CRITERIA FOR THE QUALITY ASSESSMENT OF LARGE LANGUAGE MODELS

Yuriy Khoma, D. Sc., Associate Professor, Ivan Shchudlo, PhD student
Lviv Polytechnic National University, Ukraine,
e-mail: yurii.v.khoma@lpnu.ua, ivan.o.shchudlo@lpnu.ua

https://doi.org/10.23939/istcmtm2025.03.

Abstract. The development of large language models (LLMs) with each new iteration demonstrates a significant improvement in their ability to understand and generate text, which opens up increasingly wide opportunities for their integration into information processing systems and digital business processes of enterprises and institutions. In the context of the constant growth of the complexity and functional capabilities of LLMs, the development of reliable methods for their evaluation becomes a fundamental challenge for the research community, as traditional metrics for evaluating text information are often unable to fully cover the entire depth and multifaceted nature of their potential capabilities and characteristics. The creation of comprehensive LLM quality assessment systems is designed to ensure not only an objective comparison of different models but also to form critical feedback for targeted technology improvement and the prevention of potential risks associated with their large-scale implementation. The article is devoted to the systematization of criteria for assessing the quality of large language models that have become widespread in natural language processing tasks. The article aims to create a comprehensive approach to LLM quality assessment that covers the main aspects of their functioning. The paper defines LLM quality criteria such as precision and completeness of responses, naturalness of speech, consistency, toxicity, bias, security vulnerabilities, among others. Three main approaches to assessing LLM quality are analyzed in detail: expert assessments, comparison with reference data, and automated methods without reference data. For each quality criterion, the most effective assessment methods are determined, and their advantages and disadvantages in different application contexts are noted. In conclusion, it is emphasized that, despite the high reliability of expert assessments, automated methods are becoming increasingly important for large-scale LLM evaluation, especially for subjective criteria such as toxicity or bias.

Key words: Large Language Model, Quality Assessment of Large Language Models, Quality Assessment Criteria.

1. Introduction

LLM (Large Language Models) are artificial neural networks that process natural language, taking text information or an audio signal with spoken language as input in the case of the latest multimodal LLMs. The LLM generates text as output, predicting the next words or phrases that are most likely to match the given context. This may include answering questions, creating descriptions, or formulating recommendations.

Prompts in the context of large language models are text queries or instructions entered into the model to obtain desired responses or information. They can range from simple queries to complex structured instructions that tell the model exactly how to interact with the input data. System prompts are specific types of prompts that set general rules or settings for interacting with the model. They may include instructions or parameters that regulate how the model should respond to queries.

RAG (Retrieval-Augmented Generation) is an approach to building software solutions that combines the retrieval of relevant information from external data sources with the generative capabilities of LLM [1]. The RAG architecture consists of three main components:

- a vector store for storing and indexing documents,
- a semantic search system for finding relevant information
- a mechanism for integrating the found information with the generative model through an augmented prompt.

The main advantages of RAG are increased accuracy of responses due to the use of up-to-date external sources, reduction of model hallucinations, and the ability to work with private corporate data. Technical limitations of RAG include dependence on the quality of indexed data, limitations on the size of the contextual (input) text window, and additional computational costs for searching and processing external information. Practical application of RAG systems covers a wide range of tasks, including corporate user support systems, documentation analysis, business process automation, and scientific research, where accuracy and relevance of information are critical.

Large language models (LLMs) have two main types of applications based on the nature of tasks:

- structured responses, which include clearly formatted output (e.g., form filling, text classification, or extracting specific data from documents) [2]
- unstructured responses, which involve generating free text (such as writing articles, creating descriptions, or conducting dialogue).

In the context of working with LLM queries, models can process:

- closed-ended queries that require precise, factual answers (e.g., "Who wrote 'Kobzar'?" or "What is the capital of France?"), and
- open-ended questions that require detailed, analytical answers with elements of reasoning and synthesis of information (e.g., "How can artificial intelligence affect the future of education?").

Structured responses are frequently used in business processes where standardization and automation of data processing are required, while unstructured responses are more common in creative and analytical tasks. When working with specific queries, LLMs demonstrate high accuracy and reliability, as such queries have clearly defined correct answers that can be verified. Open-ended questions, conversely, demand greater flexibility from the model, the ability to analyze and synthesize information, and the skill to generate well-

reasoned judgments and conclusions. An important aspect of LLM applications is their ability to adapt to various contexts and interaction formats, allowing them to be used as a universal tool for a wide range of tasks – from simple classification to complex analysis and content generation.

