Evaluation and Benchmarking

Part 2: Natural language generation evaluation

Vladislav Mikhailov

Language Technology Group, University of Oslo



Table of contents



1. Background

2. Metrics-based Evaluation

3. Human-based Evaluation

4. Model-based Evaluation

5. Limitations

6. References

Language generation task formulations.

- Open-ended generation tasks:
  - Story generation;
  - Code generation;
  - Creative writing;
  - and many more.
- E.g. (Rajani et al., 2023):
  - Write a funny, short story about someone who will stop at nothing to go green and reduce their carbon footprint.
  - Can you show me a basic implementation of a linked list in C?

Language generation task formulations.

- Sequence-to-sequence tasks:
  - Machine translation;
  - Text summarization;
  - Text rewriting;
  - and many more.
- E.g. (Rajani et al., 2023):
  - Rewrite this article so someone brand new to the topic could understand it.
  - Summarize the points for and against the universal voucher system.
  - I need a TL:DR on the following document.

Background

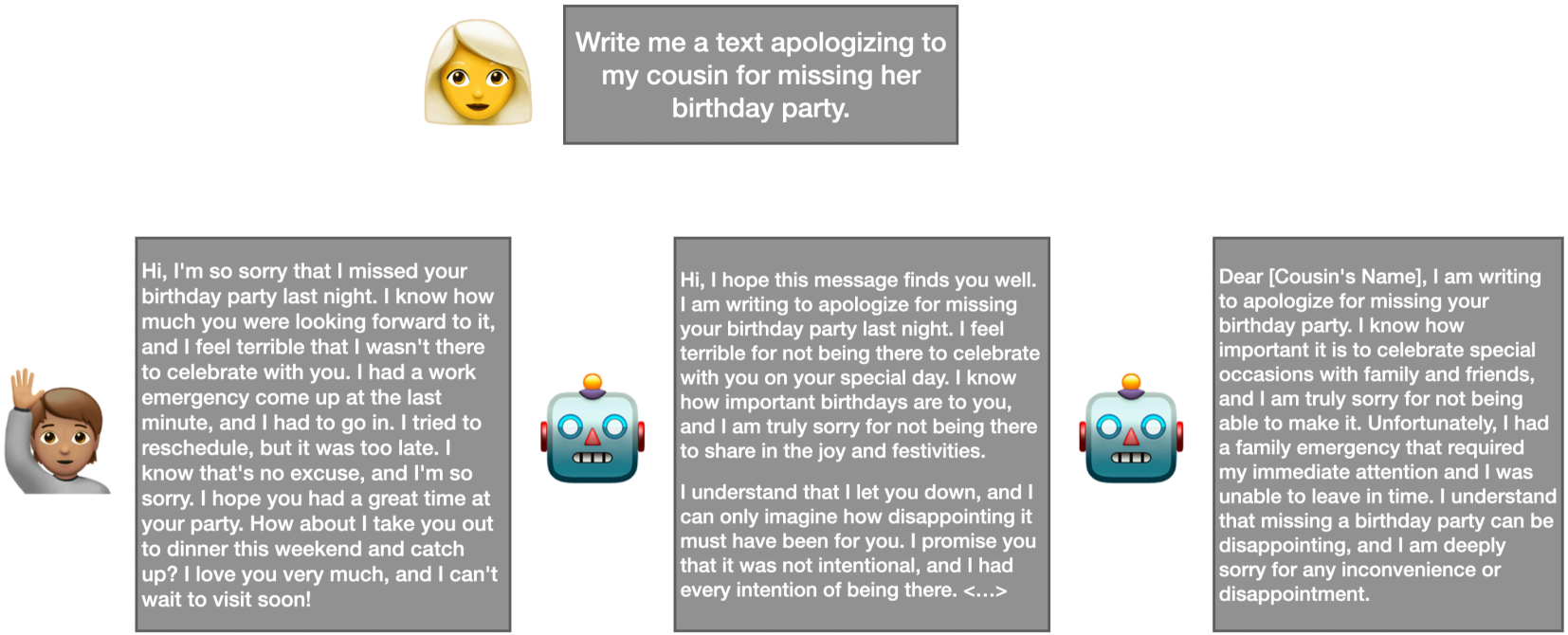


Figure 1: An example of the input, human reference, and multiple candidates.

How can we evaluate the generated outputs?

- Metrics-based evaluation:
  - **String-based metrics:** compute the overlap of textual units (N-grams, word sequences);

How can we evaluate the generated outputs?

