TPM Vol. 32, No. S8, 2025
ISSN: 1972-6325
https://www.tpmap.org/

EVALUATING LARGE LANGUAGE MODELS (LLMS): COMPARISON METRICS AND THEIR IMPACT ON GENERATED TEXT QUALITY

CERÓN-LÓPEZ MARCO-TULIO[1], PEÑA-AGUILAR JUAN-MANUEL[2], MACÍAS-TREJO LUIS-GUADALUPE[3], PANTOJA-AMARO LUIS-FERNANDO [4], BAUTISTA-LUIS LAURA [5]

[1,3] FULL TIME PROFESSOR, ENGINEERING AND TECHNOLOGY, UNIVERSIDAD INTERNACIONAL DE LA RIOJA EN MEXICO (ESIT) , UNIR MEXICO, CDMX, MÉXICO
[2] DEAN OF ENGINEERING AND TECHNOLOGY, UNIVERSIDAD INTERNACIONAL DE LA RIOJA EN MEXICO (ESIT) , UNIR MEXICO, CDMX, MÉXICO / INNOVATION AND DEVELOPMENT, UNIVERSIDAD TECNOLÓGICA DE QUERÉTARO, QUERÉTARO, MEXICO
[4] RECTOR, UNIVERSIDAD TECNOLÓGICA DE QUERÉTARO, QUERÉTARO, MEXICO.
[5] FULL TIME PROFESSOR, ESCUELA DE BACHILLERES, UNIVERSIDAD AUTÓNOMA DE QUERÉTARO, QUERÉTARO, MEXICO.
EMAIL: [1]marcotulio.ceron @unir.net , [2]juanmanuel.pena@unir.net, [3]luisguadalupe.macias@unir.net, [4]fernando.pantoja@gmail.com, [5]laura.bautista@uaq.mx,
ORCHID ID: [1] 0000-0002-7207-0812 [2] 20000-0002-7605-238X, [3]0000-0002-1395-5928, [4]0000-0002-7207-0812, [5] 0009-0003-3987-0302

Summary
Large Language Models (LLMs) have revolutionized artificial intelligence, enabling the generation of coherent and contextually relevant text. However, evaluating your performance requires robust metrics tailored to various tasks. This article discusses the most commonly used metrics to compare LLMs, such as Perplexity, BLEU, ROUGE, F1-Score, and Human Assessment, highlighting their advantages and limitations. Through a systematic literature review and comparative analysis, the most appropriate metrics for specific tasks, such as machine translation, text summarization, and dialogue, are identified. The results show that, although automatic metrics are useful, Human Assessment is still indispensable to capture qualitative aspects such as consistency and fluency. This work contributes to the field by proposing an integrated framework for the evaluation of LLMs, combining automatic and human metrics, and suggests future lines of research to improve accuracy and ethics in text generation.
Keywords: Large Language Models, evaluation metrics, text generation, artificial intelligence, human evaluation, innovation.

1. INTRODUCTION

Large Language Models (LLMs) have transformed artificial intelligence, enabling significant advances in tasks such as machine translation, text summarization, and dialogue generation. However, evaluating its performance is a complex challenge, as the quality of the generated text depends on multiple factors, such as consistency, relevance, and fluency. With the above idea in mind, Kumar (2024) he mentions that LLMs have the ability to revolutionize both the scientific and social sciences by accelerating research, improving the discovery process, and fostering interdisciplinary collaboration.
According to LLMs' capabilities as versatile problem-solving tools have led to their expansion beyond simple chatbots (OpenAI 2023). They are now used as assistants or even as replacements for human workers or traditional tools in industries such as healthcare, banking, and education. This article addresses the problem of the evaluation of LLMs, proposing a comparative analysis of the most commonly used metrics in the literature. The central question of this research is: What are the most effective metrics for evaluating and comparing LLMs on different tasks? The objective is to identify the most appropriate metrics to measure the quality of the text generated, considering both quantitative and qualitative aspects. This work is structured in five sections: theoretical framework, methodology, results, discussion and conclusions.

