NLG Evaluation: Past, Present, Future

Ehud Reiter
Dept of Computing Science
University of Aberdeen
Aberdeen, UK
e.reiter@abdn.ac.uk

Abstract

Natural Language Generation (NLG) evaluation has changed dramatically since 1990, and will continue to evolve in the future. In 1990, when NLG had close ties to linguistics, there was very little formal experimental evaluation in the modern sense. In 2026, when NLG is closely linked to machine learning, experimental evaluation is expected and indeed fundamental to research. Many evaluation techniques were developed over this period, including most recently LLM-as-Judge. I expect NLG evaluation will continue to evolve in the future. In particular, impact, qualitative, and safety evaluation will become more important as large numbers of people routinely use NLG technology.

1 Introduction

The evaluation of Natural Language Generation (NLG) systems has changed dramatically over my career. In 1990, when I got my PhD in NLG, most NLG research papers did not include a quantitative experimental evaluation of a research question. By 2026, NLG research papers are expected to include structured experimental evaluations of hypotheses, although the quality and validity of these evaluations is variable. I expect that by 2036, impact, safety, and qualitative evaluations will be much more important, because NLG technology will be widely used by large numbers of people. Table 1 summarises my view of NLG evaluation at different points in time.

2 NLG Evaluation in the Past

2.1 1990: Little quantitative experimental evaluation

The International NLG (INLG) conference in 1990 had 25 papers. *None* of them included a structured quantitative hypothesis test. Instead, these papers mostly presented an algorithm, technique, resource, or system, and justified it on engineering or linguistic criteria. For example, McCoy et al. (1990) proposed combining tree-adjoining and systemic grammars, and justified this by arguing that their approach did a better job of handling long-distance dependencies (linguistics) and also makes it easier to build grammars (engineering). Their argument was qualitative, no numbers were given.

In the broader NLP world, the speech recognition community had adapted the idea of quantitative comparisons of the performance of systems (Waibel and Lee, 1990), but this was unusual in the rest of NLP. Perhaps the most important NLP paper in 1990 was Brown et al. (1990), which introduced statistical machine translation, but even it did not provide quantitative comparisons of the sort we expect in 2026.

2.2 2000: Wide range of evaluation techniques

INLG in 2000 had 38 papers, and these included many different kinds of evaluation, as well as one of the first paper that was *about* evaluation (Bangalore et al., 2000). Types of evaluation included

- Human evaluation (Cheng and Mellish, 2000)
- Metric-based evaluation (Minnen et al., 2000)
- Task-based evaluation (Carenini, 2000)

There were also papers which continued to assess their contribution using engineering or linguistic arguments, as in 1990.

In short, by 2000 experimental evaluations was recognised as being important. However there were no widely accepted standard evaluation techniques in NLG.

A similar mix was seen at larger NLP events such as ACL. Evaluation was clearly regarded as important, but many techniques were being tried. The broader NLP community focused more on metric-based evaluation, including Gildea and Jurafsky

year	NLG evaluation	example paper and its evaluation
1990	non-quantitative evaluation, often using linguistic or engineering arguments	McCoy et al. (1990): qualitative argument that their grammatical approach handles long-distance dependencies better
2000	wide mix of different techniques, including metrics, human ratings, and task-based	Cheng and Mellish (2000): use human ratings to evaluate different ways of expressing causal and temporal relationships
2010	standardised evaluation techniques and shared tasks based on these	Belz and Kow (2010): results of GREC shared task on generating referring expressions
2020	research on evaluation becomes an important research area	Howcroft et al. (2020): gives recommendations for reporting human evaluations, based on meta-analysis of published evaluations
2026	LLM-as-Judge, annotation by human experts, safety, interdisciplinary	Bean et al. (2026): use medical evaluation techniques to assess system that answers health queries
2036	impact, qualitative, safety evaluation	*not yet written*

Table 1: NLG evaluation over the years

(2000), which won a Test of Time award. However ACL in this period also included papers reporting complex task-based evaluations (Mani et al., 1999; Reiter et al., 2001).

2.3 2010: Shared tasks and standard evaluations

INLG in 2010 had 37 papers, many of which were shared task submissions. Shared tasks (such as the GREC challenge for generating referring expressions (Belz and Kow, 2010)) had become an accepted part of NLG as well as NLP research, and used metrics and/or human evaluations to evaluate the performance of submissions. Some papers also began to describe evaluations in considerable detail (Murray et al., 2010).