- Metrics-based evaluation:
  - **String-based metrics:** compute the overlap of textual units (N-grams, word sequences);
  - **Embedding-based metrics:** compute the similarity measure over contextualized text representations;
  - **Learned metrics:** train a model to predict a score.

How can we evaluate the generated outputs?

- Metrics-based evaluation:
  - **String-based metrics:** compute the overlap of textual units (N-grams, word sequences);
  - **Embedding-based metrics:** compute the similarity measure over contextualized text representations;
  - **Learned metrics:** train a model to predict a score.
- Human-based evaluation:
  - **Scale-based evaluation:** grade the output on a given scale w.r.t. text quality criteria;

How can we evaluate the generated outputs?

- Metrics-based evaluation:
  - **String-based metrics:** compute the overlap of textual units (N-grams, word sequences);
  - **Embedding-based metrics:** compute the similarity measure over contextualized text representations;
  - **Learned metrics:** train a model to predict a score.
- Human-based evaluation:
  - **Scale-based evaluation:** grade the output on a given scale w.r.t. text quality criteria;
  - **Side-by-side comparison:** collect human preferences.

How can we evaluate the generated outputs?

- Metrics-based evaluation:
  - **String-based metrics:** compute the overlap of textual units (N-grams, word sequences);
  - **Embedding-based metrics:** compute the similarity measure over contextualized text representations;
  - **Learned metrics:** train a model to predict a score.
- Human-based evaluation:
  - **Scale-based evaluation:** grade the output on a given scale w.r.t. text quality criteria;
  - **Side-by-side comparison:** collect human preferences.
- Model-based evaluation:
  - **Scoring-based evaluation**: assign the generated candidate a score, and rank the candidates by the score or compare the scores between models;
  - **Side-by-side comparison:** collect model preferences.

Table of contents



1. Background

2. Metrics-based Evaluation

3. Human-based Evaluation

4. Model-based Evaluation

5. Limitations

6. References

Metrics-based Evaluation



String-based metrics.

- Main idea: evaluate based on surface features.

Metrics-based Evaluation

String-based metrics.

- Main idea: evaluate based on surface features.
- Reference-free metrics, e.g., (Fabbri et al., 2021):
  - **Novelty**: the fraction of N-grams in the output not found in the input.
  - **Redundancy**: the fraction of N-grams in the output which repeat in the input.

Metrics-based Evaluation

String-based metrics.

- Main idea: evaluate based on surface features.
- Reference-free metrics, e.g., (Fabbri et al., 2021):
  - **Novelty**: the fraction of N-grams in the output not found in the input.
  - **Redundancy**: the fraction of N-grams in the output which repeat in the input.
- Reference-based metrics, e.g.:
  - **ROUGE**: measures the number of overlapping textual units;
  - **chrF**: measures character-based N-gram overlap between the output and the reference;
  - **METEOR**: accounts for non-exact matches based on word stems, synonyms, and paraphrases.

Metrics-based Evaluation



Embedding-based metrics.

- Main idea: evaluate based on contextualized representations and similarity measures.

Embedding-based metrics.

- Main idea: evaluate based on contextualized representations and similarity measures.
- BERTScore (Zhang et al., 2019):