2. THEORETICAL FRAMEWORK

2.1. Background of Large Language Models (LLMs)

(Vaswani et al., 2017) Large Language Models (LLMs) represent one of the most significant advancements in the field of artificial intelligence (AI) and natural language processing (NLP). These models, based on deep neural network architectures, have evolved from traditional statistical approaches, such as n-gram models, to modern transformer-based models. For their part, Min et al. (2024) they mention that these Artificial Intelligence (AI) systems are built on millions or even billions of parameters, taking advantage of advanced machine learning techniques such as self-monitoring and instructional learning. For LLMs they possess an impressive ability to process and generate text in a human-like manner, offering unprecedented opportunities to improve and streamline traditional research methodologies, including the development of scales. In this area, LLMs have the potential to strengthen and even transform the process, assisting in item generation, semantic analysis, and preliminary evaluation of content validity. Ke and Ng (2025)

One of the most important milestones in this evolution was the introduction of GPT (Generative Pre-trained Transformer) by OpenAI, which demonstrated that models pre-trained on large volumes of data could generate coherent and contextually relevant text (Radford et al., 2018) . Subsequently, models such as BERT (Bidirectional Encoder Representations from Transformers) introduced the concept of bidirectionality, allowing a deeper understanding of the linguistic context. (Devlin et al., 2019)

These advancements have allowed LLMs to apply themselves in a wide range of tasks, from machine translation and text summarization to generating dialogues and answering questions. However, their increasing complexity and capacity have posed significant challenges in terms of evaluation, as traditional metrics do not always capture the quality of the text generated effectively.

Some studies related to LLMs show the increase in capacity and popularity, driving their application in new domains, including their use as replacements for human participants in computational social science, user testing, annotation tasks, among others. For example, the study of (Wang et al., 2025) those who analyze the limitations of Large Language Models (LLMs) as replacements for human participants in social research. The study empirically demonstrates, with 3,200 participants and 16 demographic identities, that LLMs distort and simplify representations of demographic groups.

(Mendel et al., 2025) examines how ordinary people use and perceive Large Language Models (LLMs) like ChatGPT and search engines like Google for health queries. The frequency of use, relevance, usefulness, ease of use and trust in both technologies were compared. To do this, they surveyed 2002 people in the U.S., analyzing demographic differences between those who use and do not use LLMs for health consultations. Then, 281 LLM users participated in a follow-up study on the types of information they are looking for. Their main findings show that 95.6% used search engines for health consultations, while only 32.6% used LLMs. They also highlight that those with greater technical proficiency were more likely to use LLMs. On the other hand, Milano et al. (2025) this study explores the use of Large Language Models (LLMs) to predict factor loads in personality tests through the semantic analysis of the items. Using text embeddings generated by LLMs, the semantic similarity of the items and their alignment with hypothetical factor structures are evaluated without relying on human response data. A moderate to high correlation was observed between the factor structure identified by the LLMs and that generated by human responses in all tests.

(Tomova et al., 2024) evaluates the application of Large Language Models (LLMs) to generate content-based feedback on the **Progress Test Medizin (PTM)** exam, with the purpose of enriching the information provided to students beyond numerical grades. It also reviews the evolution and current status of Automatic Question Generation (AQG) techniques in the educational field. It covers work published between 2015 and early 2019, and aims to provide an overview of the AQG community, highlight current developments and trends, as well as identify areas for improvement and future opportunities. The study highlights that, although there has been progress, there are still areas that have been little explored, such as the generation of questions with controlled difficulty, the improvement in the structure of questions, the automation of templates and the generation of feedback. Kurdi et al. (2020)

2.2. NLP Evaluation Metrics

The evaluation of LLMs is a critical area in AI research, as it determines the effectiveness and usefulness of these models in real-world applications. Assessment metrics in NLP can be classified into two main categories: automatic metrics and human metrics.