Assessing the quality of large language models is a comprehensive process that includes analyzing their ability to understand context, generate relevant and useful content, reason, solve problems, and ensure ethical reliability and safety. Systematizing the criteria for the quality assessment of large language models is critically important for ensuring objective comparison of models and determining their advantages and disadvantages in various applied aspects such as safety, ethics, and meaningfulness [3]. The current state of the problem is characterized by fragmented approaches to evaluation, where various benchmarks are used, but a single comprehensive methodology covering all significant aspects of LLM functioning is lacking. The research community is actively working on creating multidimensional evaluation systems, such as HELM (Holistic Evaluation of Language Models) [4], which attempt to integrate indicators of accuracy, reasonableness, safety, confidence calibration, and social impact of models. The difficulties of standardization of evaluation are exacerbated by the rapid evolution of language models' capabilities, which requires constant updating of "benchmarks" to prevent their obsolescence and loss of discriminatory ability when evaluating models of varying quality. For example, in previous generations of generative language models, a metric such as the number of spelling and punctuation errors per 1000 characters could have been an informative way to assess model quality, but with the advent of LLMs, it loses its discriminatory ability as all LLMs generate texts without errors, and for all texts, the expected quality results for this criterion consistently reached 100%. Creating a comprehensive, standardized, and dynamic LLM evaluation system remains an open scientific challenge, the solution of which will have a significant impact on the development of the industry and the practical application of these technologies.

2. Criteria For The Quality Assessment Of Large Language Models

As noted above, the first step towards unifying LLM quality assessment is to establish appropriate quality criteria that could cover the multidimensional nature of the functional capabilities and application areas of these models. This paper reviews and systematizes the relevant criteria.

1. **Accuracy of the generated response.** The accuracy criterion calculates the percentage of correct answers in tasks such as classification or question answering [5]. This can also include metrics such as:

- **Recall**: This is the ratio of the number of correctly identified correct responses to the total number of actual correct responses in the sample.
- **F1 score**: This is the harmonic mean of precision and recall, which provides a balanced assessment of the quality of classification or prediction of a machine learning model.

This criterion is used to evaluate LLM responses to specific questions, as well as to evaluate LLM responses in a structured format.

Example question: "In what year was Taras Shevchenko born?"

Accurate LLM answer: "1814"

2. **Completeness of the generated response.**

Assessing the quality of generated text in LLMs includes analyzing coherence, semantic connectivity, and contextual relevance, which is performed by comparing the generated text with reference samples or expert assessments. Additionally, quantitative metrics such as BLEU, ROUGE [6], and perplexity are used to quantitatively measure the accuracy, informativeness, and naturalness of machine-generated text.

This criterion is applied when evaluating LLM responses to open-ended questions where the answer may include one or more paragraphs.

Example question: "What is Taras Shevchenko famous for?"

Example of complete LLM response: "Taras Shevchenko is known as a great Ukrainian poet who wrote poems in his native language and fought for the rights of ordinary people. He was also a talented artist and created many paintings and drawings depicting the life of the Ukrainian people."

3. **Natural Language.** Natural Language criterion determines the model's ability to generate grammatically correct and stylistically appropriate text that is indistinguishable from human-written text. This includes the correct use of slang, idioms, and context-dependent expressions when appropriate. The model should adapt its communication style to the context, from formal to conversational. An important aspect is avoiding mechanical, artificial constructions and the ability to convey emotional nuances through appropriate tone and vocabulary. [7]

4. **Consistency.** The model's ability to provide identical or logically consistent answers to similar questions, even if formulated differently. An important aspect is maintaining the stability of the model's statements throughout the dialogue, without sudden changes in position or contradictory statements [8]. Consistency also includes the ability to adhere to established constraints and instructions throughout the interaction. This criterion is particularly important for long-term dialogues and tasks that require the consistent application of certain rules or methodology.

5. **Toxicity.** Toxicity indicates the presence of offensive, hateful, racist, sexist, or other unacceptable language in the model's response. Assessing toxicity is important for ensuring the ethical use of LLMs. Mechanisms are needed to detect and minimize such expressions [9].

6. **Bias.** Bias is a potential problem that can arise from training on unrepresentative or biased data. Assessing the presence of bias in model responses is important to ensure fairness and equality in communication. For example, the model should avoid stereotypes regarding different social, cultural, or ethnic groups [10].

7. **Security vulnerabilities.** Security vulnerabilities refer to system flaws that can be exploited for manipulation or deception. This may include unwanted disclosure of sensitive data by the model, as well as existing vulnerabilities that allow bypassing restrictions imposed on the model by system prompts. Assessing the presence of such vulnerabilities is crucial for developing secure LLMs [11].