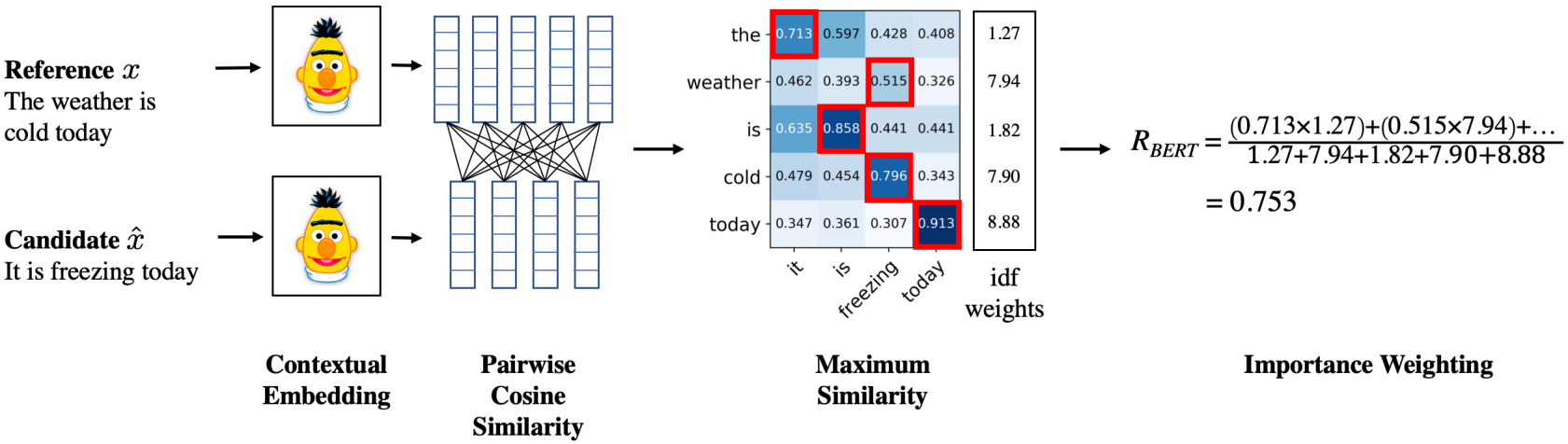


Figure 2: Illustration for computing BERTScore.

Learned metrics.

- Main idea: evaluate with a regression or classifier model trained to predict a score or a label.



Learned metrics.

- Main idea: evaluate with a regression or classifier model trained to predict a score or a label.
- RUSE (Shimanaka et al., 2018): an MLP regressor over sentence embeddings;

Learned metrics.

- Main idea: evaluate with a regression or classifier model trained to predict a score or a label.
- RUSE (Shimanaka et al., 2018): an MLP regressor over sentence embeddings;
- BLEURT (Sellam et al., 2020):

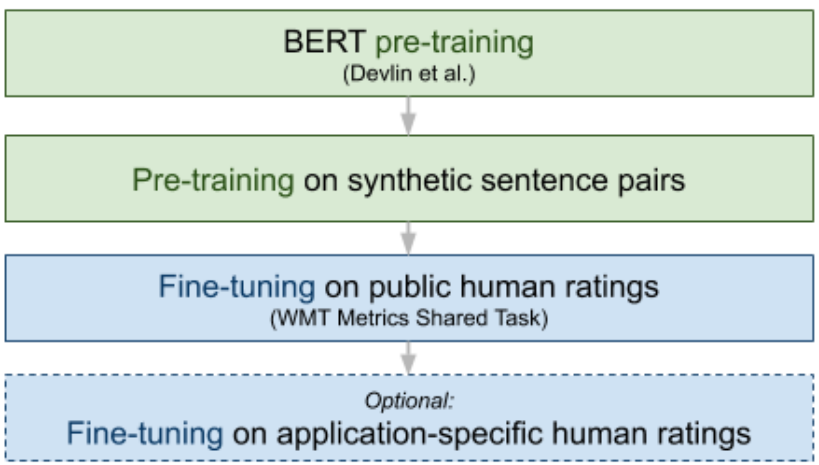


Figure 3: The training pipeline for BLEURT.

Table of contents



1. Background

2. Metrics-based Evaluation

3. Human-based Evaluation

4. Model-based Evaluation

5. Limitations

6. References

Human-based Evaluation



Scale-based evaluation.

- Main idea: identify evaluation criteria and grade the outputs on a given scale.

Human-based Evaluation



Scale-based evaluation.

- Main idea: identify evaluation criteria and grade the outputs on a given scale.
- Examples of evaluation criteria (Howcroft et al., 2020):
  - **Relevance to the input**: the degree to which an output is accurate to the input;
  - **Grammaticality**: the degree to which an output is free of grammatical errors;
  - and many more.
- Examples of scales:
  - Binary scale: a yes/no voting;
  - Likert scale: e.g., a 1-5 point scale.

Human-based Evaluation

Scale-based evaluation.

