Mapping Out the NLP Evaluation Landscape with a Standard Taxonomy of Quality Criteria

Anya Belz, Simon Mille and **Craig Thomson**
DCU Natural Language Generation Research Group
ADAPT, Dublin City University, Dublin, Ireland
anya.belz@dcu.ie

Abstract

Prior research shows that when papers report results from system evaluations in terms of a quality criterion such as Fluency, answers to two questions are normally less clear than they should be: (i) was it really Fluency that was evaluated; and (ii) was the same aspect of quality evaluated as in other evaluations also claiming to evaluate Fluency. Answers to these questions are crucial if meaningful conclusions about the Fluency of systems, independently and as compared to others, are to be drawn. We map a combined total of 1,002 individual evaluations identified in three surveys of 310 NLP papers to the standardised QCET inventory of quality criterion names and definitions. Standardisation results in up to 76% reduction in evaluation criteria names, revealing a lot of spurious difference in evaluation naming. We argue that conclusions drawn from NLP system evaluations are only fully interpretable and comparable if grounding in a standard inventory of quality criterion names and definitions forms part of experiment design and reporting, and we propose a way of achieving this.

1 Introduction

Research has consistently shown (van der Lee et al., 2019; Belz et al., 2020; Howcroft et al., 2020; Belz et al., 2025b) that what was actually evaluated in human evaluations is commonly at odds with the name and/or definition of the quality criterion given in a paper. For example, Van de Cruys (2020) reported evaluations for Fluency, but defined it as outputs being "grammatical and syntactically well-formed," which assesses grammaticality rather than fluency.[1] Given this kind of misalignment (which Howcroft et al. (2020) showed to be pervasive in NLP), not only is it problematic to report, say, that a system improves Fluency, but any comparisons with Fluency assessments conducted in other experiments are also unsafe. For human evaluations, the situation is compounded by the fact that few details of evaluation experiments are usually reported beyond the quality criterion name (Belz et al., 2023a,b; Ruan et al., 2024; Schmidtova et al., 2024, 2025). As has been argued (van der Lee et al., 2019; Howcroft et al., 2020; Belz et al., 2020; van der Lee et al., 2021; Gehrmann et al., 2023), this is an not an acceptable state of affairs in particular for human evaluation which has always been treated as the Litmus test of quality in NLP.

It is hard to see how the current misalignments between (i) what is actually evaluated vs. what it is named, and (ii) what different researchers mean by the same quality criterion name, can be addressed other than by a standard reference set of quality criteria names and definitions that those actually in use can be grounded in.[2] We have previously proposed such a standard reference set in the form of the QCET *Q*uality *C*riteria for *E*valuation *T*axonomy (Belz et al., 2025b), derived descriptively from extensive surveys of the NLP literature.

In this paper, we report a validation exercise we carried out following the completion of the QCET Taxonomy which had the threefold purpose of (i) validating the taxonomy in one of its three main deployment modes, i.e. grounding existing evaluations in standard quality criteria for comparability; (ii) creating a large dataset of evaluation experiments mapped to QCET quality criteria to facilitate analyses of the NLP evaluation landscape; and (iii) evaluating the agreement between different users when mapping evaluation experiments to QCET quality criteria, following common guidelines.

We start below by describing the annotation scheme we used in the validation and the basic concepts behind it (Section 2). Next we present the

[1] A look at almost any patent text confirms that grammatically well-formed text is not necessarily fluent.

[2] In a way not dissimilar from grounding in named entity recognition.

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 999–1014

three datasets we annotated and mapped to QCET quality criteria as well as some first analyses based on them (Sections 3–5). We present results from the user-agreement assessment (Section 6), and conclude with some discussion (Section 7).

2 Basic Concepts and Annotation Process

During the development of QCET (Belz et al., to appear), we had used three collections of papers at different stages in different ways to support our efforts to work out what aspects of quality were really being evaluated in the experiments reported in the papers, to name these aspects of quality, and to structure them into a hierarchy that reflected the relationships between them.

This did not however mean that by the time we finished the taxonomy, we had a fully annotated collection of papers mapped to the quality criteria (QCs) in the taxonomy. Creating such a data resource from the three collections of papers (which we call **Survey 1–3** below) was the aim of the validation exercise reported in this and the next three sections, as well as identifying any remaining gaps in the taxonomy in the process.

QCET Taxonomy: The *Q*uality *C*riteria for *E*valuation *T*axonomy (Belz et al., 2025a) consists of an interactive taxonomy browser, an at-a-glance diagram showing node IDs and names only (for ease of reference copied in Appendix C, Figure 1), description and usage guidance.[3]

Quality criteria: QCET, and by implication the work we present, is based on the notion of quality criterion, i.e. the specific aspect of system quality that is assessed in an evaluation (Belz et al., 2025a); see Appendix B for details. It is this basic concept of quality criterion that enables the standardisation, hence grounding of evaluation criterion names found in the literature in standardised labels, to support meaningful reporting and interpretation of results.

Annotation steps and scheme: In **Step 1** we systematically selected papers for review and annotation. This was done slightly differently for Surveys 1 and 2 on the one hand, and Survey 3 on the other, as described at the start of each of Sections 3–5 below. In **Step 2**, evaluation experiments in papers were identified and split by evaluation criterion name and dataset, with a row created in the annotation spreadsheet for each combination. In **Step 3**, the authors annotated, and in the case of Survey 3 *re*annotated, the evaluations identified in Step 2, using the following annotation scheme (for more details see Appendix A; for more on QCET, see next paragraph below):

1.-6. Paper details (auto-filled): **id, assigned_to, year, url, author, title**.

7. **type**: Type of evaluation annotated, one of {*human, metric, hybrid, none found*} where *none found* meant no evaluation of system output quality was found.
8. **location**: Part of paper where information used for 9 and 10 can be found; list of {*Section, Table, Figure, Footnote, Appendix*} *N*; or *N/A* if **type**=*none found*.
9. **verbatim_qc_name**: Name used by authors for what was evaluated, *None given*, or *N/A* if **type**=*none found*. NB not necessarily at the level of granularity of a QCET QC.
10. **verbatim_qc_definition**: Definition or explanation of what was evaluated, *None given* if no definition given, or *N/A* if **type**=*none found*; uses the question put to evaluators where given.
11. **qcet_node**: Standardised QCET quality criterion ID and name if matching node found, *Not found*, if no match found, *MULTIPLE* if multiple QCET QCs are evaluated by the same single evaluation measure, or *N/A* if **type**=*none found*.
12. **external_location_consulted**: Where information provided in paper is not enough to perform mapping in 11, URL(s) of the paper(s) additionally consulted in performing the mapping, or left blank if none consulted.
13. **proposed_qc_name**: If **qcet_node**=*Not found*, proposed ID and name for new QC to be added to QCET. Else, *n/a*.
14. **proposed_qc_definition**: If **qcet_node**=*Not found*, proposed definition for new QC to be added to QCET. Else, *n/a*.
15. **proposed_qcet_branch**: If **qcet_node**=*Not found*, branch ID and name where proposed new QCET node should be inserted. Else, *n/a*.