The relevance of some of the above criteria depends on the task itself. For example, when evaluating a document search system for banking or legal institutions, the main criteria will be security vulnerabilities and bias, whereas for a voice bot in a fast-food restaurant, the most important will be natural language and consistency, and for an agent system assisting programmers - security vulnerabilities and consistency.

3. Approaches to assessing LLM quality

The next step after defining and categorizing specific quality criteria is to systematize the approaches based on which these criteria can be assigned appropriate numerical values. In general, methods for evaluating LLM quality can be divided into three categories:

1. Using expert assessment - "manual" approach

Assessing the quality of LLMs using expert assessments involves analyzing model responses by specialists to determine their effectiveness. In turn, methods of expert assessments can be divided into individual and group, as well as assessments with and without instructions.

Individual expert assessment involves a single specialist who evaluates responses according to defined criteria, ensuring methodological consistency, speed of assessment, and cost-effectiveness of the process. This approach is optimal for assessing objective criteria, such as factual accuracy, relevance to the query, and adherence to specific instructions, where there is a clear distinction between correct and incorrect answers.

Group expert assessment involves several specialists, which allows for higher reliability of results due to averaging individual discrepancies and statistically determining consistency among experts. For assessing subjective criteria, such as content ethics, stylistic appropriateness, and overall usefulness, group assessment is more methodologically sound.

Assessment with instructions involves providing experts with detailed instructions with clear descriptors for each assessment level, which increases consistency among them and reduces the influence of subjective interpretations of criteria.

In contrast, **assessment without instructions** allows experts to apply their own criteria and approaches, contributing to the identification of unforeseen aspects of model performance quality, although it complicates quantitative comparison of different systems. Modern methodologies at various stages of research often combine individual and group assessment, with and without instructions, to ensure the comprehensiveness of the analysis.

Adaptive assessment methods involve gradual refinement of criteria based on feedback from experts and analysis of intermediate results. To increase the reliability of expert assessments, it is recommended to conduct calibration sessions before the main assessment stage. The involvement of experts from various fields of knowledge to form interdisciplinary panels capable of assessing a wide range of different aspects of LLM functioning is becoming increasingly popular. The application of statistical methods, such as Kendall's coefficient or Spearman's coefficient [12], allows for quantifying consistency among experts and verifying the reliability of assessments.

2. Comparison against benchmark - semi-automated approach

Assessing the quality of LLMs by comparing them with benchmark answers is a semi-automatic method, as it requires the initial creation of a database of benchmark answer data set by experts, which is a time-consuming manual process that provides a reference for subsequent automated comparison. The effectiveness of this method depends on the quality and representativeness of the collected reference data, as a limited set of samples can lead to an erroneous assessment of the model's ability to generalize and work with new types of queries. Despite certain limitations, comparison with benchmark data remains an important component of comprehensive LLM evaluation, especially for tasks with clearly defined correctness criteria.

Assessing the quality of large language models through comparison with benchmark data can be divided into two approaches: binary and non-binary assessment.

Binary assessment is primarily applied to structured responses where the model must provide specific information in a well-defined format, for example, when extracting facts, classifying, or solving mathematical problems with a single correct answer. In binary evaluation, the model's answer is considered to be correct only if it fully matches the reference, which allows for objective determination of the model's accuracy and calculation of metrics such as accuracy, precision, and recall.

However, binary evaluation has significant limitations when working with generative models, which can formulate correct answers in different ways, leading to the need for non-binary evaluation methods.

Non-binary assessment includes a wide range of methods for measuring partial correspondence between the model's answer and the reference, from simple lexical metrics to complex semantic comparisons. Simple lexical similarity metrics, such as Levenshtein distance, BLEU, ROUGE, or METEOR [6], measure similarity at the character or n-gram level but do not account for semantic equivalence and contextual differences. More complex methods, based on semantic analysis, use vector embeddings to compare semantic similarity, even when the wording differs significantly. Modern approaches often combine several metrics for a comprehensive assessment of answer quality; for example, BERTScore [13] uses contextual embeddings to compare semantic similarity at the token level.

For evaluating answers where different formulations with the same meaning are possible, methods based on neural network models trained to determine the semantic equivalence of texts, such as Sentence-BERT or Universal Sentence Encoder, are used. A critical aspect when using reference answers is the formation of a representative set of references that covers various potentially correct formulations. Hybrid approaches combine automatic metrics with expert assessment, where algorithmic methods are first applied to filter out obviously incorrect answers, and then experts analyze complex cases that require a deeper understanding of the context.

