

Defining and Measuring Quality in Automatically Generated Language: A Systematic Review

Abderrahmen TAHRI¹

¹Mohamed Boudiaf University - Msila, Algeria

E-mail: abderrahmane.tahri@univ-msila.dz

ORCID : <https://orcid.org/0009-0004-0042-8899>

Received: 18/07/2026, Accepted: 20/07/2026, Published: 26/07/2026

ABSTRACT: *The rapid proliferation of natural language generation (NLG) systems has intensified the need for robust frameworks to define and assess the quality of their content. This systematic review synthesizes research published between 2018 and 2023 to examine how quality is conceptualized and operationalized across key NLG tasks, including machine translation, summarization, dialogue systems and creative text generation. A total of 87 peer-reviewed studies were analysed, revealing three dominant evaluation paradigms: human-centred approaches (e.g., fluency, adequacy, and coherence), automated metrics (e.g., BLEU and BERTScore), and task-specific composite measures. The findings highlight persistent discrepancies between human judgments and automated scores, an overreliance on surface-level metrics, and a lack of standardized taxonomies that incorporate domain-specific and ethical dimensions, such as factual consistency and bias. In response, this review proposes a multilayered evaluation framework that integrates quantitative metrics with qualitative human assessment. This study provides a foundation for developing more reliable and comprehensive quality assurance practices for applied natural language processing systems.*

KEYWORDS: Natural language generation, Evaluation metrics, Content quality assessment, Human–AI collaboration, Systematic review.

* Corresponding author: Tahri Abderrahmen

Introduction

Automatic language generation has moved from a specialised area of computational linguistics into a technology that now supports translation, summarization, dialogue, writing assistance, education, customer service, medical documentation and creative production. This shift has made the question of quality more urgent and more difficult. In earlier NLG settings, evaluation often meant checking whether a system produced a fluent sentence or selected the right pieces of information from a structured input. Current systems, especially neural and large language model based systems, can produce long, polished and context-sensitive texts. The surface of the output may look convincing even when the content is incomplete, poorly grounded, biased or pragmatically unsuitable. As a result, quality can no longer be treated as a simple matter of grammatical correctness or lexical overlap with a reference text.

The difficulty is partly conceptual. Different NLG tasks require different notions of success. In machine translation, quality has often been discussed through adequacy and fluency, because the generated text is judged against a source text and a communicative target. In summarization, the same generated sentence must be readable, concise, relevant and faithful to the source material. In dialogue systems, quality includes coherence over turns, appropriateness, user satisfaction and sometimes safety. In creative generation, diversity, originality and stylistic control become more prominent, but they cannot replace basic coherence and acceptability. Studies on NLG metrics have therefore stressed that evaluation cannot be reduced to one universal score that works equally well across tasks, domains and users (Sai, Mohankumar, & Khapra, 2020; Van Der Lee, Gatt, Van Miltenburg, Wubben, & Krahmer, 2019).

The difficulty is also methodological. Human evaluation is usually regarded as the most direct way to judge generated language, but it is expensive, time-consuming and sensitive to the design of the rating protocol. Small changes in the wording of criteria, the number of annotators, the scale used, the expertise of raters or the examples shown before the task can change the resulting scores. Automatic metrics are attractive because they are fast, reproducible and easy to integrate into model development, but many of them measure convenient proxies rather than quality itself. Traditional overlap-based metrics reward similarity to a reference and may penalise valid paraphrases. Embedding-based metrics such as BERTScore were introduced to capture semantic similarity more flexibly, yet they still do not fully solve problems of factuality, pragmatic usefulness or ethical risk (Zhang, Kishore, Wu, Weinberger, & Artzi, 2019).

This systematic review examines how quality in automatically generated language was defined and measured in peer-reviewed research published between 2018 and 2023. The period is important because it covers the consolidation of neural generation, the spread of pretrained language models, and the early evaluation literature around large language models. The review focuses on four high-impact NLG task families: machine translation, summarization, dialogue systems and creative or open-ended text generation. It is guided by four questions: how do studies conceptualise quality; which human and automatic methods are used to operationalise it; where do human judgments and automatic metrics diverge; and what type of evaluation framework can better support applied NLP systems?