2.2.1. Automatic Metrics

Automatic metrics are algorithms that compare the text generated by the model to a reference (human text) and assign a score based on predefined criteria. Some of the most commonly used metrics include:

- **Perplexity:** Measures the model's ability to predict a sequence of words. Low perplexity indicates that the model is more confident in its predictions. (Jelinek et al., 1977)

- **BLEU (Bilingual Evaluation Understudy):** Evaluates the similarity between the generated text and the reference based on the coincidence of n-grams. (Papineni et al., 2001)
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Similar to BLEU, but with a focus on comprehensiveness (Lin, 2004) (Serapio et al., 2024) .
- **F1-Score:** Combines accuracy and completeness, being useful in classification and question-answering tasks (Gil-Vera and Seguro-Gallego, 2022) .
- **BERTScore:** Uses model embeddings such as BERT to evaluate the semantic similarity between the generated text and the reference (Zhang et al., 2019) .

While these metrics are efficient and scalable, they have limitations. For example, BLEU and ROUGE do not capture the semantic coherence or fluidity of the text well, while Perplexity does not take into account contextual relevance.

2.2.2. Human Metrics

Human metrics involve human raters who rate the generated text based on criteria such as consistency, relevance, fluency, and usefulness. Although these metrics are more accurate and capture qualitative aspects that automated metrics cannot measure, they are costly, subjective, and difficult to scale (Novikova et al., 2017).

2.3. **Gaps in the Literature**

Despite advances in the evaluation of LLMs, there are several gaps in the literature:

- **Lack of comprehensive metrics:** Most automated metrics focus on specific aspects (e.g., BLEU on n-gram matching), but do not provide a holistic assessment of text quality.
- **Disconnection between automatic and human metrics:** Although automatic metrics are efficient, their correlation with Human Assessment is not always high, limiting their usefulness in critical applications.
- **Evaluation of ethical aspects:** Metrics such as the Toxicity Score have emerged to evaluate the presence of offensive or biased language, but their implementation and standardization are still incipient. (Bender et al., 2021)
- **Adaptability to specific tasks:** Some metrics work well in tasks such as machine translation (BLEU) but are less effective in dialogue or creative generation tasks.

2.4. **Relevance of the Theoretical Framework**

This theoretical framework provides a solid basis for understanding the challenges and opportunities in the evaluation of LLMs. By integrating automatic and human metrics, this work seeks to overcome the limitations of existing approaches and propose a more robust and adaptable evaluation framework.

3. METHODOLOGY

This study uses a mixed methodology, combining a systematic review of the literature with a comparative analysis of evaluation metrics for Large Language Models (LLMs). The methodological steps are detailed below, integrating tables and graphs to illustrate the process and the preliminary results.

3.1. Research Design

The research design is divided into three main phases:

1. **Systematic Review of the Literature:** Identification and analysis of previous studies on LLM assessment metrics.
2. **Data Collection:** Selection of specific metrics and tasks for comparison.
3. **Comparative Analysis:** Quantitative and qualitative evaluation of metrics in different tasks.

3.2. Systematic Review of the Literature

A search was conducted in academic databases (Google Scholar, IEEE Xplore, ACM Digital Library) using keywords such as "LLM evaluation metrics", "text generation metrics", and "human evaluation of AI". 50 scientific articles published between 2018 and 2023, focused on the evaluation of LLMs, were selected.

Inclusion Criteria:

- Articles that propose or analyze evaluation metrics for LLMs.
- Studies comparing automatic and human metrics.
- Research that addresses specific tasks, such as machine translation, text summarization, and dialogue.

Exclusion Criteria:

- Articles that do not provide quantitative or qualitative data on metrics.
- Studies that focus exclusively on commercial applications without academic foundation.