The wider NLP community had embraced ngram-based metrics for evaluation of text production, and the BLEU and ROUGE metrics had effectively become standards. Papers in machine translation were expected to use BLEU, and papers in summarisation were expected to use ROUGE.

Human evaluation had become unusual in ACL conferences, although the annual WMT shared task continued to use it. The NLG community, however, insisted on using human evaluations, and could point to papers which suggested that metrics were not reliable in NLG (Reiter and Belz, 2009). When doing human evaluations, most researchers either used Likert scales or asked subjects to rank a set of texts by a quality criteria; these became standard techniques for human evaluation of generated texts.

2.4 2020: Evaluation is important research area

INLG in 2020 had 46 papers (it has not seen the exponential growth that ACL has had in recent years). Perhaps the most notable change compared to 2010 was that evaluation has become a very important part of the community's research agenda. Indeed both of the INLG2020 best papers were about evaluation (Belz et al., 2020; Dušek and Kasner, 2020), and there were several other papers about evaluation methodology (Howcroft et al., 2020; Thomson and Reiter, 2020) in INLG2020.

The wider NLP community also placed increasing importance on evaluation as an important research theme. For example the ACL 2020 best paper was about testing (Ribeiro et al., 2020), and one of the two honourable mention papers was about evaluation (Mathur et al., 2020).

In short, evaluation was now not just something which researchers had to do, but also an important research topic in its own right.

3 NLG Evaluation in 2026

Reiter (2025a) summarised NLG evaluation in 2025, including links to papers that gave best practice suggestions. Large language model (LLM) technology had become widespread and this had changed NLG evaluation and introduced new challenges.

3.1 Evaluation challenges from LLMs

There are many challenges in evaluating LLMs, including the following.

Higher quality generated texts: Texts produced by LLMs are usually higher quality than texts produced by previous technologies (rule-based, LSTM), and can in some cases be human quality, or even better-than-human. This means that many traditional evaluation techniques, such as metrics that compare generated texts against human-written reference texts, no longer work well. If we expect a generated text to be better than human, then evaluating it by comparing it to a human-written reference text does not make sense.

Semantic and pragmatic evaluations: Texts produced by LLMs are almost always fluent and readable, so evaluating readability is less useful. Instead, there is more emphasis on evaluating semantic and pragmatic quality criteria (Reiter, 2025a), such as accuracy/hallucinations, omissions, and contextual appropriateness.

Data contamination: Since LLMs are trained on the Internet, an evaluation that uses Internet data may not mean much, since the LLM may have memorised the test data (Balloccu et al., 2024).

Worst-case and safety evaluation: The growing real-world usage of LLMs in safety-critical contexts such as medicine (where flawed texts could harm patients) means that we need techniques that evaluate 'worst-case' performance of LLMs (Reiter, 2025a). If a medical LLM gives good output in 99.9% of cases but harmful output in 0.1% of cases, this is not acceptable.

Interdisciplinary interest and usage: The growing real-world usage of LLMs means that other disciplines (such as medicine and law) want LLM-based NLG systems to be evaluated using their methodologies and expectations (Duggan et al., 2025).

3.2 Changes in NLG evaluation

The above challenges have changed the way NLG is evaluated.

LLM as Judge: The above problems have stimulated interest in reference-free metrics which work for semantic and pragmatic quality criteria, including in particular using LLMs to evaluate the quality of texts produced by other LLMs (Gao et al., 2025); this is called *LLM as Judge*. This seems to work well in some cases but not others; unfortunately many researchers use LLM evaluators without checking that they are effective in their use case.

Human evaluation using expert annotations: Human evaluations in NLG have traditionally used Likert-type rating scales. This seems to work less well when evaluating semantic and pragmatic problems in high-quality LLM texts, especially with crowdworkers (who may cheat by using LLMs to do the evaluation task (Asher et al., 2026)). We are seeing more human evaluations that instead ask knowledgable people to annotate specific problems in a generated text (Thomson et al., 2023).

Private test data: In 2020, test data sets were typically published (e.g., on GitHub repos), which made replication easier. But in 2026, data contamination concerns mean that test data is sometimes not published or shared.

Safety evaluations: Many techniques have been proposed for safety evaluation. This area is heavily influenced by cyber security, and includes techniques for risk analysis (such as red teaming), risk mitigation (e.g., monitoring), and risk governance (such as incident reporting) (Bengio et al., 2026).