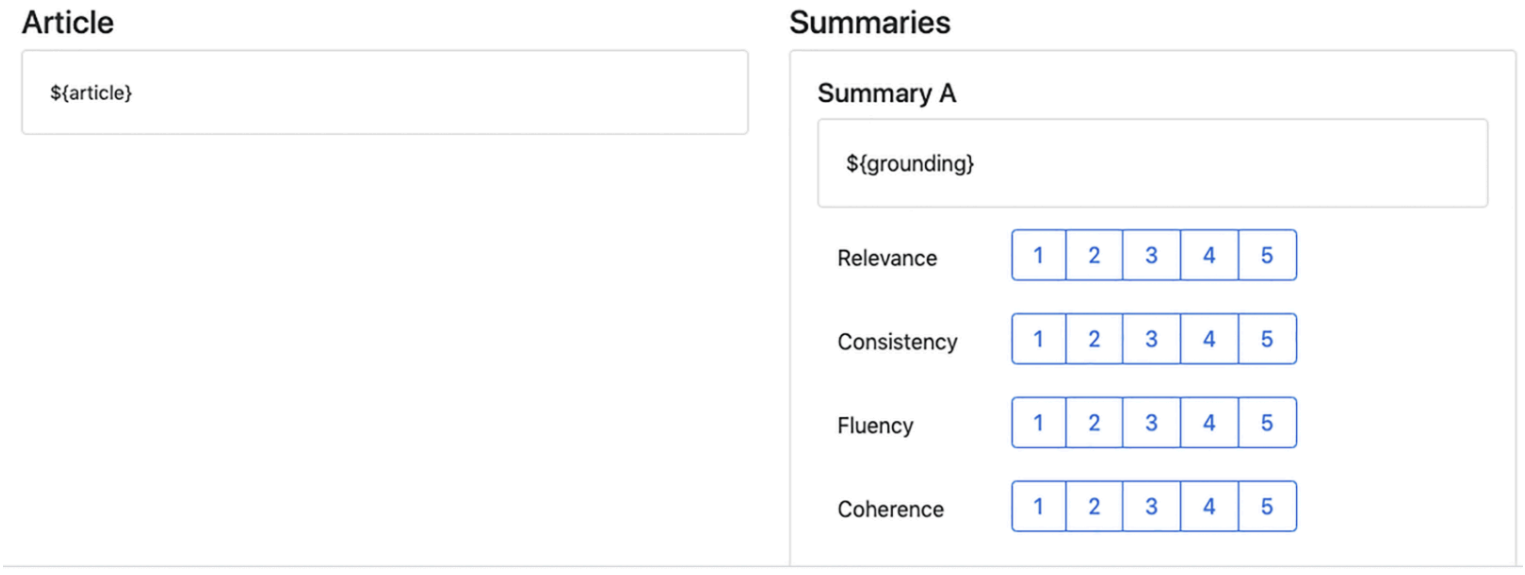


Figure 4: The annotation interface using a Likert scale (Fabbri et al., 2021).

Human-based Evaluation

Side-by-side comparison.

- Main idea: given two candidates, an annotator selects which candidate is better or tie/none otherwise.

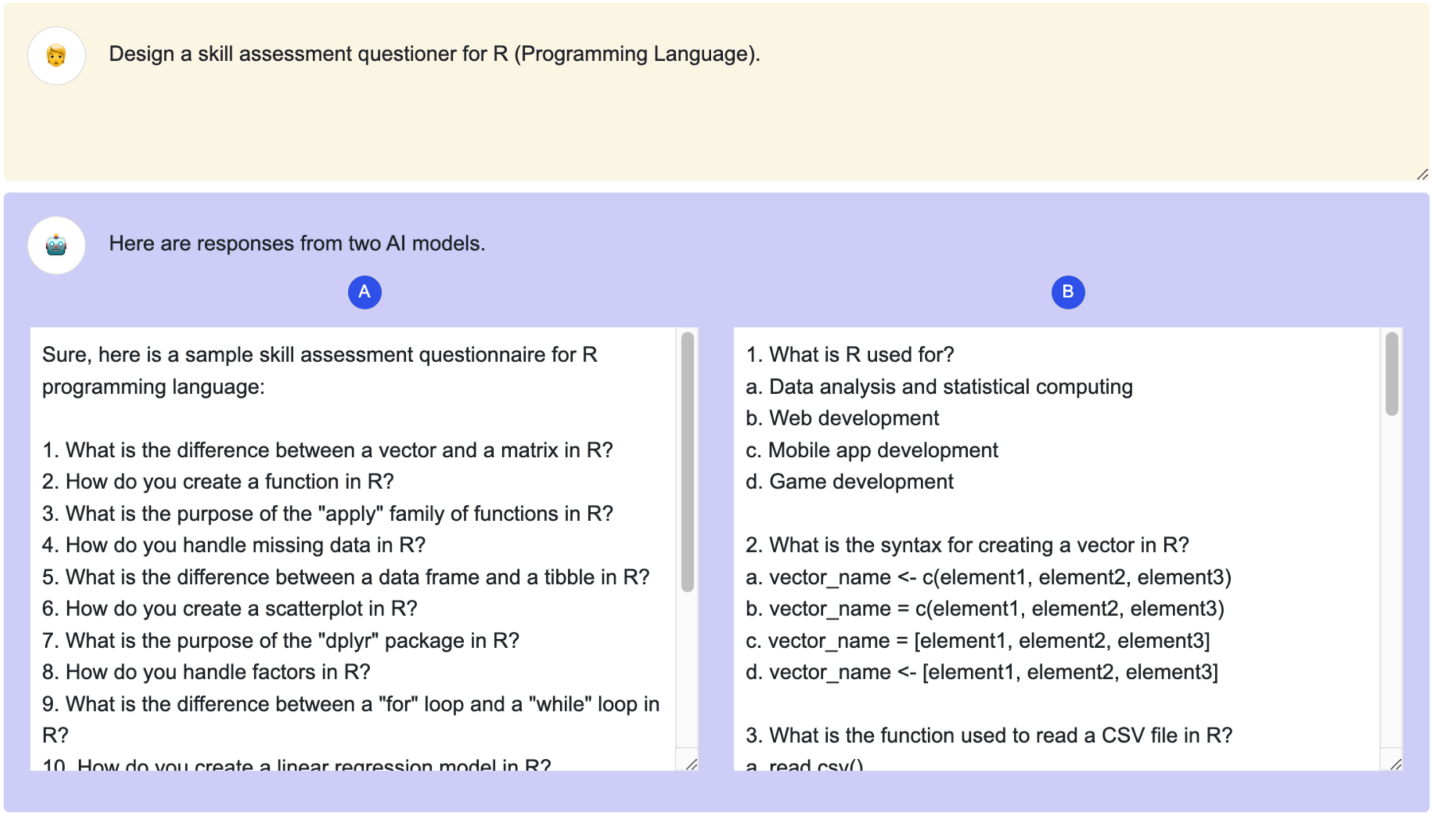


Figure 5: An example of side-by-side comparison (Wang et al., 2024).

Table of contents



1. Background

2. Metrics-based Evaluation

3. Human-based Evaluation

4. Model-based Evaluation

5. Limitations

6. References

Model-based Evaluation

Scoring-based evaluation.

- Main idea: ask an LM to score the output.

```
[System]
Please act as an impartial judge and evaluate the quality of the response provided by an
AI assistant to the user question displayed below. Your evaluation should consider factors
such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of
the response. Begin your evaluation by providing a short explanation. Be as objective as
possible. After providing your explanation, please rate the response on a scale of 1 to 10
by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

[Question]
{question}

[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]
```

Figure 6: A prompt example for grading a model output (Zheng et al., 2024).

Model-based Evaluation

Side-by-side comparison.

- Main idea: given two candidates, an LM selects which candidate is better or tie/none otherwise.

Question: If the FED buys bonds in the secondary market
(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:	**Assistant B:**
The Federal Reserve buys bonds in the secondary market to increase the money supply.	(A) the money supply will increase.

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:	**Assistant B:**
The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by: 1. Increasing the money supply. 2. Increasing the money supply and lowering the discount rate. 3. Increasing the money supply and lowering the discount rate and increasing the money supply.	When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples: 1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money. 2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities. 3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.
On the other hand, **Assistant B** provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. **Assistant B**'s response is more helpful, relevant, and detailed.

Figure 7: Multi-turn side-by-side comparison with GPT-4 (Zheng et al., 2024).

Table of contents



1. Background

2. Metrics-based Evaluation

3. Human-based Evaluation

4. Model-based Evaluation

5. Limitations

6. References

Limitations

Metrics-based evaluation.

- Automatic metrics:
  - do not capture the diversity of plausible outputs and are similar to each other (Colombo et al., 2023);
  - can poorly correlate with human ratings w.r.t. text quality criteria (Chhun et al., 2022);
  - but still can be very useful (Nimah et al., 2023)

Limitations

Human-based evaluation.

- No standard methodologies on the annotation design and evaluation criteria (Howcroft et al., 2020);
- Time- and labor-expensive, especially for medium-resource and low-resource languages;
- Low reproducibility of the results (Karpinska et al., 2021).

Model-based evaluation.

- Position bias (Wang et al., 2023);
- Scores correlate with the number of tokens (Wang et al., 2024).

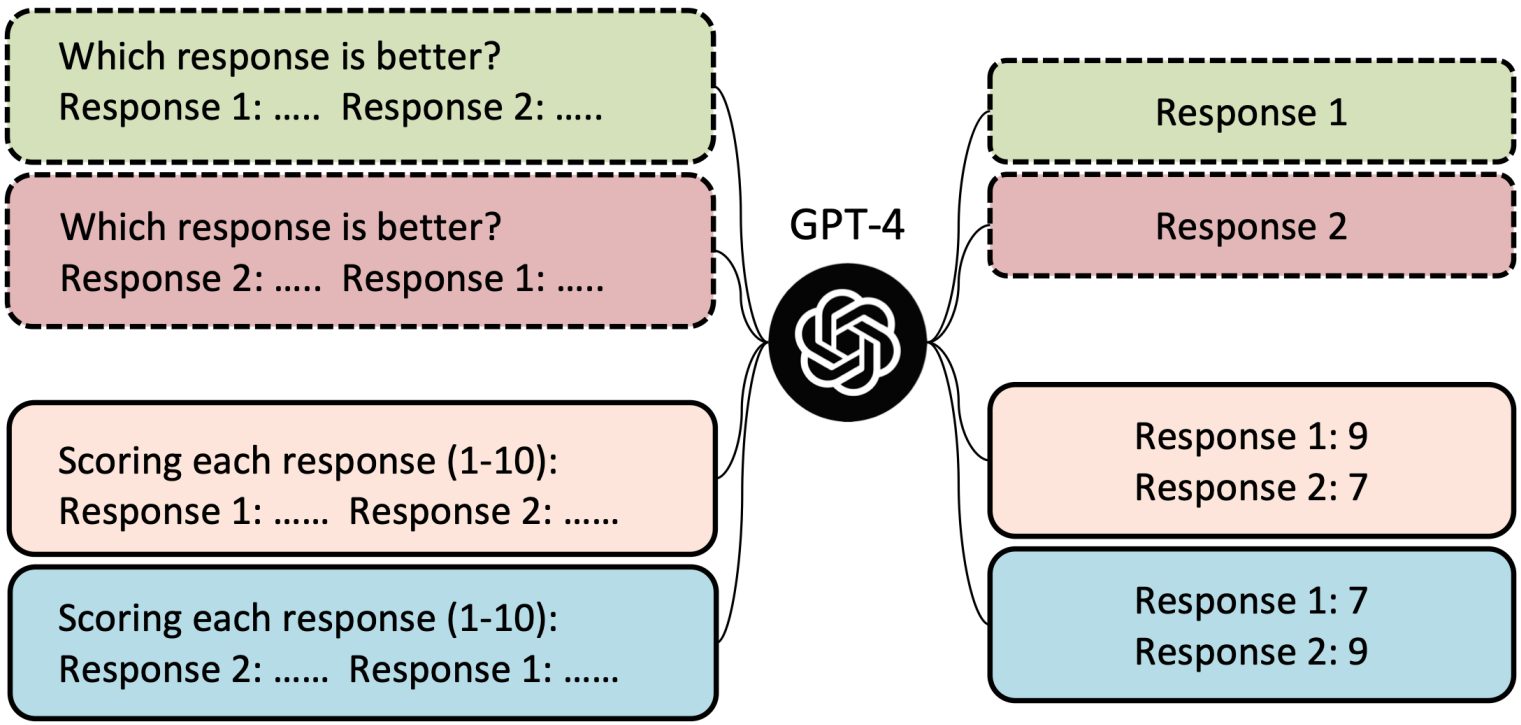


Figure 8: An example of the position bias (Wang et al., 2023).

Table of contents



1. Background
2. Metrics-based Evaluation
3. Human-based Evaluation
4. Model-based Evaluation
5. Limitations
6. References

References I

Chhun, C., Colombo, P., Suchanek, F. M., and Clavel, C. (2022). Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Colombo, P., Peyrard, M., Noiry, N., West, R., and Piantanida, P. (2023). The glass ceiling of automatic evaluation in natural language generation. In Park, J. C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., and Krisnadhi, A. A., editors, *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 178–183, Nusa Dua, Bali. Association for Computational Linguistics.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., and Rieser, V. (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.

References II



Karpinska, M., Akoury, N., and Iyyer, M. (2021). The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nimah, I., Fang, M., Menkovski, V., and Pechenizkiy, M. (2023). NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1240–1266, Toronto, Canada. Association for Computational Linguistics.

Rajani, N., Tunstall, L., Beeching, E., Lambert, N., Rush, A. M., and Wolf, T. (2023). No robots. *https://huggingface.co/datasets/HuggingFaceH4/no_robots*.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Shimanaka, H., Kajiwara, T., and Komachi, M. (2018). RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. (2023). Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

References III

Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K., Wadden, D., MacMillan, K., Smith, N. A., Beltagy, I., et al. (2024). How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.