The contribution of the article is threefold. First, it maps the main quality dimensions used across NLG research and shows that the field has gradually moved from surface-level criteria towards broader concerns such as factual consistency, robustness and bias. Second, it compares three dominant evaluation paradigms: human-centred assessment, automatic metrics and task-specific composite measures. Third, it proposes a multilayered evaluation framework that combines quantitative metrics with qualitative human review. The intention is not to reject automatic evaluation, which remains essential for large-scale research and development, but to place it within a more transparent and accountable quality assurance process.

Methodology

This study was designed as a systematic review of research on the quality evaluation of automatically generated language. The review covered studies published from January 2018 to December 2023. Foundational work published before 2018 was used only when it clarified the development of evaluation practice, especially the long-standing debate about the validity of automatic NLG metrics (Reiter & Belz, 2009). The review was restricted to peer-reviewed articles, conference papers and well-established computational linguistics proceedings dealing directly with NLG evaluation, automatic metrics, human evaluation protocols or task-specific quality assessment.

The search strategy combined terms related to natural language generation and evaluation, including natural language generation, generated text, text generation, machine translation evaluation, summarization evaluation, dialogue evaluation, factual consistency, human evaluation, automatic metrics, quality assessment, BERTScore, factuality and alignment. The search was oriented towards the major venues and indexing sources in computational linguistics and applied NLP, particularly those where NLG evaluation studies are commonly published. After screening titles, abstracts and full texts, 87 studies were retained for qualitative synthesis. Because the studies differed widely in task, dataset, metric, model type and human-evaluation protocol, a statistical meta-analysis was not appropriate. The synthesis therefore followed a descriptive and comparative approach.

The inclusion criteria were as follows: the study had to evaluate generated natural language or evaluate an evaluation metric for generated language; it had to include at least one identifiable quality criterion; it had to report either human judgments, automatic metric results or a comparison between them; and it had to belong to one of the selected task families or present a general NLG evaluation method. Studies were excluded when they focused only on model architecture without an evaluation component, when they evaluated non-linguistic outputs, when the evaluation criteria were not described, or when the paper was not accessible in sufficient detail to identify the quality dimensions used.

Each retained study was coded using a structured extraction grid. The grid recorded the task domain, the type of generated output, the definition of quality, the human evaluation criteria, the automatic metrics, the presence or absence of reference texts, the use of expert or crowd annotators, the scale or ranking procedure, and any reported relationship between automatic scores and human judgments. A second layer of coding looked for broader issues that have become increasingly important in NLG evaluation: factuality, robustness, explainability, bias, safety and domain-specific usefulness. This coding scheme was informed by best-practice discussions in human evaluation (Van Der Lee et al., 2019), surveys of NLG metrics (Sai et al., 2020), and recent work on the limits of metric correlation and metric preference (Ni'mah, Fang, Menkovski, & Pechenizkiy, 2023; Sai, Dixit, Sheth, Mohan, & Khapra, 2021).

The analysis proceeded in three stages. First, the studies were grouped according to their dominant evaluation paradigm: human-centred assessment, automatic metrics or task-specific composite measures. Second, the reported quality criteria were clustered into recurring dimensions such as fluency, adequacy, coherence, relevance, factual consistency, diversity and usefulness. Third, the review compared the strengths and weaknesses of the evaluation practices found in each task family. Particular attention was given to cases where automatic scores appeared to conflict with human judgments, because such cases reveal what a metric actually captures and what it leaves outside the measurement process.

Several limitations should be made explicit. The review relies on information reported by the original authors; many papers do not describe annotator training, inter-annotator agreement or the exact wording of the evaluation prompt in enough detail. In addition, the term quality is used unevenly across the literature. Some papers treat quality as a global rating, whereas others break it into detailed sub-criteria. The findings should therefore be read as a synthesis of reported evaluation practices rather than a

direct remeasurement of all systems. This limitation is not a weakness of the review alone; it reflects a wider problem in the NLG evaluation literature.

Conceptualizing Quality in Generated Language

Across the reviewed studies, quality appears less as a single property and more as a layered construct. The first layer is linguistic acceptability. This includes grammar, fluency, readability, naturalness and the absence of awkward or broken phrasing. Many studies still begin with this layer because a text that fails at the level of form is unlikely to be accepted by users. Tools such as GRUEN were proposed to evaluate linguistic quality by combining signals related to grammaticality, focus, coherence and non-redundancy (Zhu & Bhat, 2020). Such work is useful because it moves beyond simple word overlap, but linguistic acceptability remains only one part of quality. A fluent hallucination may still be a poor output.