Interdisciplinary evaluations: High-quality evaluations of NLG systems are appearing in other fields, notably medicine, that use medical evaluation techniques such as randomised controlled trials (which are very rare in the NLP literature (Reiter, 2025b)). Sometimes these give different results from classical NLP evaluation, which raises important questions about the best way to evaluate NLG .

3.3 Ongoing challenges for NLG evaluation

The new evaluation techniques described above are being adopted by many researchers and help in addressing some of the new evaluation challenges of LLMs. But there are many problems and concerns that still need to be addressed. These include

- *Experimental rigour:* Unfortunately, many experiments are poorly designed, poorly executed, or distorted by bugs (Thomson et al., 2024).
- *Replicability:* Many experiments cannot be replicated, in part because their authors do not support replication (Belz et al., 2023).
- *Construct validity:* Many evaluation techniques, especially benchmarks, do not measure what they claim to measure (Bean et al., 2025).

- *Cheating:* LLMs engage in behaviour such as reward hacking (Arx et al., 2025), which is essentially cheating. Asadi et al. (2026) show that LLMs can get very high benchmark scores even when input data is withheld, by picking up on subtle clues in the wording of questions in the benchmark.

- *Commercial bias and incentives:* A lot of evaluation research and development is funded by AI companies such as OpenAI, who have an interest in ensuring that their systems do well on these evaluations (Cheng et al., 2025).

- *Evolving benchmarks:* New evaluation benchmarks are constantly being proposed, and existing benchmarks often become saturated (Akhtar et al., 2026) and hence useless. It is difficult for many researchers to stay up-to-date on the best benchmark to use.

A generic challenge is that the research culture in NLP is often not very supportive of high quality evaluation. Many people feel pressure to publish large numbers of papers, and reviewers often show limited interest in quality of data sets, validity of evaluation metrics, experimental rigour, etc. This encourages researchers to conduct 'quick and dirty' evaluations.

4 NLG Evaluations in the Future

What will NLG evaluation be like in ten years time (2036)? The above challenges will hopefully be addressed, but more generally we also need to go beyond measuring performance on a test set, which dominates NLG and NLP evaluation in 2026. If we care about how our technology affects the real world, we need to do more of the following:

- Directly measure the real-world **impact** of NLG systems.

- Use **qualitative** techniques to get insights about the effectiveness of our techniques in messy and complex real-world contexts.

- Analyse what happens in worst-case or adversarial contexts, especially for **safety** criteria.

These techniques will help ensure that evaluation is relevant and meaningful in a future world where NLG is a widely-used technology.

Note that impact, qualitative, and safety evaluations are not new, they are already being done in 2026 to a limited degree in NLG; they are much more common in Medicine, perhaps because medical research has had real-world consequences for decades or indeed centuries. So the challenge for the NLG community is to embrace these types of evaluation and learn how to do them well in an NLG context.

The spread of more types of NLG evaluation may lead to the evolution of evaluation frameworks, which show how different types of evaluation can be combined to obtain a holistic understanding of what a system can do (Reddy et al., 2021).

4.1 Impact evaluation

As discussed by Reiter (2025a), there is very little evaluation of real-world impact in the NLP and NLG research literature, by which we mean how real-world usage of an NLG system changes key performance indicators (KPIs). As NLG technology improves and becomes more widely used, we need more impact studies, especially if we want to measure utility in messy real-world contexts.

A good example is Bean et al. (2026), which measured how well LLMs can respond to health queries based on scenarios. LLMs do well at this task if given the scenario directly, or if they interact with an LLM-simulated user. However, if they interact with human users (who often communicate in a confused way), their performance is much worse. Hence if we want to genuinely evaluate how well an LLM can respond to health queries, we need to measure what happens when real people interact with the LLM. Ideally, this should be based on real patients asking about their health problems (Brodeur et al., 2026).

There are many ways to evaluate impact, including randomised controlled trials (RCT), A/B tests, before-and-after (pre-post) studies, and observational studies (Reiter, 2025b). By 2036, we hope that such evaluations will much more common. Most NLG evaluations will probably still use simpler and cheaper techniques, but a significant number will evaluate real-world impact.

4.2 Qualitative evaluation

Evaluation in NLG and NLP is almost always quantitative, and typically uses statistical hypothesis testing. Such evaluation is very important, but should be supplemented by qualitative evaluation, which can provide additional insights which are very useful in complex real-world contexts (Greenhaigh and Taylor, 1997; Tisdell et al., 2025).

Some qualitative evaluation techniques are already used in NLG, including error analysis (van Miltenburg et al., 2023) and analysis of free-text comments from participants (van der Lee et al., 2021). But many other techniques are rarer, including data collection techniques such as focus groups (Sun et al., 2026) and (semi-)structured interviews (Zhou et al., 2022), and analysis techniques such as thematic analysis (Guest et al., 2011) and content analysis (Sambaraju et al., 2011).

As NLG systems become more capable and are used in a wider variety of complex contexts, we expect that qualitative evaluation and insights will become more important, especially since many quantitative results will quickly become dated as newer models are released.

4.3 Safety evaluation

Safety evaluation is not new, it is a rapidly growing area of evaluation, which looks at whether AI systems can harm individuals (for example by encouraging suicide[1] or giving dangerous medical advice (Bickmore et al., 2018)) or society (e.g., by empowering hackers or terrorists) (Bengio et al., 2026).

We expect that safety will become one of the main foci of evaluation research. Ultimately, performance evaluation is of interest primarily to companies and academics who develop NLG technology, whereas safety evaluation is of interest to everyone who *uses* NLG technology, which is a much larger number of people. Safety evaluation is probably more important to society than performance evaluation. Indeed, governments have begun to impose safety standards on AI systems[2], and this may lead to formal government involvement in AI evaluation methodology.

Safety evaluation is also more challenging than performance evaluation, because it is about worst-case behaviour, and behaviour under adversarial attack (e.g., hackers trying to break into a system). Performance evaluations usually look at average case performance, so they can be computed based on a representative sample. Safety evaluation requires looking for misbehaviour everywhere, including edge cases, which are hard to predict for complex stochastic black box neural models. It will almost certainly require monitoring of the actual behaviour of deployed systems, as well as experiments on test data or test subjects.

[1] See https://www.thehumanlineproject.org/ stories, such as Badshah (2026)

[2] https://www.gov.uk/government/publications/generative-ai-product-safety-standards

5 Conclusion

NLG evaluation has changed dramatically between 1990 (mostly lingiustic evaluation) and 2026 (LLM-as-Judge and human annotation protocols). It continues to evolve, and the next ten years should be exciting, with more focus on impact, qualitative, and safety evaluation.

Acknowledgements

Many thanks to the anonymous reviewers, the members of the Aberdeen CLAN research group, and Saad Mahamood for their very helpful comments.

References

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, Jennifer Mickel, Leshem Choshen, Marek Šuppa, Jan Batzner, Jenny Chim, Jeba Sania, Yanan Long, Hossein A. Rahmani, Christina Knight, and 18 others. 2026. When ai benchmarks plateau: A systematic study of benchmark saturation. *Preprint*, arXiv:2602.16763.

Sydney Von Arx, Lawrence Chan, and Elizabeth Barnes. 2025. Recent frontier models are reward hacking. https://metr.org/blog/2025-06-05-recent-reward-hacking/.

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. 2026. Mirage: The illusion of visual understanding. *Preprint*, arXiv:2603.21687.

Michael W. Asher, Gillian Gold, Eason Chen, and Paulo F. Carvalho. 2026. Chatbots are undermining crowdsourced research in the behavioral sciences: Detecting artificial intelligence–assisted cheating with a keystroke-based tool. *Advances in Methods and Practices in Psychological Science*, 9(1):25152459261424723.

Nadeem Badshah. 2026. Teenager died after asking chatgpt for 'most successful' way to take his life, inquest told. *The Guardian*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93.

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 1–8, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, and 1 others. 2025. Measuring what matters: Construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703; presented at Neurips 2025*.

Andrew M Bean, Rebecca Elizabeth Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera-Gómez, Sara Hincapié M, Aruna S Ekanayaka, Lionel Tarassenko, Luc Rocher, and 1 others. 2026. Reliability of llms as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, pages 1–7.

Anja Belz and Eric Kow. 2010. The GREC challenges 2010: Overview and evaluation results. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

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, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Yoshua Bengio, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Ben Bucknall, Malcolm Murray, Rishi Bommasani, Stephen Casper, Tom Davidson, Raymond Douglas, David Duvenaud, Philip Fox, Usman Gohar, Rose Hadshar, Anson Ho, Tiancheng Hu, Cameron Jones, Sayash Kapoor, Atoosa Kasirzadeh, and 73 others. 2026. International ai safety report 2026. *Preprint*, arXiv:2602.21012.

Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. *Journal of medical Internet research*, 20(9):e11510.

Peter Brodeur, Jacob M. Koshy, Anil Palepu, Khaled Saab, Ava Homiar, Roma Ruparel, Charles Wu, Ryutaro Tanno, Joseph Xu, Amy Wang, David Stutz, Wei-Hung Weng, Hannah M. Ferrera, David Barrett, Lindsey Crowley, Jihyeon Lee, Spencer E. Rittner, Ellery Wulczyn, Selena K. Zhang, and 29 others. 2026. A prospective clinical feasibility study of a conversational diagnostic ai in an ambulatory primary care clinic. *Preprint*, arXiv:2603.08448.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Giuseppe Carenini. 2000. A task-based framework to evaluate evaluative arguments. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 9–16, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Hua Cheng and Chris Mellish. 2000. An empirical analysis of constructing non-restrictive NP modifiers to express semantic relations. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 108–115, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Zerui Cheng, Stella Wohnig, Ruchika Gupta, Samiul Alam, Tassallah Abdullahi, João Alves Ribeiro, Christian Nielsen-Garcia, Saif Mir, Siran Li, Jason Orender, and 1 others. 2025. Benchmarking is broken–don't let ai be its own judge. *arXiv preprint arXiv:2510.07575, presented at Neurips2025*.

Matthew J. Duggan, Julietta Gervase, Anna Schoenbaum, William Hanson, III Howell, John T., Michael Sheinberg, and Kevin B. Johnson. 2025. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. *JAMA Network Open*, 8(2):e2460637–e2460637.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. LLM-based NLG evaluation: Current status and challenges. *Computational Linguistics*, 51:661–687.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.

T Greenhaigh and Rod Taylor. 1997. Papers that go beyond numbers (qualitative research)'. *British Medical Journal*, 315(7110):740–743.

Greg Guest, Kathleen M MacQueen, and Emily E Namey. 2011. *Applied thematic analysis*. sage publications.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 1999. The TIPSTER SUMMAC text summarization evaluation. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–85, Bergen, Norway. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Kathleen F. McCoy, K. Vijay-Shanker, and Gijoo Yang. 1990. Using Tree Adjoining Grammars systemic framework in the. In *Proceedings of the Fifth International Workshop on Natural Language Generation*, Linden Hall Conference Center, Dawson, Pennsylvania. Association for Computational Linguistics.

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 201–208, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Sandeep Reddy, Wendy Rogers, Ville-Petteri Makinen, Enrico Coiera, Pieta Brown, Markus Wenzel, Eva Weicken, Saba Ansari, Piyush Mathur, Aaron Casey, and 1 others. 2021. Evaluation framework to guide implementation of ai systems into healthcare settings. *BMJ health & care informatics*, 28(1):e100444.

Ehud Reiter. 2025a. *Natural Language Generation*. Springer.

Ehud Reiter. 2025b. We should evaluate real-world impact. *Computational Linguistics*, 51(4):1419–1431.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ehud Reiter, Roma Robertson, A. Scott Lennox, and Liesl Osman. 2001. Using a randomised controlled clinical trial to evaluate an NLG system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 442–449, Toulouse, France. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Rahul Sambaraju, Ehud Reiter, Robert Logie, Andy Mckinlay, Chris McVittie, Albert Gatt, and Cindy Sykes. 2011. What is in a text and what does it do: Qualitative evaluations of an NLG system – the BT-nurse – using content analysis and discourse analysis. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 22–31, Nancy, France. Association for Computational Linguistics.

Mengxuan Sun, Ehud Reiter, Peter Murchie, Anne E Kiltie, George Ramsay, Lisa Duncan, and Rosalind Adam. 2026. Can chatgpt give holistic and accurate patient-centred information to oncology patients? a mixed-methods evaluation with stakeholders. *medRxiv*.

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.

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

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech and Language*, 80:101482.

E.J. Tisdell, S.B. Merriam, and H.L. Stuckey-Peyrot. 2025. *Qualitative Research: A Guide to Design and Implementation*. Wiley.

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 and Language*, 67:101151.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Stephanie Schoch, Craig Thomson, and Luou Wen. 2023. Barriers and enabling factors for error analysis in NLG research. *Northern European Journal of Language Technology*, 9.

Alexander Waibel and Kai-Fu Lee. 1990. *Readings in speech recognition*. Morgan Kaufmann.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.