An important methodological tool is context-dependent assessment, which takes into account the specifics of a particular task and knowledge domain when comparing answers with references. Modern research demonstrates the effectiveness of methods that use LLMs as evaluators of other LLMs (LLM-as-judge) [14, 15], where a powerful model compares the evaluated model's answer with the reference, considering semantics, context, and pragmatic aspects of communication.

3. Automated quality assessment methods that don't use benchmark data

Fully automated quality assessment of LLMs without the use of reference data represents an innovative research direction that allows for objectively measuring model performance in various contexts on data volumes significantly larger than what can be realistically processed through expert assessments or comparisons with reference data.

The main method that has recently gained widespread use is the utilization of other **LLMs as judges**, where one language model, usually more powerful, analyzes and evaluates the responses of another model based on predefined quality criteria, such as relevance, factual accuracy, and logical consistency. To ensure the objectivity of such an assessment, specialized prompt instructions are developed for the judge model, which clearly define the quality criteria, analysis procedure, and scoring scale, minimizing the influence of potential biases of the judge model itself [14, 15].

Self-evaluation is a method where the same model that generates the response analyzes its own results, identifying potential errors, inaccuracies, or incompleteness in the response. Self-evaluation of large language models is carried out through their ability to analyze their own responses using specially designed prompts that encourage the model to critically evaluate various aspects of the generated content, including factual accuracy, logical consistency, and contextual relevance. Within this approach, models can use chain-of-thought reasoning techniques for step-by-step analysis of their own responses, and also apply internal confidence metrics to assess the reliability of their conclusions. An important component of self-evaluation is also the model's ability to identify potential limitations and uncertainties in its own responses, which is achieved through the use of special prompts for analyzing so-called edge cases.

Specialized models for assessing specific aspects of LLM quality are narrowly focused neural networks trained on specific datasets to detect characteristics such as toxicity, bias, aggressiveness, or inappropriate content, allowing for detailed automated analysis of generated texts according to specific quality criteria. Such models often use a combination of machine learning methods, including classifiers based on neural network transformers and specialized architectures for detecting subtle linguistic nuances that may indicate problematic content. A crucial aspect here is the continuous updating and retraining of these models on new examples to maintain their relevance and effectiveness. A key advantage of using specialized models is their ability to operate in real time and scale to analyze large volumes of generated content, while ensuring high accuracy and specificity of assessment for specific quality parameters.

4. Effectiveness of quality assessment methods for different criteria

The quality assessment methods presented above will have different effectiveness for different quality criteria. Table 1 provides a comparative analysis of assessment methods for different criteria, indicating which types of methods are the most effective for each criterion.

Expert assessments are the most reliable method for assessing the quality of LLMs for any of the criteria. Expert assessments are the basis for forming reference datasets and training models that perform automatic assessments. At the same time, an obvious disadvantage

of expert assessments is the cost and limited human resources compared to machine ones.

The method with reference data is best suited for assessing the precision and content of LLM responses when there is a dataset with reference responses that can be compared with actual responses.

Automatic methods can be applied to all criteria, but they will be most effective for subjective quality criteria such as naturalness, toxicity, or impartiality, for which it is impossible to formulate a representative reference dataset.

Table. The most effective methods of quality assessment for each of the criteria

	Expert assessment	Benchmark	Automated
Accuracy	Individual expert assessment.	Binary, or non-binary assessment with simple lexical similarity metrics.	LLM-as-judge or self-evaluation (ineffective compared to the benchmark approach)
Completeness	Group expert assessment, with instructions in the form of an assessment scale.	Non-binary assessment with semantic similarity analysis.	LLM-as-judge or self-evaluation (ineffective compared to the benchmark approach)
Natural nalguage	Group expert assessment, with instructions in the form of an assessment scale	Inefficient.	Specialized natural language assessment models.
Consistency	Individual expert assessment.	Binary or non-binary comparison, based on the types of assessed responses.	LLM-as-judge.
Toxicity	Group expert assessment, with instructions on the assessment scale.	Inefficient.	Specialized toxicity assessment models.
Bias	Group expert assessment, with instructions in the form of an assessment scale.	Inefficient.	Specialized bias assessment models.
Security vulnerabilities	Individual expert assessment.	Inefficient.	LLM-as-judge, for malicious prompt generation and the assessment of the model behavior.

