Natural Language Generation

Emiel van Miltenburg[a] and Chenghua Lin[b]
[a]Tilburg University, Department of Cognition and Communication, Tilburg, The Netherlands
[b]University of Manchester, Department of Computer Science, Manchester, UK

1 Introduction

The term *Natural Language Generation* (NLG), in its broadest definition, refers to the study of systems that verbalize some form of information through natural language. That information could be stored in a large database or knowledge graph (in *data-to-text* applications), but NLG researchers may also study summarisation (*text-to-text*) or image captioning (*image-to-text*), for example. As a subfield of Natural Language Processing, NLG is closely related to other sub-disciplines such as Machine Translation (MT) and Dialog Systems. Some NLG researchers exclude MT from their definition of the field, since there is no *content selection* involved where the system has to determine *what to say*. Conversely, *dialog systems* do not typically fall under the header of Natural Language Generation since NLG is just one component of dialog systems (the others being Natural Language Understanding and Dialog Management). However, with the rise of Large Language Models (LLMs), different subfields of Natural Language Processing have converged on similar methodologies for the production of natural language and the evaluation of automatically generated text.

1.1 Relation to linguistics

Historically, NLG research has had close ties with pragmatics and psycholinguistics. The reason for this is that automatically generated text should generally be as fluent and natural as possible. However, there are often many different ways to communicate the same message. How to best formulate that message depends on the context.[1] By studying human language production, we can better understand how people formulate their utterances in different situations. Psycholinguistic models of speech production (e.g. Levelt 1989) also have clear parallels with classical NLG pipelines (see §3.1). All these models can be divided into two stages: conceptualisation (*what to say*) and formulation (*how to say it*).

2 Applications

NLG technology promises to automate or streamline writing tasks in many different domains. We discuss three examples below.

2.1 NLG and business

The commercial potential of NLG has long been recognized; the first company to commercialize NLG was founded in 1990 and several different companies have followed in their footsteps (Dale, 2020). Robert Dale marks 2012 as the year when NLG entered the mainstream media conscious, with the publication of a *Wired* article on automated journalism. Following the release of ChatGPT, a full decade later (November 30, 2022), we have now seen how automatic text generation has become fully mainstream, with different companies offering the general public paid subscriptions to Large Language Models via user-friendly interfaces.

NLG techniques have traditionally been used to convert tabular data into text. For example, generating financial reports,[2] or generating product titles and descriptions based on product specifications (Mathur et al., 2017; Zou et al., 2023). Following the rise of ChatGPT, we are now seeing an explosion of interest to use language models for all sorts of purposes, including content generation (writing informative texts), and the generation of email responses (see Tafesse & Wien 2024; Tafesse & Wood 2024 for more discussion).

2.2 NLG and journalism

Like with business applications, NLG techniques have traditionally been used in journalism to produce data-driven articles to report on topics such as weather reports, sports matches, earthquakes, and elections. These articles are relatively formulaic, making it easier to develop templates where relevant values (such as the magnitude of an earthquake) can be inserted (Gatt & Krahmer, 2018). Nowadays, journalists use generative AI throughout the reporting process (news gathering, news production, news verification, distribution, and moderation) for a much wider array of articles (Cools & Diakopoulos, 2024). Still, classical rule- and template-based approaches remain the most reliable solution for standalone text generation, since they do not suffer from hallucination (see §5.1). Moreover, as Reiter (2025, Chapter 6) notes, rule-based systems may be easier to modify (e.g. make small changes to the output) and maintain.