In **Step 4** (annotation checking), the separation into evaluations from Step 2, and the annotations from Step 3, were checked by a different author. This

[3] https://nlp-qcet.github.io/

Survey	Venues	Papers	Evals	Type of evaluation				Evaluation criterion names			
				Human	Metric	Hybrid	None found	None given	as found	normalised	QCET (new)
1	ACL Main	60	233	39	192	2	(8)	(1)	170	164	43 (8)
2	ACL Main	60	293	47	246	0	(5)	(4)	203	191	48 (14)
3	ENLG, INLG	190	476	453	23	0	(19)	(53)	256	N/A	90 (6)

Table 1: High-level descriptive statistics for the three surveys analysed. None given = number of evaluations where no name was given. Last three columns show unique counts of different criterion names.

Annotators	**EM**	**F1**	**J**
X vs Y	0.842	0.844	0.931
X vs Z	0.700	0.881	0.901
Y vs Z	0.789	0.927	0.922

Table 2: Agreement between pairs of annotators by EM=Exact Match; F1 Score, and J=Jaccard score.

was followed by discussions and updating of the agreed annotation guidelines (see Appendix A), after which each original annotator checked over all their annotations again.

In **Step 5** all three annotators annotated a fresh batch of 19 papers for IAA purposes as described in Section 6 and Table 2.

3 Survey 1: 60 ACL Papers, 2022–2024

The Survey 1 dataset comprises 60 ACL 2022–2024 main conference papers, obtained by downloading the ACL Anthology as BibTex from `https://aclanthology.org` in Feb 2025, extracting all ACL main conference items for 2022, 2023, and 2024, then removing items which were not papers, and randomly sampling 20 papers from each year.

As shown in Table 1, we divided papers into 233 evaluations (Step 1), of which 39 were annotated *human*, 192 *metric*, and 2 *hybrid* in Step 2. In 8 cases, evaluations turned out to be something else (e.g. a wizard-of-oz experiment) and were excluded. There were 170 different evaluation criterion names found (verbatim_qc_name in our annotation scheme, Section 2). Light normalisation using the mapping schema shown in Appendix D resulted in a reduction to 164 names.

There were 8 unique QCs we needed to add to QCET to cover evaluations for which no matching node existed in the taxonomy yet. We mapped evaluations to just 43 different QCET nodes, including the 8 newly added ones. This is a very substantial reduction of 74.7% relative to the verbatim names (73.8% relative to the lightly normalised ones).

The first column of Table 3 shows the 10 QCET quality criteria most frequently mapped to, the second column the number of times they were found, and the last the original names (verbatim_qc_name) they correspond to. The definitions of QCET QCs we mention in this paper can be found in Appendix E. The three most common QCs found were (using QCET names) Classification Accuracy, Similarity to Target Outputs (outputs as a whole), and Complete Target Output Matching. Between them these account for 54.9% of evaluations.

4 Survey 2: 60 ACL Papers, 2022–2024

We obtained the Survey 2 data by randomly sampling another 60 papers, using the same bibtex download and methods as in Section 3, ensuring no overlap. In Step 1, these were divided into 293 evaluations, a higher rate per paper than in Survey 1.

In Step 2, 246 evaluations were annotated *metric*, 47 *human*, none *hybrid,* and in 5 cases, no evaluations were found. We needed to add a total of 14 new QCET nodes (see Appendix E for definitions of QCET QCs) to cover all evaluations from this batch. There were 203 different evaluation criterion names which light normalisation reduced to 191 names. These mapped to 48 different QCET nodes, including the 14 new ones. This is a reduction of 76.4% relative to the verbatim names (74.9% relative to the lightly normalised ones).

Table 4 shows that, as in Survey 1, the three most common QCs were Complete Target Output Matching, Similarity to Target Outputs (outputs as a whole), and Classification Accuracy, this time with reversed frequencies, and accounting for 49.1% of evaluations.

5 Survey 3: 165 INLG papers, 2000–2019

Our third survey is a reannotation of the 20Years Survey (Howcroft et al., 2020), conducted after the above two surveys and associated QCET updates. The 190 papers annotated are the same in the original and our reannotation. However, the evaluations annotated are not, because (i) we reinstated evaluations that were ruled out due to using metrics, or not being NLG, in the original survey; (ii) we corrected errors where papers had been incor-

QCET quality criterion	N	Original names
QTC-w-1:Classification Accuracy	60	accuracy, f1, macro f1, precision, recall, precision re tasks, macro precision, macro recall, micro f1, precision ner tasks, recall ner tasks, precision trig. cls, f1 score with fixed fpr, recall re tasks, recall trig. cls, true positive rate, f1 trig. cls, average accuracy (text classification), classification accuracy, average f1 trig. cls, average accuracy (relation extraction), average accuracy (ner), auprc, auc, area under coverage-f1 curve, accuracy (text), accuracy (image), acc-5, acc-3, acc-2, weighted-average f1
QTG-w-1:Similarity to Target Outputs (outputs as a whole)	48	bleu, rouge-1, rouge-2, rouge-l, sacrebleu, f1, meteor, spearman, cider, bleu-1, matthews correlation coefficient, sentencepiece-bleu, sari, rouge, rlasp, pearson, nist-4, morpheme-level f1, mean token f1, comet, hit, easse (keep), easse (delete), easse (add), bleu-4, bleu-2, accuracy, accuracy vqa-hat-cp, spice
QTC-w-3:Complete Target Output Matching	20	exact match, word error rate, holder f1, task accuracy, target f1, sentiment graph f1, relation strict f1, qa em, non-polarity sentiment graph f1, hits@1, guess, guess (strict), f1, expressions f1, entity f1, answer coverage rate, whole-word accuracy
QOF-w-1:Diversity/Non-diversity (outputs as a whole)	13	c-dist-1, c-dist-2, dis-1, dis-2, dis-3, dist-1, diversity, dist-2, diversity, diversity, ent-1, ent-2, ent-3, s-dist-1, s-dist-2
QEG-w-8:Performance of an Embedding/Downstream System/Component	6	helpful, llmarena+average reward, llmarena+trueskill ratings, none given, spearman correlation on multilingual sts 2017, vwsd performance
MULTIPLE	6	fluency, coh-con.score, fpvg, harmful and relevant, overall score
QIF-w-4:Similarity/Dissimilarity to Input (outputs as a whole)	5	average precision, energy distance, f0 distance, false positive rate, mel-ceptral distortion
QEF-w-1:Similarity/Dissimilarity to Non-target Reference (outputs as a whole)	4	novelty-1, novelty-2, novelty-3, novelty-4
QTG-c-1:Similarity to Target Outputs (content/meaning)	4	bert score, agreement
QTC-w-4:Retrieval Accuracy	4	average f1 trig. id, f1 trig. id, precision trig. id, recall trig. id

Table 3: **Survey 1**: 10 most frequent QCET quality criteria, and the original names they were mapped from.

rectly divided into evaluations; and (iii) we deleted rows where evaluations of manually created text had been included. This resulted in 476 evaluations annotated compared to the original 482.

Note that the annotations for this third survey were not checked as rigorously as for the other two, therefore they are more likely to contain errors.

Of the evaluations, 23 were *metric*, 453 *human*, and in 19 cases, no evaluations were found. The reason for the far fewer metric evaluations was that the 20Years survey was designed as one of human evaluations. The few metric evaluations in our reannotation were due to errors or n/a annotations in the original survey. We added 6 unique new QCET nodes. 23 evaluations mapped to multiple different QCs, and 19 were out of scope.

We found 256 different original evaluation criterion names which we did not normalise, because human evaluations use more shared terms (Fluency, Grammaticality, Adequacy, etc.), and mapped to 90 QCET nodes, including the 6 new ones.

Table 5 shows that in contrast to the first two surveys, there is no distinct set of top most frequent QCs that account for a large majority of all evaluations. The four most common QCs found were (in their standardised versions) Usefulness for Task/Information Need, Understandability, Grammaticality, Clarity, and Fluency. Some of these were mapped from a relatively small number of original names (e.g. Fluency), while others were mapped from a varied array of different original names (e.g. Usefulness for Task/Information Need).

6 Assessing Annotator Agreement

To calculate agreement between annotators, we selected 19 papers from Survey 3 to be annotated by all three annotators. We then compared our annotations using Exact Match (strict correctness), Micro-F1, and Jaccard score (intersection-over-union between annotator label sets). Each is defined in the below equations.

$$\text{EM} = \mathbb{1}\{A_f^{\text{ann}_\text{A}} = A_f^{\text{ann}_\text{B}}\} \tag{1}$$

$$\text{Micro-F1} = \frac{2\left|A_f^{\text{ann}_\text{A}} \cap A_f^{\text{ann}_\text{B}}\right|}{\left|A_f^{\text{ann}_\text{A}}\right| + \left|A_f^{\text{ann}_\text{B}}\right|} \tag{2}$$

$$\text{Jaccard} = \frac{|A_f^{\text{ann}_\text{A}} \cap A_f^{\text{ann}_\text{B}}|}{|A_f^{\text{ann}_\text{A}} \cup A_f^{\text{ann}_\text{B}}|} \tag{3}$$

For agreement scores see Table 2, with annotators pseudonymised to X, Y, and Z. We see strong agreement by all methods, especially between X and Y.

QCET quality criterion	N	Original names
QTC-w-3:Complete Target Output Matching	55	accuracy, exact match, f1, precision, node type recall, link prediction precision, link prediction recall, macro f1, micro f1, node arg_whether accuracy, node missing_be accuracy, node type accuracy, node type precision, none given, link prediction f1, optimal f1, rationale-f1, recall, temporal e-n, temporal s-n, omission errors, graph structure accuracy, hits@1 percentage, f1 (triple), acc. per example, acc. per sent. bound, addition errors, auc, average accuracy, character error rate, correctness - answer accuracy, detection of additions f1, detection of additions precision, detection of additions recall, detection of omissions f1, detection of omissions precision, detection of omissions recall, edge prediction precision, edge prediction recall, entity recall, f1 (entity), word error rate
QTG-w-1:Similarity to Target Outputs (outputs as a whole)	52	bleu, rouge-l, bleu-4, meteor, f1, rouge-1, rouge-2, bleu-2, chrf++, labelled attachment score, unigram f1, sari, ref-bleu, metricx-23xxl, metricx-23xl, message-f1, clas, hamming distance, f1 over rouge-1 and bleu-1, edit distance, comet, cider, bleu-1, "average bleu" (over languages), wsclas
QTC-w-1:Classification Accuracy	37	accuracy, micro f1, none given, f1, internal classifier accuracy, true positive rate, therapy accuracy, test accuracy, temporal binary, strategy accuracy, r, p, pos type statistics, macro f1, external classifier accuracy, false positive rate, error pos cases, emotion accuracy, correctness - nli, controllability, average prediction accuracy, avg, auroc, agency lexicon accuracy, acc, unanswerable accuracy
QEF-w-9:Likelihood According to External Model	11	perplexity, fluency, average perplexity, fluency, independence, silhouette, sufficiency
QEG-w-9:Multi-task Performance	10	accuracy - coreference resolution / sentence completion, accuracy - nli, accuracy, advglue score, average accuracy, average rouge-l, average score, glue score, macro f1, macro-f1 - agg.
QTC-w-2:Sequence Labelling Accuracy	10	accuracy, f1 - character level, f1 - word level, pos accuracy, precision - character level, precision - word level, precision, recall - character level, recall - word level, recall
QIF-w-4:Similarity/Dissimilarity to Input (outputs as a whole)	10	bleu (src), lexical diversity, minimal edits, percentage (%) of 1-grams in the test input sequence that appear in the verbalized idea., percentage (%) of 2-grams in the test input sequence that appear in the verbalized idea., percentage (%) of 3-grams in the test input sequence that appear in the verbalized idea., percentage (%) of 4-grams in the test input sequence that appear in the verbalized idea., percentage (%) of 5-grams in the test input sequence that appear in the verbalized idea., self-bleu, similarity
QTC-w-4:Retrieval Accuracy	9	mrr, accuracy of selecting the neologism or distractor option, exact match, f1, ndcg@3, recall@100, recall, retrieval performance (ret)
QTG-c-1:Similarity to Target Outputs (content/meaning)	8	bert score, accuracy of correct definitions, mover score
QOF-w-1:Diversity/Non-diversity (outputs as a whole)	6	distinct-1, distinct-2, bert score, bleu

Table 4: **Survey 2**: 10 most frequent QCET quality criteria, and the original names they were mapped from.

7 Discussion and Conclusion

We analysed 310 NLP papers reporting 1,002 human and metric evaluations, and mapped the criteria evaluated to standardised names. We found that metrics are dominated by three (standardised) QCs: Classification Accuracy, Similarity to Target Outputs, and Complete Target Output Matching. These account for 55/49% of evaluations in Surveys 1/2. In Survey 3 (mainly human evaluations), the top three QCs account for only 17.9% of evaluations, reflecting greater QC variation.