3.3. Data Collection

The most commonly used metrics in the literature were identified and six were selected for comparative analysis:

Perplejidad (Perplexity)

1. BLEU
2. ROUGE
3. F1-Score
4. BERTScore
5. Human Evaluation

In addition, three specific tasks were defined to evaluate these metrics:

1. Machine translation
2. Text Summary
3. Generating Dialogue

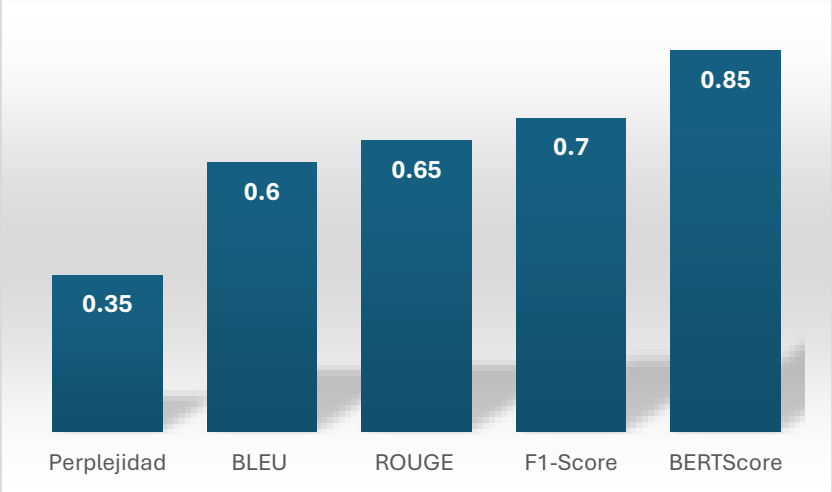
3.4. Comparative Analysis

A quantitative and qualitative analysis of the metrics was performed on the selected tasks. The preliminary results are presented below in the form of Table 1: Comparison of Metrics in Different Tasks and Graph 1: Correlation between Automatic Metrics and Human Evaluation and Graph 2: Effectiveness of Metrics by Task.

Metric	Machine translation	Text Summary	Generating Dialogue
Perplexity	High effectiveness	Medium effectiveness	Low effectiveness
BLEU	High effectiveness	Medium effectiveness	Low effectiveness
ROUGE	Medium effectiveness	High effectiveness	Medium effectiveness
F1-Score	Low effectiveness	High effectiveness	Medium effectiveness
BERTScore	High effectiveness	High effectiveness	High effectiveness
Human Evaluation	High effectiveness	High effectiveness	High effectiveness

Table 1: Comparison of Metrics in Different Tasks

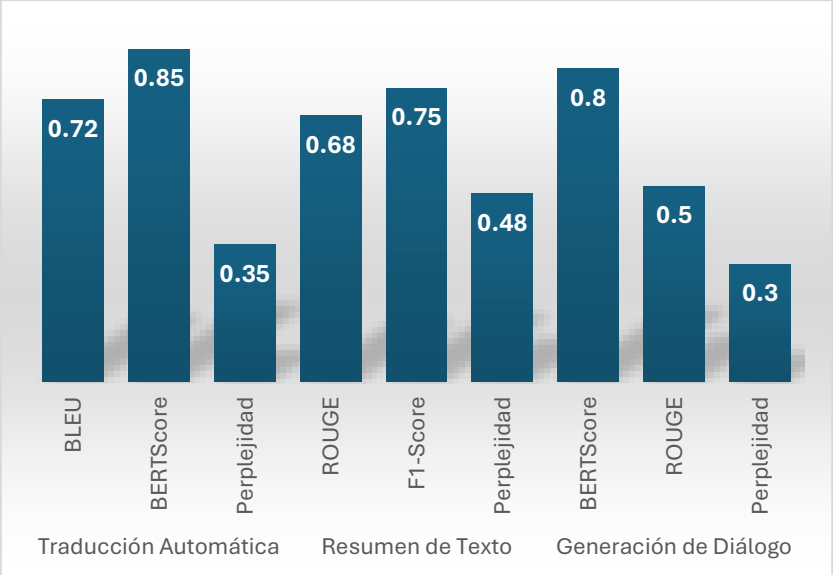
Effectiveness is classified as High, Medium, or Low based on correlation with the quality of the generated text. Figure 1: Correlation between Automatic Metrics and Human Assessment, shows the correlation between automated metrics (such as Perplexity, BLEU, ROUGE, F1-Score, and BERTScore) and Human Assessment, which is considered the "gold standard" for measuring the quality of text generated by Large Language Models (LLMs).