The second layer is semantic adequacy. In translation, this means that the target text should preserve the meaning of the source. In data-to-text generation, it means that the output should express the relevant input facts without adding unsupported information. In summarization, adequacy overlaps with relevance and coverage: a good summary should retain the central information from the source without distorting it. Machine translation studies have long recognized the tension between automatic and human measures of adequacy and fluency, and systematic reviews of translation quality assessment show that human and automatic approaches often answer different evaluation questions (Chatzikoumi, 2019; Rivera-Trigueros, 2021).

The third layer is discourse-level coherence. A generated text may contain correct sentences but still fail as a text if the information is poorly ordered, if referents are unclear, or if the argument does not develop logically. This is especially important in open-ended generation, dialogue and long-form summarization. Human evaluation studies of AMR-to-English generation, for example, show that outputs can be assessed not only for sentence-level fluency but also for whether they preserve meaning and produce coherent English realizations (Manning, Wein, & Schneider, 2020). Coherence is difficult for automatic metrics because it depends on relations across sentences, not just local similarity to a reference.

The fourth layer is factual consistency. This dimension has become central with the rise of neural summarization and large language models. A system may generate a grammatically polished summary that includes facts not present in the source or changes relations between entities. Research on factuality in summarization and long-form generation has therefore developed tools and benchmarks that examine whether generated claims are supported by evidence (Chern et al., 2023; Goodrich, Rao, Saleh, & Liu, 2019; Min et al., 2023; Muhlgay et al., 2023; Tam, Mascarenhas, Zhang, Kwan, Bansal, & Raffel, 2022). Factual consistency is not simply another metric; in many applied contexts, it is a condition for safe use.

The fifth layer is pragmatic and task-specific usefulness. A generated answer, explanation, summary or dialogue turn must serve a user need. In natural language explanations, for example, a text may be fluent and semantically related to the input but still fail if it does not help the user understand a model decision or a task outcome. Clinciu, Eshghi, and Hastie (2021) show that explanation evaluation requires attention to criteria that are not always captured by generic NLG metrics. Similarly, dialogue quality involves appropriateness, consistency over turns and user satisfaction, while creative generation may require novelty, diversity and style control. Montahaei, Alihosseini, and Baghshah (2019) address this issue by jointly considering diversity and quality, reminding us that repeated safe outputs are not necessarily good outputs in creative settings.

Finally, quality increasingly includes ethical and social dimensions. Bias, toxicity, stereotyping, privacy risk and misleading confidence are not peripheral concerns when generated language is deployed in education, administration, health, media or public communication. Yet the review shows that these dimensions are often added after the main evaluation rather than integrated into the definition of quality

from the beginning. This is a serious gap. If generated language is meant to support human work, its quality must include not only what the text says and how well it reads, but also what effects it may have on users, institutions and represented groups.

Dominant Evaluation Paradigms

The first dominant paradigm is human-centred evaluation. It includes expert review, crowd annotation, pairwise comparison, ranking, Likert-scale scoring and user studies. Its main advantage is that it can capture aspects of language quality that are not easy to formalize, such as naturalness, usefulness, tone, coherence and perceived trustworthiness. Human evaluation is also necessary when there is no single reference answer. Van Der Lee et al. (2019) argue that human evaluation needs careful design, including clear criteria, appropriate raters and transparent reporting. Without these elements, human scores may appear authoritative while remaining difficult to interpret or reproduce.

In the reviewed literature, human evaluation is strongest when the task criteria are explicitly separated. A single overall score is easy to report, but it hides the reason for failure. A translation may be fluent but inaccurate. A summary may be concise but unfaithful. A dialogue response may sound natural but ignore the user's request. Multi-criteria rubrics reduce this ambiguity by asking raters to judge, for example, fluency, adequacy, coherence and factuality separately. Pairwise comparisons can also be useful because humans often find it easier to choose the better of two outputs than to assign an absolute number. However, pairwise designs become costly when many systems are compared.

The second paradigm is automatic evaluation. Automatic metrics make it possible to compare systems at scale and to monitor progress during model development. Older reference-based metrics reward similarity between generated output and one or more human references. Their practicality explains their persistence, especially in machine translation and summarization. However, they are limited when the task allows multiple valid outputs. Semantic metrics such as BERTScore attempt to reduce this problem by comparing contextual embeddings rather than exact word overlap (Zhang et al., 2019). Learned or model-based metrics go further by predicting human preferences or assessing more specific dimensions such as factual alignment.

Recent work shows both the promise and the instability of automatic metrics. Hashimoto, Zhang, and Liang (2019) propose ways to connect human and statistical evaluation, while Sai et al. (2021) use perturbation checklists to test whether metrics respond sensibly to controlled changes in generated text. Ni'mah et al. (2023) argue that correlation with human judgments is not enough; a metric should also be examined through preferences, robustness and task behaviour. This line of work is important because high correlation on one benchmark can create a false impression of general validity. A metric may perform well for one dataset, one task or one error type, yet fail on another.

The third paradigm is task-specific composite evaluation. Here, researchers combine several indicators because no single measure is considered sufficient. In factuality evaluation, for example, systems may be judged by claim extraction, evidence retrieval, entailment, atomic fact scoring or tool-augmented verification. FacTool evaluates factuality across multiple tasks and domains by using tools to support the checking process (Chern et al., 2023). AlignScore aims to measure factual consistency through a unified alignment function (Zha, Yang, Li, & Hu, 2023). FActScore evaluates long-form generation by decomposing text into atomic factual units (Min et al., 2023). These methods are more aligned with the real risk of generated text, but they also introduce new design decisions that must be reported clearly.

Table 1 summarises the three paradigms and the type of quality evidence they usually provide. The table also shows why the review treats evaluation as a triangulation problem. Each paradigm sees part of the object. Human judgment captures social and communicative acceptability, automatic metrics provide

scale and repeatability, and composite measures bring task-specific sensitivity. Reliable evaluation requires a deliberate combination of these views rather than a competition between them.

Paradigm	Typical criteria	Main strength	Main limitation
Human-centred assessment	Fluency, adequacy, coherence, usefulness, tone, satisfaction	Captures communicative and social acceptability	Costly; sensitive to rater selection, rubric wording and reporting quality
Automatic metrics	Reference similarity, semantic similarity, linguistic quality, learned preference	Fast, reproducible and scalable across systems	Often measures proxies; may miss factuality, pragmatic usefulness and bias
Task-specific composite measures	Factuality, alignment, diversity, task success, domain risk	Better adapted to task-specific failure modes	Requires careful validation and transparent design choices

Table 1. Dominant paradigms in the evaluation of automatically generated language.

Findings and Synthesis

The first major finding is that quality is usually defined through familiar linguistic criteria, but the meaning of these criteria changes across tasks. Fluency is common, yet a fluent translation, a fluent summary and a fluent dialogue response are not evaluated in the same communicative situation. Adequacy is central in translation and data-to-text generation, while relevance and faithfulness dominate summarization. Coherence appears across almost all tasks, but it is often measured weakly, either through human impressions or indirect automatic signals. This confirms a recurrent problem in NLG evaluation: the field uses stable labels, but the operational meaning of those labels is not always stable.

The second finding is a persistent discrepancy between human judgments and automatic scores. This does not mean that automatic metrics are useless. Rather, it means that their interpretation should be narrow and evidence-based. Reiter and Belz (2009) had already questioned the validity of some automatic metrics for NLG, and the studies reviewed here show that the problem continues in neural generation. Metrics may reward lexical similarity, semantic closeness or learned preference patterns, but human raters often react to a wider set of features, including tone, missing context, unsupported claims and practical usefulness. When papers report only automatic scores, they risk presenting a partial picture of quality.

The third finding is the overuse of surface-level proxies. Overlap-based evaluation remains common because it is simple, familiar and computationally cheap. It is not wrong to use such metrics, especially for diagnostic comparison, but it is problematic when they are presented as full measures of quality. Sai et al. (2020) show that NLG systems have been evaluated through a wide range of metrics, many of which were borrowed from earlier tasks and then reused in new settings. This borrowing is useful, but it can become uncritical. An evaluator should ask what kind of error the metric can detect, what kind of valid output it may penalise, and what kind of failure it may ignore.

The fourth finding is that factuality has become a central quality dimension between 2018 and 2023. This is especially visible in summarization and long-form generation. Goodrich et al. (2019) examine factual accuracy in generated text, while Gabriel et al. (2020) study meta-evaluation of factuality in summarization. Later work, including FacTool, AlignScore, FActScore and factuality benchmarks, indicates a clear shift from reference similarity towards evidence-based checking (Chern et al., 2023; Min et al., 2023; Muhlgaay et al., 2023; Zha et al., 2023). This shift is methodologically important because

factual errors can be hidden by fluent wording. In applied settings, a factual error may have higher cost than a minor grammatical problem.