A key finding is that standardisation reduced QC names by around 74–76% in Surveys 1 and 2 (dominated by metrics), and by 65% in Survey 3 (mostly human evaluations), revealing that a large proportion of naming differences do not reflect actual differences in what is evaluated. This starkly brings home the difficulties of interpreting evaluation results, in particular relative to other evaluation results, in the absence of a standard nomenclature.

Spurious naming differences obscure findings: a standard inventory of quality criterion names and definitions such as QCET is essential if we want conclusions drawn from NLP system evaluations to be fully interpretable and comparable.

In the work reported here, we used QCET to *retrospectively* standardise the aspects of quality as-

QCET quality criterion	N	Original names
QEG-w-3.1:Usefulness for Task/Information Need	37	usefulness, none given, helpfulness, task completion, productivity, user task success, text usefulness, task success, task performance, task ease/success, summary useful (q5), successful trials, ru consultations, required effort (q2), performance, preference, percentage of appropriate actions, parts replaced, ru replacements, learning value, indicator consultations, increased understanding, helped stay on right track, footnote, adequacy of content, worth it
QOG-w-5:Understandability	25	understandability, clarity, comprehensibility, comprehension, readability, easiness perceived, easy to understand, naturalness, none given, text understandability, understandability accuracy, understood paper (q1)
MULTIPLE	23	none given, quality, coherence, fluency, naturalness, readability, all correct, conciseness, elegance, engagement, grammatical correctness and/or understandability, language fluency, readability & understandability, semantic adequacy, writing quality
QOC-f-1:Grammaticality	22	grammaticality, fluency, grammatical correctness, grammatical errors, none given, syntax, error rate, grammar, grammaticality (verb conjugations), grammatically correct or not, syntactic correctness, syntatic correctness = grammticality and fluency
QOG-w-5.1:Clarity	17	understandability, none given, clarity, ease of visualisation, utility, multiple choice score, not misleading, repetition requests
QOG-w-3:Fluency	15	fluency, fluency (document), fluency (sentence), grammatical fluency, phrasing, text quality
QEG-w-7:Clarity of Referents	15	accuracy, effect of text restructuring on a text's discourse level structure (stylistic change evaluation), performance, adequacy, clarity of referential, clarity, error rate, gain, identification time, more suitable, none given, referential clarity
QEF-w-5:Detectability of Speaker/Author Trait	14	personality, agreeableness, competence of dialogue agent, conscientiousness, friendliness of dialogue agent, fun-ness of dialogue agent, informedness of dialogue agent, intelligence of the robot, politeness, relative power, social distance, system understanding, warmth of dialogue agent
QIF-c-2:Similarity/Dissimilarity to Input (content/meaning)	14	adequacy, importance, correctness, humanlikeness assessed with pyramid/modified pyramid scores, information coverage, meaning preservation, meaning similarity, none given, relevance, semantic adequacy, sufficiency
QOG-w-2:Readability	13	readability, reading speed, easy to read, easy, flesh reading ease score, reading errors

Table 5: **Survey 3**: 10 most frequent QCET quality criteria, and the original names they were mapped from.

sessed in existing NLP evaluations, thereby grounding them in a shared nomenclature. However, for new evaluations the idea is to take QCET as a starting point when designing new evaluations and ground aspects of quality assessed in standard QCs *by design*. This would build in comparability and interpretability across evaluation results.

Acknowledgments

Mille's work was funded by the Irish Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media via the eSTÓR project. Thomson's contribution was funded by the ADAPT SFI Centre for Digital Media Technology. Our work has also benefitted more generally from being carried out in the wider context of the ADAPT SFI Centre for Digital Media Technology which is funded by Science Foundation Ireland through the SFI Research Centres Programme, and co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

Limitations

As is always the case, taking a sample from a large population risks sampling bias, and it is possible that taking a different sample of papers would have led to different results. Evaluating a large number like 310 papers should mitigate this somewhat, but clearly we are limited in how many papers can be annotated for 15 different properties by hand. Moreover, it is encouraging that the two different random samples of 60 papers drawn from recent ACL papers produced mostly very similar results.

Ethical Statement

As a survey of published, peer-reviewed papers that have previously undergone ethical review, we consider the ethical risk attached to conducting a survey of such papers minimal.

Appendix

A Paper Annotation Scheme

Below we provide additional details of the 15 dimensions on which we annotated each paper. Note that dimensions 1–6 are automatically extracted from the meta data on the ACL Anthology.

1. **id**: Unique identifier for the paper source, e.g. feng-etal-2022-legal.
2. **assigned_to**: Anonymised annotator ID, identifying the annotator who carried out the first round of annotations, e.g. A1.
3. **year**: Publication year of the paper, e.g. 2022.
4. **url**: Hyperlink to online PDF copy of paper in the ACL Anthology, e.g. https://aclanthology.org/2022.acl-long.48.pdf.
5. **author**: List of paper authors, e.g. Feng, Yi and Li, Chuanyi and Ng, Vincent.
6. **title**: Paper title, e.g. Legal Judgment Prediction via Event Extraction with Constraints.
7. **type**: Type of evaluation annotated, one of {human, metric, hybrid, none found}.
   (a) none found: no evaluation of system output quality was found in the paper;
   (b) human: evaluations where scores or other assessments are assigned by human participants (other aspects may be automated);
   (c) metric: fully automatic evaluations, such as BLEU or perplexity calculated with a given model;
   (d) hybrid: evaluations where scores or other assessments are assigned by a combination of human and metric techniques (other aspects may be manual or automated).
8. **location**: The part of the paper where information included in, or otherwise used for determining values for, **verbatim_qc_name** and **verbatim_qc_definition**, can be found. The value entered is a list of items where each item is one of {Section, Table, Figure, Footnote, Appendix} followed by an integer *N*. E.g.: Section 5, Appendix G, Table G.1.

   If **type**=none found then n/a is selected for location.

   At a minimum, this dimension identifies the part(s) of the paper providing evidence that evaluations were carried out with results reported in the paper. At most, it also identifies the part(s) of the paper where the name and definition of the evaluation criterion used can be found.
9. **verbatim_qc_name**: The name used by authors for what was evaluated, e.g. Macro-F1. If **type**=none found then n/a is selected for **verbatim_qc_name**.

   It does happen that no name is given, for example when just the benchmark dataset is named. In those cases we select none given.

   Often a paper will have two or more variants of an evaluation measure name, e.g. a long form and a short form, the latter often being used as column headers in results tables. E.g. Recall and R, Perplexity and PPL, Reading Comprehension and Comprehension. Here the long form should be entered as the annotation, unless the full name is not mentioned in the paper, or the abbreviation or acronym is the standard given name for the metric, e.g., BLEU and ROUGE.

   If in doubt, multiple variants of the name can be entered, separated by commas.

   NB the values selected here are not necessarily at the level of granularity of a QCET QC. E.g. human evaluation experiments sometimes map to more than one QCET QC.
10. **verbatim_qc_definition**: Definition or explanation of what was evaluated, or None given if no definition or explanation given, surrounded by double quotes.