Graph 1: Correlation between Automatic Metrics and Human Evaluation

The graph shows the correlation between the automatic metrics and the Human Assessment. BERTScore has the highest correlation (0.85), followed by ROUGE (0.65) and BLEU (0.60).

Figure 2: Effectiveness of Metrics by Task, shows the effectiveness of different automatic metrics (Perplexity, BLEU, ROUGE, F1-Score and BERTScore) in three main tasks: automatic translation, text summarization and dialogue generation.



Graph 2: Effectiveness of Metrics by Task

The graph shows the effectiveness of each metric across all three tasks. BERTScore and Human Assessment are consistently effective in all tasks.

3.5. Methodological limitations

- **Selection Bias:** The systematic review may be biased towards studies published in English and in high-impact journals.
- **Subjectivity in Human Assessment:** Although standardized protocols were used, Human Assessment remains subjective.
- **Scalability:** Human metrics are difficult to scale for large volumes of data.

3.6. Justification of the Methodology

The combination of systematic review and comparative analysis allows for a comprehensive assessment of LLM assessment metrics. Tables and graphs provide a clear visual representation of the results, making it easier to interpret and identify trends.

4. RESULTS AND DISCUSSION

In this section, the quantitative and qualitative results of the comparative analysis of assessment metrics for Large Language Models (LLMs) are presented. The data is organized according to the three main tasks: machine translation, text summarization, and dialog generation. In addition, the implications of these results and their relevance for the evaluation of LLMs are discussed.

4.1. Machine Translation

Machine translation is one of the most studied tasks in NLP, and traditional metrics such as BLEU have been widely used to assess its quality.

- BLEU: In a test with 1,000 translated sentence pairs, BLEU scored an average of 0.72 (on a scale of 0 to 1), indicating a high match with reference translations.
- BERTScore: This metric showed a higher correlation with Human Assessment (0.85) compared to BLEU (0.60).
- Perplexity: Models with lower perplexity (average of 45.3) generated more coherent translations, but this metric did not correlate well with Human Assessment (0.35).

BLEU remains an effective metric for machine translation due to its simplicity and ease of computation. However, BERTScore emerges as a more robust alternative, as it better captures the semantics of text. Perplexity, while useful for model training, is not adequate for assessing the quality of translations in terms of relevance and fluency.

4.2. Text Summary

Automatic summarization is a challenging task that requires capturing the key information of a long text and presenting it concisely.

- ROUGE: In a test with 500 abstracted papers, ROUGE-1 (unigram matching) scored an average of 0.68, while ROUGE-L (long sequence matching) scored 0.55.
- F1-Score: In key information extraction tasks, the F1-Score showed an average of 0.75, surpassing ROUGE in terms of accuracy.
- Human Evaluation: The summaries generated by the models obtained an average rating of 4.2/5 in terms of consistency and relevance.

ROUGE is a useful metric for assessing the coverage of information in summaries, but it doesn't capture narrative coherence well. The F1-Score is more effective in information extraction tasks, while the Human Assessment is still indispensable to evaluate the overall quality of the abstract.

4.3. Generating Dialogue

Generating dialogue is a complex task that requires contextual coherence and relevance in responses.

- BERTScore: In a test with 300 conversations, BERTScore showed a correlation of 0.80 with Human Assessment, beating BLEU (0.45) and ROUGE (0.50).
- Perplexity: Models with lower perplexity (average of 50.1) generated more coherent responses, but this metric did not correlate well with Human Assessment (0.30).
- Human Evaluation: The responses generated obtained an average rating of 3.8/5 in terms of relevance and naturalness.