The fifth finding concerns the rise of large language models as evaluators. G-Eval uses GPT-4 for NLG evaluation and reports stronger alignment with human judgments than many traditional metrics (Liu, Iter, Xu, Wang, Xu, & Zhu, 2023). This approach is promising because large models can interpret instructions, apply rubrics and produce explanations. Nevertheless, it raises concerns. A model-based evaluator may inherit biases, show sensitivity to prompt wording, prefer certain styles, or give high scores to confident but unsupported text. It may also be difficult to reproduce if the underlying model changes. LLM-based evaluation should therefore be treated as an assisted human-evaluation tool or a learned metric requiring calibration, not as an unquestionable judge.

The sixth finding is the lack of a common taxonomy that integrates domain-specific and ethical dimensions. Many studies evaluate fluency and adequacy, fewer evaluate factuality, and fewer still evaluate bias, safety, fairness or user harm as part of quality. This creates a gap between research benchmarks and real deployment. A system used to assist a journalist, teacher, doctor, translator or public administrator must be judged not only by how natural the text sounds but also by whether it is accurate, traceable, non-discriminatory and fit for the institutional context. Evaluation methods have begun to move in that direction, but the integration remains uneven.

The final finding is that reporting practices remain inconsistent. Some papers provide detailed annotator instructions, examples, scales and agreement scores. Others give only a short statement that human evaluation was performed. The same problem appears in metric reporting: papers may list scores without explaining what each metric is expected to capture or why it is appropriate for the task. This weakens comparability across studies. For a field that depends heavily on benchmarks, the lack of reporting discipline is not a minor issue. It affects whether results can be interpreted, replicated and responsibly applied.

A Multilayered Framework for Quality Evaluation

On the basis of the review, this article proposes a multilayered framework for evaluating automatically generated language. The framework is built on a simple assumption: quality should be assessed at the same level at which the generated text can fail. A text can fail linguistically, semantically, factually, pragmatically or ethically. A single metric cannot reliably cover all these levels. The proposed framework therefore combines automatic indicators, structured human review and task-specific risk checks.

The first layer is task definition. Before choosing metrics, evaluators should state what the system is supposed to do, who will use it, what sources it may rely on and what type of error is most costly. The second layer is linguistic quality, covering fluency, grammar, readability and non-redundancy. Automatic tools may be useful here, but human sampling remains necessary for tone and naturalness. The third layer is semantic and task adequacy. This includes whether the output preserves the source meaning, selects the right content and satisfies the prompt. The fourth layer is discourse quality, including coherence, cohesion, structure and continuity across turns or paragraphs.

The fifth layer is factual and evidential reliability. For summarization, question answering, explanation and long-form generation, evaluators should check whether claims are supported by the source or by retrievable evidence. Metrics such as AlignScore, FacTool or FActScore can support this layer, but the final interpretation should consider the domain and the seriousness of error (Chern et al., 2023; Min et al., 2023; Zha et al., 2023). The sixth layer is human usefulness. It asks whether the generated text helps the intended user complete a task. The seventh layer is ethical and social risk, including bias, harmful stereotypes, privacy leakage, toxicity and misleading confidence. The last layer is

reporting transparency. Every evaluation should describe its criteria, metrics, raters, prompts, datasets and known limitations.

Table 2 presents the framework as a practical checklist. It is deliberately modular. A low-risk creative writing tool may require more attention to diversity and style, while a medical or legal summarization system would require stricter factual and expert review. The framework does not prescribe the same evaluation package for every system; instead, it asks evaluators to justify which layers are essential for the task and how each layer is measured.

Evaluation layer	Main question	Suggested evidence	Example indicators
Task definition	What should the system do and for whom?	Use-case description and risk profile	User group, task type, acceptable error threshold
Linguistic quality	Is the text readable and natural?	Automatic screening and human sampling	Grammar, fluency, readability, non-redundancy
Semantic adequacy	Does the output preserve intended meaning?	Reference comparison, expert review, rubric scoring	Adequacy, relevance, content coverage
Discourse coherence	Does the text work as a connected whole?	Human rating and discourse checks	Cohesion, logical order, turn consistency
Factual reliability	Are claims supported by evidence?	Factuality metrics, evidence checks, expert audit	Entailment, source alignment, atomic fact precision
Human usefulness	Does the output help the user complete the task?	User study, expert assessment, task success data	Helpfulness, actionability, satisfaction
Ethical and social risk	Could the output mislead or harm?	Bias and safety review, red-team prompts	Bias, toxicity, privacy, harmful stereotyping
Transparency	Can the evaluation be interpreted and repeated?	Full reporting of methods and limitations	Rater details, prompts, metrics, datasets, agreement

Table 2. A multilayered framework for NLG quality assessment.