    If **type**=none found then n/a is selected for **verbatim_qc_definition**.

    For metrics, the most reliable evidence papers provide is the metric's mathematical definition, or a pseudo-algorithmic or narrative walk-through. If these are present, they should be included here. Most often, a paper provides just a gloss, or nothing if the metric is well established (e.g. BLEU).

For human evaluations, the most reliable level of evidence is provided by the question and/or instructions put to evaluators, if given verbatim. More often, a less detailed gloss of what was evaluated is provided. As a last resort, we record anything that describes what was done beyond just giving a name for what was evaluated.

E.g. "Task accuracy is measured as the percentage of instances for which the model selects a proposal whose intersection-over-union (IoU) with the ground-truth box is at least 0.5."

11. **qcet_node**: The standardised QCET quality criterion ID and name if a matching node is found in the QCET taxonomy; MULTIPLE if multiple QCET QCs are evaluated by the same single evaluation measure; not found, if no match is found; or n/a if **type**=none found.

    In performing this mapping, we use the information entered for **type**, **location**, **verbatim_qc_name**, and **verbatim_qc_definition**, and follow the instructions in Section 5.1 of the extended QCET paper (Belz et al., 2025a).

    Occasionally evaluation measures in papers are at a higher level of granularity than QCET QCs and evaluate more than one aspect of quality, e.g. when evaluators are instructed to score both grammatical correctness and humanlikeness in the same assessment; these we annotate as MULTIPLE.

12. **external_location_consulted**: In some cases, the information provided in the paper, and captured for **type**, **location**, **verbatim_qc_name**, and **verbatim_qc_definition**, is not enough to perform the mapping required for **qcet_node**. In those cases, sometimes another paper is referenced in the paper being annotated in which sufficient information can be found. Where this is the case, this dimension provides the URL(s) of the paper(s) that was/were additionally consulted in performing the mapping. Left blank if none consulted.

    If **qcet_node**=Not found, then we additionally complete the following dimensions, to facilitate expanding the QCET taxonomy to cover them. Note that these annotations are not included in the public release of the dataset; instead we provide the **qcet_node** that was added to the taxonomy directly.

13. **proposed_qc_name**: If **qcet_node**=not found, we enter here the proposed ID and name for the new QC to be added to QCET. Else, n/a.

14. **proposed_qc_definition**: If **qcet_node**=not found, we enter here the proposed definition for the new QC to be added to QCET. Else, n/a.

15. **proposed_qcet_branch**: If **qcet_node**=not found, we enter here the branch ID and name where the proposed new QCET node should be inserted. Else, n/a.

Further details and notes on inclusion criteria, identifying individual experiments and edge case are distributed with the release of the dataset of paper annotations.

B Quality Criteria

Quality Criteria (QCs) represent the basic level at which we would expect two high-quality evaluations (that assess the same QC) to support the same conclusions about which of two systems is better. We would not expect implementational details, operationalisation, whether we're assessing one or two systems at a time, etc., to affect such coarse-grained conclusions, and this is what makes the assessed aspect of quality the right level for comparability grounding.

To support such assessments, we decompose evaluation methods M into QC, **evaluation modes** and **experiment design** which relate as follows:

> Quality criterion + evaluation mode = **evaluation measure**;
> Evaluation measure + experiment design = **evaluation method**.

Evaluation mode (Belz et al., 2020) has three dimensions: **extrinsic** (impact on something external to the system is assessed) vs. **intrinsic** (otherwise); **objective** (repeated measurements yield the same results to within some margin of precision) vs. **objective** (otherwise); and **absolute** (one system evaluated at a time) vs. **relative** (more than one system evaluated at a time).

In order to disentangle the QC, we need to peel away everything to do with evaluation mode, experiment design and implementation, to find the answer to the question: *Q: What does this evaluation consider a better system to be?*

For example, a better system is not one that is found to be better in an experiment with 8 users that access an airline schedule database to book a

specific flight while talking to the system and take less time to do it. Whether there are 8 users or a different number, where they get hold of the flight information, etc., is not part of system quality, but of experiment design or system implementation. Once we disregard such factors, we are left with the following answer for the above example: *A: A better system is one that enables the user to complete a task more quickly.*

C Diagrammatic View of QCET Taxonomy

Figure 1 shows the whole of the QCET taxonomy in diagrammatic overview, displaying node IDs and QC names only.

D Normalisation Schema for Verbatim QC Names

Table 6 shows the normalisation scheme for verbatim quality criterion names. All verbatim names were first lower-cased, then had this mapping applied.

E List of Quality Criteria from QCET mentioned in paper, with Definitions and Notes

[QTC-w-1] Classification Accuracy: A better system produces output classes that less often differ from the given target output class (from a given finite set of classes).

Example: Jarvis et al. (2013) evaluate a classifier that predicts the native language of participants by computing class Recall on 11,000 texts (in 11 languages).

Notes: The notion of accuracy in this QC is wider than just the Accuracy metric, encompassing also e.g. Recall, Precision, F-score, and other combinations of true/false positives/negatives.

[QTG-w-1] Similarity to Target Outputs (outputs as a whole): A better system produces outputs that are overall more similar to given target outputs.

Example: In the WebNLG shared tasks Gardent et al. (2017), the similarity between outputs of data-to-text generators and target system outputs is evaluated using BLEU (strict n-gram matching) and METEOR (allowing synonyms and morphological variation).

Notes: Similarity to target outputs is a very common form of evaluation in NLP where one or more target outputs (often called gold outputs or references) are provided as part of a test set, and the degree of similarity between actual system output and target output(s) is measured. Note this is different from binary same/not same assessments made e.g. in Classification Accuracy. [QTG-w-1] Similarity to Target Outputs (outputs as a whole) assesses overall similarity, not distinguishing form or content.

[QTC-w-3] Complete Target Output Matching: A better system produces outputs that less often differ from a given target output (where the set of possible outputs is not given, and is not necessarily finite).

Example: Yue et al. (2022) report the exact match (EM) rate for question-answer generation systems, where an exact match is a question that appears in the set of target output questions verbatim.

Notes: In contrast to Classification Accuracy, this QC is defined for cases where the system does *not* choose between a set of posible outputs that is known a priori. Instead the output is typically generated in some way from the input, and is often quantified as the exact match rate.

[QOF-w-1] Diversity/Non-diversity (outputs as a whole): A better system produces outputs that are either (a) more diverse, or (b) less diverse.

Example: Jin and Le (2016) ask evaluators to assess the overall diversity of a set of questions generated by a system from a given input text.

Li et al. (2015) assess the diversity of the responses produced by their conversational system with the distinct-1 and distinct-2 metrics computed as the number of distinct unigrams (bigrams) over the total number of generated tokens.

Notes: (Non)diversity of outputs as a whole captures diversity of form and content both, either at the level of individual outputs where those are longer than a sentence, or at the level of a sample of outputs. E.g. a diverse set of questions and answers generated for a given text passage would have little overlap in coverage of the text between them.

[QEG-w-8] Performance of an Embedding/Downstream System/Component: A better system produces outputs that result in outputs with better performance when used by another system or component.

Example: Reddy et al. (2017) evaluate a system that generates question-answer pairs from keywords by assessing whether its outputs can improve the performance of a semantic parser when added

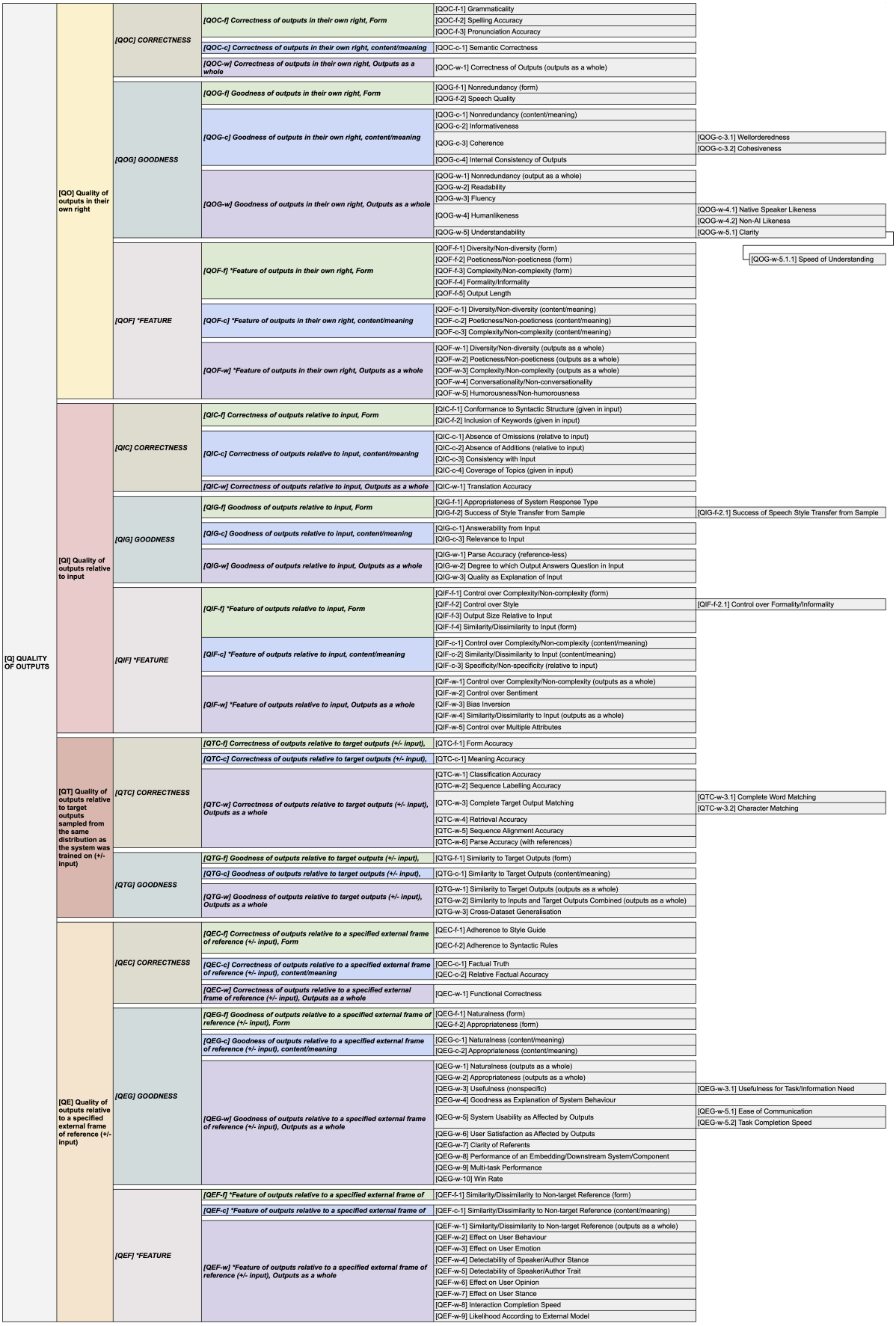

Figure 1: Diagrammatic overview of the QCET taxonomy, showing node IDs and quality criterion (QCs) names only. * = node is a class of QCs, but not a QC in its own right.

Original verbatim	Normalized form
"alignment error rate"	alignment error rate
advglue mean accuracy	advglue score
average accuracy of 8 plms	average accuracy
average glue score	glue score
bertscore	bert score
bleu - compound-splitting	bleu
bleu4	bleu-4
correctness - rougel	rouge-l
correctness - unigram f1 - wow dataset	unigram f1
em	exact match
entityf1	entity f1
f1 score	f1
factual	factuality
glue mean accuracy	glue score
gpt-2 ppl	perplexity
grammatical	grammaticality
kappa correlation coefficient	kappa
las	labelled attachment score
macro f1 score	macro f1
macro-f1	macro f1
macro-precision	macro precision
macro-recall	macro recall
micro-averaged f1	micro f1
micro-f1	micro f1
moverscore	mover score
mrr mean reciprocal rank	mean reciprocal rank
pearson correlation between gaps	pearson
pearson correlation coefficient	pearson
precision@1	precision
r@1	recall
rouge 1	rouge-1
rouge 2	rouge-2
rouge l	rouge-l
rouge1	rouge-1
spearman correlation on multilingual sts 2017	spearman
spearman correlations	spearman
spearman's correlation on semantic similarity scores	spearman
top-1 accuracy (image)	accuracy (image)
top-1 accuracy (text)	accuracy (text)
wer	word error rate

Table 6: Normalisation used for name standardisation following lower-casing.

to its training data. Comparing performance when (a) training on manual question-answer pairs only with (b) training on the manual data plus the system's question-answer pairs, they find a 5.5% improvement for the augmented training data.

Notes: Assessing a system in terms of the impact its outputs have on the performance of the bigger system it is part of, or in terms of another system that uses its outputs, is often called an extrinsic form of evaluation. Extrinsic evaluation is especially suitable for NLP components embedded in a larger system, such as a TTS component that is part of an interactive system.

[QEG-w-9] Multi-task Performance: A better system produces outputs that obtain higher aggregated scores on a given set of task datasets and metrics.

Example: Zhou et al. (2023) explore data leakage in LLM assessment, evaluating different models on the MMLU benchmark of 57 different tasks that require real-world knowledge and problem-solving abilities.

Notes: Multi-task benchmarks have become increasingly common in NLP, particularly in LLM evaluation. [QEG-w-10] Multi-task Performance covers any case where aggregated results expressing performance at multiple tasks is reported.

[QEF-w-9] Likelihood According to External Model: A better system produces outputs that are estimated to be more likely by a given external model.

Example: Yedetore et al. (2023) train various models (5-gram model, LSTMs, Transformers) on child-directed language data, and use perplexity (a standard formulation as well as the word-frequency normalised SLOR metric) to evaluate how well each model captures the basic structure of the train-

ing domain, finding that Transformers have the lowest perplexity, the 5-gram model the highest.

Notes: The more common use of perplexity is in evaluations where *low* perplexity as computed with a given model is desirable, where it's seen as indicative of 'natural' output. However, it can equally be desirable for outputs to have high perplexity, e.g. in situation where a different style from that encapsulated by the model is intended. [QEF-w-9] Model Perplexity is a Feature-type QC, hence captures both possibilities. Note that various metrics exist for measuring model perplexity including normalised ones such as SLOR.

[QIF-f-2] Control over Style: A better system produces outputs that are in their form more in the target style provided in the input.

Example: Gero et al. (2019) evaluate whether automatically generated sentences exhibit the style specifed in the input, by asking human annotators to label output sentences with style labels and then calculating the accuracy compared to the input labels.

Notes: Style captures aspects of form, as opposed to meaning. It relates to the way something is said, rather than what is said. Examples include formal/informal, literary/non-fiction, and more fine-grained distinctions such as newspaper house style and personal writing style. [QIF-f-2] Control over Style typically is used to evaluate systems that are trained on a specific set of alternative styles, with a control attribute in the input indicating the style that outputs are supposed to be generated in.

[QTG-c-1] Similarity to Target Outputs (content/meaning): A better system produces outputs that are in their content/meaning more similar to given target outputs.

Example: In their study about experiment design for the evaluation of dialogue system ouptuts, Santhanam and Shaikh (2019) compute the cosine similarity between embeddings of (a) system responses and (b) target system responses'.

Mille et al. (2018) evaluate multilingual surface realisers that take syntactic or semantic trees as input by asking raters to assess the meaning similarity between system outputs and the target outputs (i.e. the original sentences previously parsed to get the inputs).

Notes: Similarity to target outputs is a very common form of evaluation in NLP where one or more target outputs (often called gold outputs or references) are provided as part of a test set, and the degree of similarity between actual system output and target output(s) is measured. Note this is different from binary same/not same assessments made e.g. in Classification Accuracy. [QTG-c-1] Similarity to Target Outputs (content/meaning) assesses similarity in terms of content units or semantic representations.

[QEC-c-1] Factual Truth: A better system produces texts with fewer real-world untruths.

Example: Thomson and Reiter (2020) evaluate the outputs of their sports summarisation system by asking participants (a) to mark up factual errors as determined by open web search as non-overlapping word spans, then (b) to categorise the word spans. They report an average of 19 erorrs per summary.

Notes: In assessing [QEC-c-1] Factual Truth, the aim is to establish the real-world truth or untruth of output content. In contrast to [QEC-c-2] Relative Factual Accuracy, specific information sources (not expected to contain contradictory information) are not normally provided in evaluation. More typically, a process is described whereby truth is to be established for the purposes of the evaluation which may involve resolving any amount of contradictory information.

[QTC-w-2] Sequence Labelling Accuracy: A better system produces sequences of output labels that less often differ from the given target output label (from a given finite set of labels).

Example: de Vries et al. (2022) compute the part-of-speech (POS) tagging accuracy achieved by a task-tuned model on pairs of languages where one was seen during task-tuning and the other was not. They report POS tagging accuracy for a large number of pairs of languages some of which were seen during model pretraining, some were not.

[QIF-c-2] Similarity/Dissimilarity to Input (content/meaning): A better system produces outputs that in their content/meaning are (a) more similar to the input, (b) less similar to the input, or (c) more at the target level of similarity to the input, where that target level is provided in the input.

Example: In their survey of text style transfer research, Hu et al. (2022) identify some of the main ways in which previous work has measured the meaning (dis)similarity between source text and transferred text: cosine similarity between embeddings, word overlap (excluding style-related words), and human assessment of meaning (dis)similarity.

Notes: [QIF-c-2] Similarity/Dissimilarity to Input (content/meaning) is the same as [QEF-c-1] Similarity/Dissimilarity to Non-target Reference (content/meaning), but here the comparison is against the input, rather than a system-external reference. Typical NLP tasks where this QC is assessed include paraphrasing and style transfer.

[QOG-w-3] Fluency: A better system produces outputs that are more fluent.

Example: Resendiz and Klinger (2025) evaluate the fluency of affective text generation systems via LLM-prompting with the following prompt: "Assess the text's fluency, assigning a score from 1 to 5, with 5 representing the highest level of fluency. Do not give an explanation of the selection."

Notes: Fluency captures how well text or speech flows, being absorbed readily without bringing the reader or listener up short, and without in the case of speech, hesitations, filler, or overly long pauses. For high fluency, language does not necessarily need to be simple, cf. [QOG-w-2] Readability.

[QEG-w-3.1] Usefulness for Task/Information Need: A better system produces outputs that are more useful for the user's task and/or information need.

Example: Qu and Green (2002) assess a cooperative mixed-initiative dialogue system for information-seeking dialogue via a user study where they measure the agreement between "the user's recorded solution for each task" and "the user's original information need" with the kappa statistic (Carletta, 1992).

Notes: [QEG-w-3.1] Usefulness for Task/Information Need shares the same characteristics of its parent QC[QEG-w-3] Usefulness (nonspecific), but is more specific, assessing usefulness for a given task, such as following generated instructions to trouble-shoot a malfunctioning app, or for a given information need, e.g. are the accommodation descriptions on a website useful in selecting a holiday rental.

[QOG-w-5] Understandability: A better system produces outputs that are more understandable.

Example: Hershenhouse et al. (2024) assess LLMs in terms of their ability to communicate medical information to the public by asking crowdworkers to demonstrate their understanding of generated texts through multiple-choice questions.

Notes: Understandability captures whether an output can be understood and is commonly evaluated in terms of whether it has been understood (via comphrehension questions). Cf. sub-QC [QOG-w-5.1] Clarity for which an output is assessed in terms of the higher-threshold criterion whether it can be *easily* understood.

[QOC-f-1] Grammaticality: A better system produces texts with fewer grammatical errors.

Example: Humphreys et al. (2001) informally evaluate 200 outputs manually for Grammaticality reporting 4% of outputs with grammatical errors for their combined parser/generator.

Notes: [QOC-f-1] is Grammaticality as judged by native speakers, i.e. it's a human-assessable only QC. Cf. [QIC-f-1] Matching Syntactic Structure (given in input), and [QEC-f-2] Adherence to Syntactic Rules which can be assessed either with metrics or humans.