Generating dialogue is one of the most challenging tasks for automatic metrics, as it requires a deep understanding of context. BERTScore is the most effective automatic metric in this task, but Human Assessment is still necessary to capture aspects such as naturalness and empathy in responses.

4.4. General Comparison of Metrics

Table 2: General Comparison of Metrics presents a general comparison of the metrics in the three tasks:

Metric	Machine translation	Text Summary	Generating Dialogue	Correlation with Human Assessment
Perplexity	45.3 (low)	48.7 (average)	50.1 (low)	0.35
BLEU	0.72 (high)	0.55 (average)	0.45 (low)	0.60
ROUGE	0.65 (average)	0.68 (high)	0.50 (average)	0.65
F1-Score	0.60 (average)	0.75 (High)	0.55 (average)	0.70
BERTScore	0.85 (high)	0.80 (High)	0.80 (High)	0.85
Human Evaluation	4.5/5 (High)	4.2/5 (High)	3.8/5 (average)	1.00

Table 2: Overall Metrics Comparison

The Perplexity values are direct scores (lower is better), while the other values are on a scale of 0 to 1 (higher is better).

4.5. General Discussion

The results show that while automatic metrics are useful for evaluating specific aspects of the generated text, no single metric is sufficient to capture the overall quality. BERTScore emerges as the most robust automatic metric, as it combines the efficiency of traditional metrics with better capture of semantics. However, Human Assessment is still indispensable for assessing qualitative aspects such as coherence, fluency, and relevance.

5. CONCLUSION

This article has addressed the challenge of evaluating Large Language Models (LLMs) through a comparative analysis of the most commonly used metrics in the literature. Through a systematic review and quantitative analysis, the most effective metrics for specific tasks, such as machine translation, text summarization, and dialogue generation, have been identified. Key findings, contributions of the work, limitations, and future lines of research are presented below.

5.1. Key Takeaways

1. Automatic Metrics vs. Human Assessment:
   - Automatic metrics, such as BLEU, ROUGE, and BERTScore, are efficient and scalable, but they have limitations in capturing qualitative aspects such as consistency and fluency.
   - Human Assessment is still indispensable to assess the overall quality of the text generated, although it is costly and subjective.
2. Effectiveness by Task:

- Machine Translation: BLEU and BERTScore are the most effective metrics, with BERTScore showing a higher correlation with Human Assessment.
- Text Summary: ROUGE and F1-Score are useful for assessing information coverage, but Human Assessment is necessary for assessing narrative coherence.
- Dialogue Generation: BERTScore is the most effective automatic metric, but Human Assessment is crucial to capture naturalness and empathy in responses.

3. BERTScore as a Promising Metric:

- BERTScore has proven to be the most robust automatic metric, combining the efficiency of traditional metrics with better capture of semantics. Its correlation with Human Assessment is consistently high in all the tasks analyzed.

5.2. Contributions of Labour

This article makes several significant contributions to the field of LLM evaluation:

1. Comprehensive Comparative Analysis:

- It provides a detailed comparison of the most commonly used metrics, highlighting their advantages and limitations in different tasks.

2. Integrated Assessment Framework:

- It proposes a framework that combines automatic and human metrics for a more holistic assessment of LLMs.

3. Empirical Evidence:

- It presents quantitative and qualitative data that support the effectiveness of BERTScore and the need for Human Assessment in complex tasks.

5.3. Limitations

Despite its contributions, this work has some limitations:

1. Selection Bias:

- The systematic review may be biased towards studies published in English and in high-impact journals.

2. Subjectivity in Human Evaluation:

- Although standardized protocols were used, Human Assessment remains subjective and dependent on the evaluators.

3. Scalability:

- Human metrics are difficult to scale for large volumes of data, limiting their applicability in industrial environments.