2.3 NLG in a medical context

There is a long history of Natural Language Generation in healthcare (see Cawsey et al. 1997 for an early survey). Applications range from medical report generation (presenting relevant clinical data to different audiences, e.g. Gatt et al. 2009) to clinical note generation (summarising doctor-patient interactions; e.g. Ben Abacha et al. 2023) and personalised decision support tools (e.g. Hommes et al. 2019). Challenges in the medical field include working with potentially unreliable sensor data (van Deemter & Reiter, 2018), and involving different stakeholders (doctors, patients, nurses) who often differ in background knowledge, reading level, technical ability, and socio-economic status. Medical NLG research is usually carried out in interdisciplinary teams, with experts not only assessing the quality of the output, but also looking at the potentially harmful impact that systems may have on their users (e.g. Balloccu et al. 2024).

[1]We can perhaps best see this in the literature on referring expression generation, see Krahmer & van Deemter (2012).

[2]Ehud Reiter (2018) notes that all major NLG companies are involved in financial reporting, because the sector is large and use cases are similar between different organisations, making it easier to capitalise on NLG technology than in areas where more customisation is needed.

3 Approaches

This section provides a brief overview of the different approaches to Natural Language Generation, focusing on the evolution from the classical NLG pipeline, through end-to-end approaches, to the use of (more general) large language models. (For an overview of all NLG approaches before LLMs, see Gatt & Krahmer 2018).

3.1 The classical NLG pipeline

In the classical NLG pipeline, text generation proceeds in different stages. Here are the steps proposed by Reiter & Dale (2000):

- **Content determination**: deciding what information should be communicated.
- **Document structuring**: deciding how to group and order different chunks of information.
- **Lexicalisation**: deciding what words, phrases, or syntactic constructions to use.
- **Referring expression generation**: deciding what words or phrases should be used for different named entities.
- **Aggregation**: deciding how to map the different information chunks to sentences and paragraphs.
- **Realisation**: converting the abstract representation of the text-to-be-generated into a real text.

These steps can be handled by different modules that may either use (often hand-written) rules, Machine Learning, or both. The initial steps of the pipeline are context-dependent, but the final stage (realisation) is universal. As such, different projects may re-use the same *realiser*.[3]

3.2 End-to-end generation

End-to-end generation drastically simplifies the problem of natural language generation by directly mapping data to text. All steps are implicitly handled by the same model that is developed using a data-driven approach (see Castro Ferreira et al. 2019 for a discussion). Popular challenges that can be resolved through this approach are the E2E dataset (Novikova et al., 2017) and the WebNLG challenge (e.g. Cripwell et al. 2023).[4] Taking WebNLG as an example, the goal is to render a set of RDF triples in natural language. Recent years have seen the focus shift away from English (as this problem has more or less been solved) and towards low-resource languages.

3.3 Using Large Language Models

Large Language Models (LLMs), such as ChatGPT (Ouyang et al., 2022) and LLaMA (Touvron et al., 2023), typically adopt transformer-based architectures (Vaswani et al., 2017), which rely on self-attention mechanisms to capture long-range contextual dependencies in text. The training of LLMs generally follows a two-stage training pipeline. The first stage, known as *pre-training*, is an unsupervised process in which models are trained on vast amounts of unlabelled textual data containing billions or even trillions of tokens from diverse sources such as web pages, books, and encyclopaedias. This stage allows LLMs to acquire and encode general linguistic and world knowledge. The second stage, known as *Supervised Fine-Tuning* (SFT), involves further training the model on labelled datasets containing explicit instruction-output pairs. SFT aims to enhance the model's ability to understand and follow human-provided instructions, ensuring its responses are more relevant, specific, and aligned with user expectations. Techniques such as reinforcement learning from human feedback (RLHF) and direct preference optimisation (DPO) can further refine the model's outputs, improving its safety, helpfulness, and alignment with human preferences (Xu et al., 2024).

Once trained, LLMs generate text through prompting, where users provide textual inputs that may include explicit instructions, context, or examples to guide the model's responses. Prompting enables LLMs to perform a wide range of tasks, such as summarisation, translation, question answering, and creative writing, without requiring additional task-specific fine-tuning. Advanced prompting techniques, such as chain-of-thought prompting that introduces intermediate reasoning steps within the prompt, can further enhance the model's performance on complex reasoning and generation tasks (see Liu et al. (2023) for a survey).

4 Evaluation

Evaluating the quality of automatically generated text remains a fundamental challenge in natural language processing (NLP). As text generation models continue to improve, the difficulty of evaluating their outputs has also increased. In many cases, evaluating generated text is now as challenging as generating it. This difficulty arises due to the inherent variability in natural language—multiple outputs can be equally valid despite differing in lexical choice, structure, or even semantics. For instance, in machine translation, multiple translations of a given sentence may convey the same meaning while varying in word choice and phrasing. The challenge becomes even more pronounced in open-domain dialogue generation, where a single input may lead to many plausible responses with different semantics, i.e. the "one-to-many" problem (Zhao et al., 2023, 2024). Similarly, summarisation often involves subjectivity, as multiple summaries can be correct depending on the focus and interpretation of the content. Additionally, certain text generation tasks, such as humour or metaphor, require evaluators to account for demographic and cultural differences, further complicating the assessment process (Loakman et al., 2023; Wang et al., 2024). We can rougly distinguish two kinds of evaluation: using human evaluation studies, or through automatic metrics (see Celikyilmaz et al. 2020 for a survey).

4.1 Human evaluation

In human evaluation studies, participants answer questions about one or more texts (van der Lee et al., 2021). A common approach is to rate the quality of generated texts on a five-point scale (Amidei et al., 2019). There is a wide array of different dimensions that may be relevant to assess the quality of a text, for example: *fluency, correctness, completeness*, and *appropriateness* (see Belz et al. 2020 for a general taxonomy). For quality dimensions that relate to intrinsic properties of the texts, human evaluation studies are similar to traditional experiments that one may also encounter in other subfields of linguistics. For quality dimensions that are more closely related to the use of an NLG system in context, techniques from the field of human-computer interaction (such as contextual interviews,

[3]Among existing realisers, SimpleNLG (Gatt & Reiter, 2009) is a popular choice that has been translated into several different languages.
[4]Although note that these challenges do not require any signal analysis or content determination.

think-aloud tasks, or usability surveys) may be used. Finally, authors may also choose to carry out an *error analysis*, where authors identify and categorise instances where something went wrong in the generated text (van Miltenburg et al., 2021).

4.2 Automatic evaluation

While human evaluation remains the gold standard in NLG research, it can be costly and time-consuming. Therefore, different researchers have developed different ways to assess the output of NLG systems along different dimensions (e.g. *fluency, grammaticality, correctness*). Celikyilmaz et al. (2020) distinguish two kinds of automatic evaluation:

1. **Untrained automatic metrics** compare generated texts with a set of reference texts, through different similarity measures.
2. **Machine-learned metrics** are based on machine-learned models, and serve as automatic equivalents to human judges.

Recent years have seen rapid developments in the second category, with the *LLM-as-a-judge* paradigm quickly gaining popularity after showing impressive results on different benchmarks (e.g. Zheng et al. 2023). For an overview of currently used tasks and evaluation measures, readers are referred to the recurring GEM shared task (e.g. Mille et al. 2024), which is the most comprehensive community effort to assess NLG models on a wide array of different tasks.[5]

5 Challenges

5.1 Factuality versus fluency?

There is an interesting tension between factuality and fluency of automatically generated texts. Generally speaking, traditional rule-based approaches result in texts that are factually correct but not always fluent, while neural approaches tend to produce texts that are fluent but not always factually correct.[6] This is related to the problem of hallucination (see Huang et al. 2025 for a survey). It is currently unclear how to guarantee that NLG output is both fluent *and* 100% factual.

5.2 Long Text Generation

Traditionally, most text generation tasks have focused on producing relatively short outputs, such as weather reports or summaries that span a few dozen to a few hundred words. However, there is growing interest in developing models capable of generating much longer texts, often extending to thousands of words. For example, recent efforts such as AI Scientist (Lu et al., 2024) have demonstrated the ability to generate entire scientific papers. By incorporating structured scientific knowledge (e.g., experimental results), this framework can draft papers that adhere to domain-specific requirements, integrating relevant citations and conforming to disciplinary norms. Similarly, LongWriter (Bai et al., 2024) addresses long-text generation across various domains, including academic publications and monographs. It employs hierarchical attention mechanisms and fine-tuning strategies to maintain thematic consistency and produce structured arguments across extended outputs. Despite notable advancements, achieving coherence and structure in long-form content generation remains a significant challenge. This challenge stems from key issues such as the scarcity of high-quality long-form text data in the supervised fine-tuning phase, which limits the model's ability to learn effective writing patterns. Moreover, the alignment phase based on RLHF or DPO introduces complexities in reward modeling and evaluation, where assessing attributes like narrative flow, logical consistency, and thematic coherence becomes increasingly difficult as the text lengthens.

5.3 Evaluation

NLG evaluation is a hotly debated topic. Earlier studies found that there is a widespread confusion about the terminology that is used to refer to different quality dimensions (Howcroft et al., 2020). Human evaluations may always not be reproducible (Belz et al., 2023), which may be partly due to past reporting standards (Howcroft et al., 2020; Shimorina & Belz, 2022; Gehrmann et al., 2023), but we also have to consider the inherent difficulty of quantifying the objective quality of a text (if such a thing even exists). More research is needed to deepen our understanding of (the interaction between) different quality dimensions that are involved in the process of text appreciation (van Miltenburg et al., 2020).

5.4 Reproducibility

An open challenge in NLG research is how to ensure reproducible results. In other words: being able to obtain the same results as those reported in earlier studies. Although this holds for both the output generated by NLG software and the results of the evaluation procedure, most attention has gone to the repeatability of human evaluation studies (Belz et al., 2023).

5.5 Ethics: the social impact of NLG

As NLG technology is both increasingly powerful *and* increasingly widespread, we also have to contend with the real-world implications of our work. A recent survey (van Miltenburg, 2025) provides an overview of dual use issues that can arise from our research, and Solaiman et al. (2024) offer a broad taxonomy of areas that may be impacted by generative AI systems. Handling these issues requires continuous effort from both individual researchers and the community as a whole.

6 Conclusion

This article provided an overview of the field of Natural Language Generation. Due to space constraints, we have focused on the application side of the field (i.e. building and evaluating systems that convert data to text). Readers who would like to learn more about this topic can read the survey by Gatt & Krahmer (2018) or the recent book by Reiter (2025).

NLG researchers have always found inspiration in the way humans produce language. Although we have not focused on this topic, we encourage readers to also explore the more cognitively oriented work that studies human language production in context, and work that compares human and automatic language production.

[5] For an indication of the popularity of current metrics, see Schmidtova et al. 2024. But note that single metrics are always reductive; no individual metric can fully capture all different quality dimensions.

[6] Though see van Deemter & Reiter 2018, where the authors show how all types of NLG systems may deviate from the truth in their outputs. To some extent this is unavoidable in situations where the system also has to carry out signal analysis (i.e. interpreting the input, with a risk of misinterpretation) and content selection (meaning that the system cannot provide 'the whole truth').

Acknowledgments

See Also: article title article title

Bibliography

Amidei, J., Piwek, P., & Willis, A. (2019, October–November). The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. In K. van Deemter, C. Lin, & H. Takamura (Eds.), *Proceedings of the 12th international conference on natural language generation* (pp. 397–402). Tokyo, Japan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W19-8648/ doi: 10.18653/v1/W19-8648.

Bai, Y., Zhang, J., Lv, X., Zheng, L., Zhu, S., Hou, L., ... Li, J. (2024). *Longwriter: Unleashing 10,000+ word generation from long context llms.* Retrieved from https://arxiv.org/abs/2408.07055

Balloccu, S., Reiter, E., Li, K. J.-H., Sargsyan, R., Kumar, V., Reforgiato, D., ... Dusek, O. (2024, November). Ask the experts: sourcing a high-quality nutrition counseling dataset through human-AI collaboration. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 11519–11545). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.findings-emnlp.674/ doi: 10.18653/v1/2024.findings-emnlp.674.

Belz, A., Mille, S., & Howcroft, D. M. (2020, December). Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In B. Davis, Y. Graham, J. Kelleher, & Y. Sripada (Eds.), *Proceedings of the 13th international conference on natural language generation* (pp. 183–194). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.inlg-1.24/ doi: 10.18653/v1/2020.inlg-1.24.

Belz, A., Thomson, C., Reiter, E., & Mille, S. (2023, July). Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 3676–3687). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.findings-acl.226/ doi: 10.18653/v1/2023.findings-acl.226.

Ben Abacha, A., Yim, W.-w., Michalopoulos, G., & Lin, T. (2023, July). An investigation of evaluation methods in automatic medical note generation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 2575–2588). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.findings-acl.161/ doi: 10.18653/v1/2023.findings-acl.161.

Castro Ferreira, T., van der Lee, C., van Miltenburg, E., & Krahmer, E. (2019, November). Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 552–562). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1052/ doi: 10.18653/v1/D19-1052.

Cawsey, A. J., Webber, B. L., & Jones, R. B. (1997, 11). Natural language generation in health care. *Journal of the American Medical Informatics Association*, *4*(6), 473-482. Retrieved from https://doi.org/10.1136/jamia.1997.0040473 doi: 10.1136/jamia.1997.0040473.

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. *CoRR*, *abs/2006.14799*(), . Retrieved from https://arxiv.org/abs/2006.14799

Cools, H., & Diakopoulos, N. (2024). Uses of generative ai in the newsroom: Mapping journalists' perceptions of perils and possibilities. *Journalism Practice*, *0*(0), 1–19. Retrieved from https://doi.org/10.1080/17512786.2024.2394558 doi: 10.1080/17512786.2024.2394558.

Cripwell, L., Belz, A., Gardent, C., Gatt, A., Borg, C., Borg, M., ... Soto Martinez, W. (2023, September). The 2023 WebNLG shared task on low resource languages. overview and evaluation results (WebNLG 2023). In A. Gatt et al. (Eds.), *Proceedings of the workshop on multimodal, multilingual natural language generation and multilingual webnlg challenge (mm-nlg 2023)* (pp. 55–66). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.mmnlg-1.6/

Dale, R. (2020). Natural language generation: The commercial state of the art in 2020. *Natural Language Engineering*, *26*(4), 481–487. doi: 10.1017/S135132492000025X.

Gatt, A., & Krahmer, E. (2018, January). Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *J. Artif. Int. Res.*, *61*(1), 65–170.

Gatt, A., Portet, F., Reiter, E., Hunter, J., Mahamood, S., Moncur, W., & Sripada, S. (2009). From data to text in the neonatal intensive care unit: Using nlg technology for decision support and information management. *Ai Communications*, *22*(3), 153–186.

Gatt, A., & Reiter, E. (2009, March). SimpleNLG: A realisation engine for practical applications. In E. Krahmer & M. Theune (Eds.), *Proceedings of the 12th European workshop on natural language generation (ENLG 2009)* (pp. 90–93). Athens, Greece: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W09-0613/

Gehrmann, S., Clark, E., & Sellam, T. (2023, May). Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, *77*(), 103–166. Retrieved from http://dx.doi.org/10.1613/jair.1.13715 doi: 10.1613/jair.1.13715.

Hommes, S., van der Lee, C., Clouth, F., Vermunt, J., Verbeek, X., & Krahmer, E. (2019, October–November). A personalized data-to-text support tool for cancer patients. In K. van Deemter, C. Lin, & H. Takamura (Eds.), *Proceedings of the 12th international conference on natural language generation* (pp. 443–452). Tokyo, Japan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W19-8656/ doi: 10.18653/v1/W19-8656.

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., ... Rieser, V. (2020, December). Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In B. Davis, Y. Graham, J. Kelleher, & Y. Sripada (Eds.), *Proceedings of the 13th international conference on natural language generation* (pp. 169–182). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.inlg-1.23/ doi: 10.18653/v1/2020.inlg-1.23.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... Liu, T. (2025, January). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, *43*(2), . Retrieved from https://doi.org/10.1145/3703155 doi: 10.1145/3703155.

Krahmer, E., & van Deemter, K. (2012, March). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218. Retrieved from https://aclanthology.org/J12-1006/ doi: 10.1162/COLI_a_00088.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press. Retrieved from http://dx.doi.org/10.7551/mitpress/6393.001.0001 doi: 10.7551/mitpress/6393.001.0001.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, *55*(9), . Retrieved from https://doi.org/10.1145/3560815 doi: 10.1145/3560815.

Loakman, T., Maladry, A., & Lin, C. (2023). The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 6676–6689). Retrieved from https://aclanthology.org/2023.findings-emnlp.444/ doi: 10.18653/v1/2023.findings-emnlp.444.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). *The ai scientist: Towards fully automated open-ended scientific* *discovery.* Retrieved from https://arxiv.org/abs/2408.06292

Mathur, P., Ueffing, N., & Leusch, G. (2017, September). Generating titles for millions of browse pages on an e-commerce site. In J. M. Alonso, A. Bugarín, & E. Reiter (Eds.), *Proceedings of the 10th international conference on natural language generation* (pp. 158–167). Santiago de Compostela, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W17-3525/ doi: 10.18653/v1/W17-3525.

Mille, S., Sedoc, J., Liu, Y., Clark, E., Axelsson, A. J., Clinciu, M. A., ... Zhang, L. (2024, September). The 2024 GEM shared task on multilingual data-to-text generation and summarization: Overview and preliminary results. In S. Mille & M.-A. Clinciu (Eds.), *Proceedings of the 17th international natural language generation conference: Generation challenges* (pp. 17–38). Tokyo, Japan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.inlg-genchal.2/

Novikova, J., Dušek, O., & Rieser, V. (2017, August). The E2E dataset: New challenges for end-to-end generation. In K. Jokinen, M. Stede, D. DeVault, & A. Louis (Eds.), *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue* (pp. 201–206). Saarbrücken, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W17-5525/ doi: 10.18653/v1/W17-5525.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 36th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Reiter, E. (2018). *Where is nlg most successful commercially?* (Published on Ehud Reiter's blog: `https://ehudreiter.com/2018/10/30/most-successful-commercial-nlg/`)

Reiter, E. (2025). *Natural language generation.* Springer Nature Switzerland. Retrieved from http://dx.doi.org/10.1007/978-3-031-68582-8 doi: 10.1007/978-3-031-68582-8.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems.* Cambridge University Press.

Schmidtova, P., Mahamood, S., Balloccu, S., Dusek, O., Gatt, A., Gkatzia, D., ... Sivaprasad, A. (2024, September). Automatic metrics in natural language generation: A survey of current evaluation practices. In S. Mahamood, N. L. Minh, & D. Ippolito (Eds.), *Proceedings of the 17th international natural language generation conference* (pp. 557–583). Tokyo, Japan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.inlg-main.44/

Shimorina, A., & Belz, A. (2022, May). The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In A. Belz, M. Popović, E. Reiter, & A. Shimorina (Eds.), *Proceedings of the 2nd workshop on human evaluation of nlp systems (humeval)* (pp. 54–75). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.humeval-1.6/ doi: 10.18653/v1/2022.humeval-1.6.

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., ... Subramonian, A. (2024). *Evaluating the social impact of generative ai systems in systems and society.* Retrieved from https://arxiv.org/abs/2306.05949 (Forthcoming in Hacker, Engel, Hammer, Mittelstadt (eds), Oxford Handbook on the Foundations and Regulation of Generative AI. Oxford University Press)

Tafesse, W., & Wien, A. (2024). Chatgpt's applications in marketing: a topic modeling approach. *Marketing Intelligence & Planning*, *42*(4), 666–683.

Tafesse, W., & Wood, B. (2024). Hey chatgpt: an examination of chatgpt prompts in marketing. *Journal of Marketing Analytics*, *12*(4), 790-805. Retrieved from http://dx.doi.org/10.1057/s41270-023-00284-w doi: 10.1057/s41270-023-00284-w.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... Lample, G. (2023). *Llama: Open and efficient foundation language models.* Retrieved from https://arxiv.org/abs/2302.13971

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, *67*(), 101151. Retrieved from https://www.sciencedirect.com/science/article/pii/S088523082030084X doi: https://doi.org/10.1016/j.csl.2020.101151.

van Deemter, K., & Reiter, E. (2018, 11). 420lying and computational linguistics. In *The oxford handbook of lying.* Oxford University Press. Retrieved from https://doi.org/10.1093/oxfordhb/9780198736578.013.32 doi: 10.1093/oxfordhb/9780198736578.013.32.

van Miltenburg, E. (2025). *Dual use issues in the field of natural language generation.* Retrieved from https://arxiv.org/abs/2501.06636

van Miltenburg, E., Clinciu, M., Dušek, O., Gkatzia, D., Inglis, S., Leppänen, L., ... Wen, L. (2021, August). Underreporting of errors in NLG output, and what to do about it. In A. Belz, A. Fan, E. Reiter, & Y. Sripada (Eds.), *Proceedings of the 14th international conference on natural language generation* (pp. 140–153). Aberdeen, Scotland, UK: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.inlg-1.14/ doi: 10.18653/v1/2021.inlg-1.14.

van Miltenburg, E., Lu, W.-T., Krahmer, E., Gatt, A., Chen, G., Li, L., & van Deemter, K. (2020, December). Gradations of error severity in automatic image descriptions. In B. Davis, Y. Graham, J. Kelleher, & Y. Sripada (Eds.), *Proceedings of the 13th international conference on natural language generation* (pp. 398–411). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.inlg-1.45/ doi: 10.18653/v1/2020.inlg-1.45.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4 Paper.pdf

Wang, S., Zhang, G., Wu, H., Loakman, T., Huang, W., & Lin, C. (2024, November). MMTE: Corpus and metrics for evaluating machine translation quality of metaphorical language. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 11343–11358). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.emnlp-main.634/ doi: 10.18653/v1/2024.emnlp-main.634.

Xu, S., Fu, W., Gao, J., Ye, W., Liu, W., Mei, Z., ... Wu, Y. (2024). Is dpo superior to ppo for llm alignment? a comprehensive study. In *Proceedings of the 41st international conference on machine learning.*

Zhao, K., Yang, B., Lin, C., Rong, W., Villavicencio, A., & Cui, X. (2023). Evaluating open-domain dialogues in latent space with next sentence prediction and mutual information. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 562–574). Retrieved from https://aclanthology.org/2023.acl-long.33/ doi: 10.18653/v1/2023.acl-long.33.

Zhao, K., Yang, B., Tang, C., Lin, C., & Zhan, L. (2024). SLIDE: A framework integrating small and large language models for open-domain dialogues evaluation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 15421–15435). doi: 10.18653/v1/2024.findings-acl.911.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Zou, Y., Zhang, X., Zhou, J., Diao, S., Chen, J., Ding, Z., ... et al. (2023). Automatic product copywriting for e-commerce. *AI Magazine*, *44*(1), 41–53.