[QOG-w-5.1] Clarity: A better system produces outputs that are clearer.

Example: Clinciu et al. (2021) evaluate a system that generates explanations of Bayesian network graphs, e.g. in terms of clarity where evaluators are asked to indicate "[h]ow clear the meaning of an explanation is" on a 7-point scale, where 1 = unclear and 7 = very clear.

Notes: –

[QEF-w-5] Detectability of Speaker/Author Trait: A better system produces outputs that make the entity producing the output come across to an observer as having one of a range of given traits (a) more, (b) less, or (c) to the degree specified in the input.

Example: Glas and Pelachaud (2015) evaluate different strategies for dialogue agents to introduce new topics into conversation, assessing which makes the user perceive the agent as more (a) competent, (b) friendly, (c) fun, and (d) informed, i.e. four different traits.

Notes: [QEF-w-5] Detectability of Speaker/Author Trait is about the degree to which the user perceives the entity producing the outputs (which may be perceived as an interlocutor) as having a given trait. A trait in this context is usually something that can be captured in a single adjective, as in the example attestation.

[QEG-w-7] Clarity of Referents: A better system produces referring expressions that more clearly identify their referents.

Example: In the 2005 DUC shared task on summarisation, Dang (2005) manually assess systems

in terms of 'Referential clarity,' defined as follows: "It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear."

Notes: Clarity of referents is about how readily intended referents can be identified from referring expressions in texts or other representations. The explanation included in the DUC 2005 attestation provides a good explanation for the case of texts.

[QOG-w-2] Readability: A better system produces outputs that are more readable.

Example: Afsal and Kuppusamy (2024) compare the readability of texts generated with the Gemini LLM via different prompts with the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) metrics.

Notes: Readability captures 'reading ease' in the sense of text measures like Flesch and Flesch-Kincaid which aim to capture ability to easily read at different reading ages. Better readability is associated e.g. with more common words, shorter words, shorter sentences, and simpler sentence structure. Cf. [QOG-w-3] Fluency: very short sentences with repetitive structure would score highly on Readability, but not on Fluency.

[QIC-c-3] Consistency with Input: A better system produces outputs that have fewer inconsistencies with a given aspect of the input.

Example: Shu et al. (2021) propose a metric called bi-directional logic evaluation of consistency (BLEC) for evaluating the consistency between database query logic inputs and textual questions in the output.

Notes: Cf. Similarity/Dissimilarity to Input (content/meaning); Consistency with Input is not about meaning similarity, but consistency in a task-specific sense, for which absence of contradictions may be sufficient.

References

C.P Afsal and K S Kuppusamy. 2024. Assessing the readability and coherence in gemini's triple draft generation: A multi-metric approach. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Simon Mille, and Craig Thomson. 2025a. The qcet taxonomy of standard quality criterion names and definitions for the evaluation of nlp systems. *Preprint*, arXiv:2509.22064.

Anya Belz, Simon Mille, and Craig Thomson. 2025b. Standard quality criteria derived from current NLP evaluations for guiding evaluation design and grounding comparability and AI compliance assessments. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26685–26715, Vienna, Austria. Association for Computational Linguistics.

Anya Belz, Simon Mille, and Craig Thomson. to appear. Standardising evaluation criterion names and definitions in nlp via systematic surveys.

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 Krahmer, 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 *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.

Miruna Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12. Citeseer.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Katy Gero, Chris Kedzie, Jonathan Reeve, and Lydia Chilton. 2019. Low level linguistic controls for style transfer and content preservation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 208–218, Tokyo, Japan. Association for Computational Linguistics.

Nadine Glas and Catherine Pelachaud. 2015. Topic transition strategies for an information-giving agent. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 146–155, Brighton, UK. Association for Computational Linguistics.

Jacob S Hershenhouse, Daniel Mokhtar, Michael B Eppler, Severin Rodler, Lorenzo Storino Ramacciotti, Conner Ganjavi, Brian Hom, Ryan J Davis, John Tran, Giorgio Ivan Russo, et al. 2024. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. *Prostate Cancer and Prostatic Diseases*, pages 1–6.

David Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid Hasan, Saad Mahamood, Simon Mille, Sashank Santhanam, Emiel van Miltenburg, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Natural Language Generation Conference*.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *SIGKDD Explor. Newsl.*, 24(1):14–45.

Kevin Humphreys, Mike Calcagno, and David Weise. 2001. Reusing a statistical language model for generation. In *Proceedings of the ACL 2001 Eighth European Workshop on Natural Language Generation (EWNLG)*, Toulouse, France. Association for Computational Linguistics.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118, Atlanta, Georgia. Association for Computational Linguistics.

Yiping Jin and Phu Le. 2016. Selecting domain-specific concepts for question generation with lightly-supervised methods. In *Proceedings of the 9th International Natural Language Generation conference*, pages 133–142, Edinburgh, UK. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Simon Mille, Anya Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The first multilingual surface realisation shared task (SR'18): Overview and evaluation results. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12, Melbourne, Australia. Association for Computational Linguistics.

Yan Qu and Nancy Green. 2002. A constraint-based approach for cooperative information-seeking dialogue. In *Proceedings of the International Natural Language Generation Conference*, pages 136–143, Harriman, New York, USA. Association for Computational Linguistics.

Sathish Reddy, Dinesh Raghu, Mitesh M. Khapra, and Sachindra Joshi. 2017. Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 376–385, Valencia, Spain. Association for Computational Linguistics.

Yarik Menchaca Resendiz and Roman Klinger. 2025. Mopo: Multi-objective prompt optimization for affective text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5588–5606.

Jie Ruan, Wenqing Wang, and Xiaojun Wan. 2024. Defining and detecting vulnerability in human evaluation guidelines: A preliminary study towards reliable NLG evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7965–7989, Mexico City, Mexico. Association for Computational Linguistics.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Patricia Schmidtova, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondrej Dusek. 2025. Do my eyes deceive me? a survey of human evaluations of hallucinations in NLG. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 60–79, Hanoi, Vietnam. Association for Computational Linguistics.

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

Chang Shu, Yusen Zhang, Xiangyu Dong, Peng Shi, Tao Yu, and Rui Zhang. 2021. Logic-consistency text generation from semantic parses. *Preprint*, arXiv:2108.00577.

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.

Tim Van de Cruys. 2020. Automatic poetry generation from prosaic text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2471–2480.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 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, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Aditya Yedetore, Tal Linzen, Robert Frank, and R Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Xiang Yue, Ziyu Yao, and Huan Sun. 2022. Synthetic question value estimation for domain adaptation of question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1340–1351, Dublin, Ireland. Association for Computational Linguistics.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *Preprint*, arXiv:2311.01964.