Discussion

The review suggests that NLG evaluation is entering a more mature but also more demanding phase. The field no longer lacks metrics; it lacks agreement on when a metric is appropriate, how different forms of evidence should be combined, and how quality should be reported for real-world use. The popularity of automatic scores is understandable. Research communities need benchmarks, model developers need fast feedback and institutions need scalable quality checks. Yet the history of NLG evaluation shows that speed can produce a narrow view of performance. The more fluent generated text becomes, the more dangerous it is to rely on surface impressions.

For researchers, the main implication is that evaluation design should be treated as part of the scientific contribution, not as a routine appendix. A paper that introduces a new generation system should explain why its chosen criteria match the task. If the paper uses an automatic metric, it should state what that metric measures and what it does not measure. If it reports human evaluation, it should provide

enough detail for readers to understand the rating process. This is consistent with the best-practice recommendations of Van Der Lee et al. (2019) and the metric-evaluation concerns raised by Sai et al. (2021) and Ni'mah et al. (2023).

For practitioners, the implication is that quality assurance should be risk-based. A customer-service chatbot, a classroom writing assistant and a clinical note generator should not be evaluated with the same checklist or the same tolerance for error. Applied systems need minimum thresholds for linguistic clarity, factual reliability, user usefulness and safety. They also need procedures for monitoring performance after deployment, because user populations and input distributions change. Human-AI collaboration is central here. Human evaluators should not simply provide labels for metric training; they should help define what quality means in context.

For the wider NLP community, the key challenge is to build evaluation cultures that reward transparency. Model-based evaluators such as G-Eval may help reduce evaluation cost and improve alignment with human preferences, but they should be documented and audited like any other measurement instrument (Liu et al., 2023). Factuality tools and atomic evaluation methods can make errors more visible, but they should not be presented as final substitutes for expert judgment in high-stakes settings. A robust evaluation culture will accept that quality is multidimensional and that some dimensions require human interpretation.

Conclusion

This systematic review examined how quality in automatically generated language was defined and measured in research published between 2018 and 2023. Across 87 studies, three broad evaluation paradigms were identified: human-centred assessment, automatic metrics and task-specific composite measures. The synthesis shows that NLG quality is commonly associated with fluency, adequacy and coherence, but recent work increasingly adds factual consistency, usefulness, diversity, explainability and ethical risk. This expansion reflects a real change in the field: generated language is no longer evaluated only as text, but as language used in human tasks and institutional contexts.

The review also shows that no single evaluation method is sufficient. Human judgments remain indispensable, but they require careful protocol design and transparent reporting. Automatic metrics are valuable for scale and comparison, but they must be interpreted according to what they actually measure. Composite and factuality-focused methods respond to important weaknesses in traditional metrics, yet they also need validation and domain-sensitive use. The proposed multilayered framework offers a practical way to combine these forms of evidence. It encourages evaluators to begin with the task, identify the likely failure levels, and then select human and automatic methods that match those levels.

The future of NLG evaluation should move away from the search for a perfect universal score. Quality in generated language is not one number. It is a structured judgment about form, meaning, evidence, usefulness and responsibility. Reliable evaluation will depend on combining metrics with human expertise, documenting the limits of each method and adapting assessment to the risks of each application.

Declaration of AI Refined: This document has benefited from the application of AI-driven tools, including ChatGPT, Grammarly and Paperpal Chat, to refine its linguistic aspects. These tools were utilized to correct grammar and spelling and improve the overall writing style. It is acknowledged that the use of these technologies may introduce certain AI-generated linguistic patterns. However, the core intellectual content, data, and conclusions presented remain the sole work of the authors.

References

- Chatzikoumi, E. (2019). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26, 137-161. <https://doi.org/10.1017/s1351324919000469>
- Chern, E., Chern, S., Chen, S., Yuan, W., Feng, K., Zhou, C., et al. (2023). FacTool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *ArXiv*, abs/2307.13528. <https://doi.org/10.48550/arxiv.2307.13528>
- Clinciu, M., Eshghi, A., & Hastie, H. (2021). A study of automatic metrics for the evaluation of natural language explanations. *ArXiv*, abs/2103.08545. <https://doi.org/10.18653/v1/2021.eacl-main.202>
- Gabriel, S., Celikyilmaz, A., Jha, R., Choi, Y., & Gao, J. (2020). GO FIGURE: A meta evaluation of factuality in summarization. 478-487. <https://doi.org/10.18653/v1/2021.findings-acl.42>
- Goodrich, B., Rao, V., Saleh, M., & Liu, P. (2019). Assessing the factual accuracy of generated text. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. <https://doi.org/10.1145/3292500.3330955>
- Hashimoto, T., Zhang, H., & Liang, P. (2019). Unifying human and statistical evaluation for natural language generation. 1689-1701. <https://doi.org/10.18653/v1/n19-1169>
- Leiter, C., Lertvittayakumjorn, P., Fomicheva, M., Zhao, W., Gao, Y., & Eger, S. (2022). Towards explainable evaluation metrics for natural language generation. *ArXiv*, abs/2203.11131. <https://doi.org/10.48550/arxiv.2203.11131>
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. *ArXiv*, abs/2303.16634. <https://doi.org/10.18653/v1/2023.emnlp-main.153>
- Manning, E., Wein, S., & Schneider, N. (2020). A human evaluation of AMR-to-English generation systems. 4773-4786. <https://doi.org/10.18653/v1/2020.coling-main.420>
- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P., et al. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *ArXiv*, abs/2305.14251. <https://doi.org/10.48550/arxiv.2305.14251>
- Montahaei, E., Alihosseini, D., & Baghshah, M. (2019). Jointly measuring diversity and quality in text generation models. *ArXiv*, abs/1904.03971. <https://doi.org/10.18653/v1/w19-2311>
- Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., et al. (2023). Generating benchmarks for factuality evaluation of language models. 49-66. <https://doi.org/10.48550/arxiv.2307.06908>
- Ni'mah, I., Fang, M., Menkovski, V., & Pechenizkiy, M. (2023). NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. 1240-1266. <https://doi.org/10.48550/arxiv.2305.08566>
- Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35, 529-558. <https://doi.org/10.1162/coli.2009.35.4.35405>
- Rivera-Trigueros, I. (2021). Machine translation systems and quality assessment: A systematic review. *Language Resources and Evaluation*, 56, 593-619. <https://doi.org/10.1007/s10579-021-09537-5>
- Sai, A., Dixit, T., Sheth, D., Mohan, S., & Khapra, M. (2021). Perturbation CheckLists for evaluating NLG evaluation metrics. *ArXiv*, abs/2109.05771. <https://doi.org/10.18653/v1/2021.emnlp-main.575>

- Sai, A., Mohankumar, A., & Khapra, M. (2020). A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)*, 55, 1-39. <https://doi.org/10.1145/3485766>
- Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M., & Raffel, C. (2022). Evaluating the factual consistency of large language models through news summarization. 5220-5255. <https://doi.org/10.18653/v1/2023.findings-acl.322>
- Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. 355-368. <https://doi.org/10.18653/v1/w19-8643>
- Zha, Y., Yang, Y., Li, R., & Hu, Z. (2023). AlignScore: Evaluating factual consistency with a unified alignment function. 11328-11348. <https://doi.org/10.48550/arxiv.2305.16739>
- Zhang, T., Kishore, V., Wu, F., Weinberger, K., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. *ArXiv*, abs/1904.09675
- Zhou, W., & Xu, K. (2020). Learning to compare for better training and evaluation of open domain natural language generation models. 9717-9724. <https://doi.org/10.1609/aaai.v34i05.6521>
- Zhu, W., & Bhat, S. (2020). GRUEN for evaluating linguistic quality of generated text. *ArXiv*, abs/2010.02498. <https://doi.org/10.18653/v1/2020.findings-emnlp.9>

Author's biography

Tahri Abderrahmen (b. 1983) is an Inspector of National Education. He holds a master's in Linguistics (2018). Actually, he is a PhD student (Applied Linguistics) at Mohamed Boudiaf University, Msila. His career spans from his initial licence in 2006 to his current supervisory role.



مَشَقَات