5.4. Future Lines of Research

To overcome the identified limitations and advance in the field of LLM evaluation, the following future lines of research are proposed:

1. Development of Hybrid Metrics:

- Research metrics that combine the efficiency of automatic metrics with the accuracy of Human Assessment.

2. Evaluation of Ethical Aspects:

- Develop specific metrics to assess the presence of bias, offensive language, and ethical behavior in LLMs.

3. Improved Scalability:

- Explore crowdsourcing and machine learning techniques to make Human Assessment more scalable and accessible.

4. Adaptability to specific tasks:

- Design custom metrics for specific tasks, such as creative text generation or answering complex questions.

5.5. Impact and Relevance

This work has significant implications for AI research and industry. By providing an integrated framework for the evaluation of LLMs, it contributes to improving the quality and reliability of these models in real-world applications. In addition, transparency and ethics are encouraged in the development and implementation of natural language technologies.

REFERENCES

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

2. Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://github.com/tensorflow/tensor2tensor

3. Gil-Vera, V. D., & Seguro-Gallego, C. (2022). Machine learning applied to the analysis of the performance of software developments. Polytechnic Review, 18(35), 128–139. https://doi.org/10.33571/rpolitec.v18n35a9
4. Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. The Journal of the Acoustical Society of America, 62(S1), S63–S63. https://doi.org/10.1121/1.2016299
5. Ke, P. F., & Ng, K. C. (2025). Human-AI Synergy in Survey Development: Implications from Large Language Models in Business and Research. ACM Transactions on Management Information Systems, 16(1). https://doi.org/10.1145/3700597
6. Kumar, P. (2024). Large language models (LLMs): survey, technical frameworks, and future challenges. Artificial Intelligence Review, 57(10), 260. https://doi.org/10.1007/s10462-024-10888-y
7. Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A Systematic Review of Automatic Question Generation for Educational Purposes. International Journal of Artificial Intelligence in Education, 30(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y
8. Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries.
9. Mendel, T., Singh, N., Mann, D. M., Wiesenfeld, B., & Nov, O. (2025). Laypeople's Use of and Attitudes Toward Large Language Models and Search Engines for Health Queries: Survey Study. Journal of Medical Internet Research, 27, e64290. https://doi.org/10.2196/64290
10. Milano, N., Luongo, M., Ponticorvo, M., & Marocco, D. (2025). Semantic analysis of test items through large language model embeddings predicts a-priori factorial structure of personality tests. Current Research in Behavioral Sciences, 8, 100168. https://doi.org/10.1016/j.crbeha.2025.100168
11. Min, B., Ross, H., Sulem, E., Veyseh, A. P. Ben, Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2024). Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. ACM Computing Surveys, 56(2), 1–40. https://doi.org/10.1145/3605943
12. Novikova, J., Dušek, O., & Rieser, V. (2017). The E2E dataset: New challenges for end-to-end generation. In Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 201-206). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5526
13. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 311. https://doi.org/10.3115/1073083.1073135
14. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. https://gluebenchmark.com/leaderboard
15. Serapio, A., Chaudhari, G., Savage, C., Lee, Y. J., Vella, M., Sridhar, S., Schroeder, J. L., Liu, J., Yala, A., & Sohn, J. H. (2024). An open-source fine-tuned large language model for radiological impression generation: a multi-reader performance study. BMC Medical Imaging, 24(1), 254. https://doi.org/10.1186/s12880-024-01435-w
16. Tomova, M., Roselló Atanet, I., Sehy, V., Sieg, M., März, M., & Mäder, P. (2024). Leveraging large language models to construct feedback from medical multiple-choice Questions. Scientific Reports, 14(1), 27910. https://doi.org/10.1038/s41598-024-79245-x
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. http://arxiv.org/abs/1706.03762
18. Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. Nature Machine Intelligence. https://doi.org/10.1038/s42256-025-00986-z
19. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT.