- Number of types (van der Lee et al., 2023)
- Percentage of novel texts (van der Lee et al., 2023)
- Syntactic Novelty (dependency role) (McCoy et al., 2023)
- Syntactic Novelty (labeled dependency arc) (McCoy et al., 2023)
- Syntactic Novelty (sentence level) (McCoy et al., 2023)
- Unique Sentence Count (Xu et al., 2023a)
- Δ CR (Hirsch et al., 2023)

C.10.3 N-gram Repetition

N-gram repetition metrics.

- 4-gram Repetition (Li et al., 2023d)
- Bigram Repetition (Li et al., 2023d)
- Lexical Repetition (Xie et al., 2023; Liang et al., 2023)
- Repetition rate (Han et al., 2023b)
- Trigram Repetition (Surya et al., 2023; Liu et al., 2023a; Li et al., 2023d)

C.10.4 Sequence Length

Various measures of generated sequence length.

- Average Length (Sheng et al., 2023)
- Average Sentence Length (van der Lee et al.,

2023)

- HAUSER Informativeness (He et al., 2023a)
- Length (Xie et al., 2023)
- Sentence Count (Xu et al., 2023a)
- Sentence Length (Bhandari and Brennan, 2023)
- Shortening Factor (SF) (Nawrot et al., 2023)
- Standard deviation of the sentence length (van der Lee et al., 2023)

D Paper and Code Resources

This section adds further detail to the results discussed in subsection 4.4.

D.1 Code Releases Annotation Procedure

For each paper we annotated with the following procedure:

1. If the paper provides a link to a code or data release.
2. If the link actually contains the release resulting labels *no code*, *delivered*, *missing* (Figure 10).
3. We annotated if the authors come from *Academia* or *Industry*. The mixed authoring teams received the labels *Academia Industry*, *Industry Academia* depending on the first authors, resulting in four labels.
4. We retrieved the GitHub Stars for each release since all except one paper was released on GitHub (Figures 15 and 16).
5. We annotated if the *Installation Instructions* were provided as follows (Figure 11):
 - *None* - no attempt at providing installation instructions seen.
 - *Some* - installation instructions are visible but lack the necessary detail.
 - *Detailed* - clearly states dependencies and exact (minimal) versions so we believe the computational environment can be easily replicated.
6. We checked the clarity of the experiment structure if the experiments mentioned in the paper are *discoverable* (Figure 12).
 - *None*: we have no idea how to start any experiment.
 - *Some*: we easily found how to replicate only the main experiments.

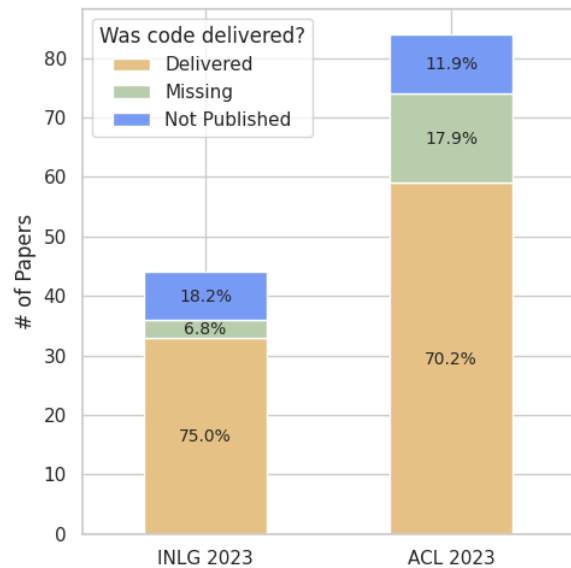


Figure 10: Each paper either did not link any source code (or data) or linked it and delivered or failed to deliver it – ‘missing’.

- *Many*: we found out how to run experiments even for all the ablation groups.

7. We labeled the level of documentation detail with the following (Figure 13):

- *None*: no introduction to the codebase.
- *Basic*: it was clear what the main commands do, including the most important arguments.
- *Detailed*: it was clear what most hyperparameters mean and how one could change them.

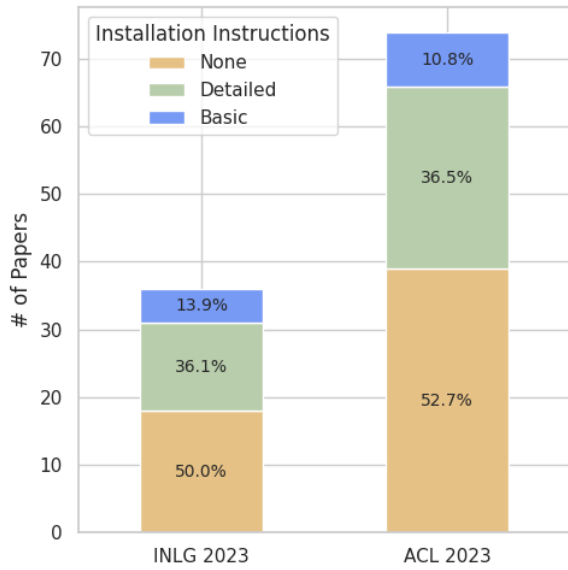


Figure 11: The quality of installation instructions annotated as *None*, *Basic*, *Detailed*.

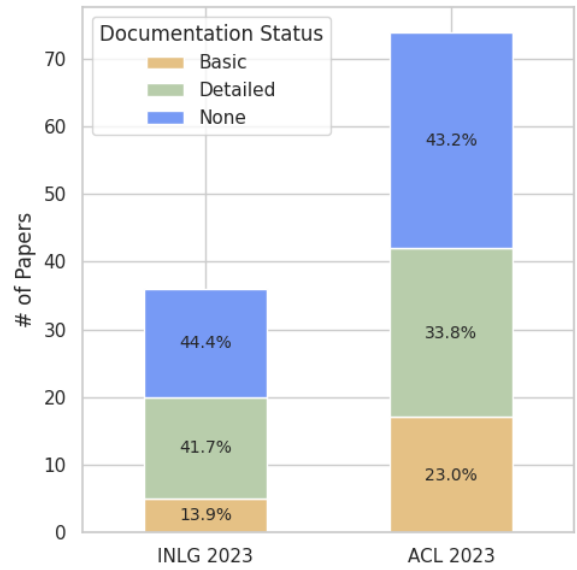


Figure 13: The quality of documentation annotated as *None*, *Basic*, *Detailed*.

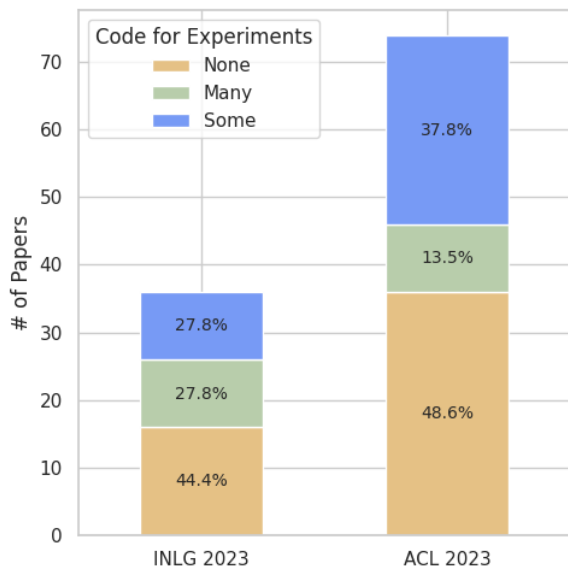


Figure 12: The quality of linking experiments in paper and code annotated as found *None*, *Some*, and *Many*.

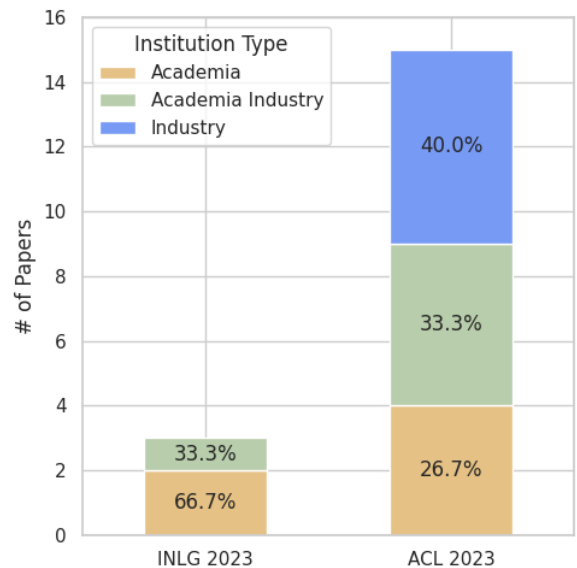


Figure 14: How the teams from academia or industry behind the papers with missing code are represented?

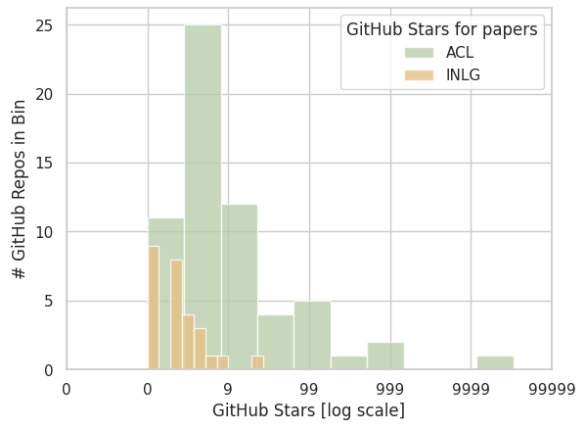


Figure 15: Distribution of GitHub Stars for INLG and ACL papers

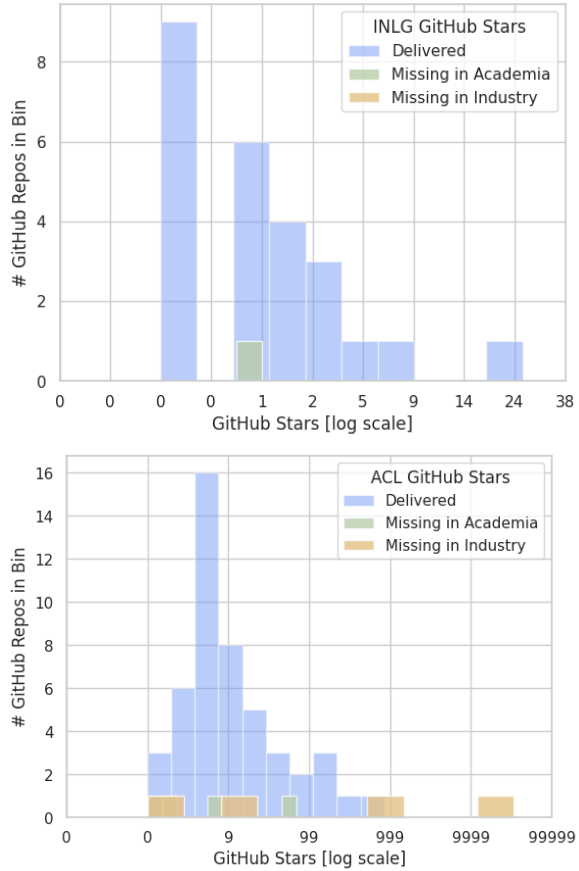


Figure 16: Distribution of the GitHub Stars for ACL for groups with groups missing for Industry and Academia