5. Conclusion

The article addresses a new problem that has emerged in the context of the rapid development of generative artificial intelligence models and large language models – the multidimensional and comprehensive quality assessment of generated responses. The paper proposes a set of seven criteria for assessing LLMs, which opens up the possibility of formalizing and unifying a comprehensive system for assessing the quality of large language models. The proposed approach allows for a systematic coverage of various aspects of LLM functioning – from technical accuracy to ethical and safety parameters.

The analysis of methods for LLM quality assessment identified three main approaches: expert assessments (manual method), comparison with benchmark data (semi-automated method), and fully automated methods without the use of benchmarks. During the research and analysis, it was established that expert assessments remain the most reliable, though least scalable, assessment method for all quality criteria. Methods of comparison with benchmark data show high efficiency for assessing precision and content relevance of responses, but have limitations when working with subjective criteria, such as toxicity or bias. Automated methods, such as LLM-as-judge and specialized assessment models, demonstrate the best results for fuzzy criteria where the creation of representative reference datasets is practically impossible.

The differences in the effectiveness of quality assessment methods for different types of queries and responses are also emphasized. For structured responses and specific queries, methods using reference data are most effective, while for unstructured responses and open questions, a combined approach with elements of expert assessment is more appropriate.

In summary, the development of combined assessment methodologies that combine the advantages of different approaches for a comprehensive assessment of LLM quality, taking into account the specific nature of their application, is important. The creation of standardized, comprehensive, and dynamic LLM evaluation systems remains an open scientific challenge, the solution of which will have a significant impact on the development of the artificial intelligence industry and the implementation of these technologies in critically important information processing systems.

Conflict of Interest

The authors state that there are no financial orother potential conflicts regarding this work.

References

[1] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A.. A Comprehensive Overview of Large Language Models. 2023. [Online]. Available: https://arxiv.org/abs/2307.06435

[2] Li, Diya & Zhao, Yue & Wang, Zhifang & Jung, Calvin & Zhang, Zhe. (2024). Large Language Model-Driven Structured Output: A Comprehensive Benchmark and Spatial Data Generation Framework. ISPRS International Journal of Geo-Information. 13. 405. 10.3390/ijgi13110405.

[3] Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Yu, L., Liu, Y., Li, J., Xiong, B., & Xiong, D.. Evaluating Large Language Models: A Comprehensive Survey. 2023. [Online]. Available: https://arxiv.org/abs/2310.19736

[4] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Hudson, D. A., Zelikman, E., . Koreeda, Y.. Holistic Evaluation of Language Models. 2022. [Online]. Available: https://arxiv.org/abs/2211.09110

[5] Banerjee, D., Singh, P., Avadhanam, A., & Srivastava, S.. Benchmarking LLM powered Chatbots: Methods and Metrics. 2023. [Online]. Available: https://arxiv.org/abs/2308.04624

[6] Wu, N., Gong, M., Shou, L., Liang, S., & Jiang, D.. Large Language Models are Diverse Role-Players for Summarization Evaluation. 2023. [Online]. Available: https://arxiv.org/abs/2303.15078

[7] Ni, X., & Li, P.. A Systematic Evaluation of Large Language Models for Natural Language Generation Tasks. 2024. [Online]. Available: https://arxiv.org/abs/2405.10251

[8] Cui, W., Zhang, J., Li, Z., Damien, L., Das, K., Malin, B., & Kumar, S.. DCR-Consistency: Divide-Conquer-Reasoning for Consistency Evaluation and Improvement of Large Language Models. 2024. [Online]. Available: https://arxiv.org/abs/2401.02132

[9] Zhao, Y., Zhu, J., Xu, C., & Li, X.. Enhancing LLM-based Hatred and Toxicity Detection with Meta-Toxic Knowledge Graph. 2024. [Online]. Available: https://arxiv.org/abs/2412.15268

[10] L. Baresi, C. Criscuolo and C. Ghezzi, "Understanding Fairness Requirements for ML-based Software," 2023 IEEE 31st International Requirements Engineering Conference (RE), Hannover, Germany, 2023, pp. 341-346, doi: 10.1109/RE57278.2023.00046.

[11] Kolchenko, V., Khoma, V., Sabodashko, D., & Perepelytsia, P. (2024). Exploring large language models' security threats with automated tools. Social Development and Security, 14(6), 81-96. https://doi.org/10.33445/sds.2024.14.6.9

[12] Kendall, M.G. & Gibbons, J.D. (1990). *Rank Correlation Methods* (5th ed.). Oxford University Press. https://archive.org/details/rankcorrelationm0000kend

[13] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. 2019. [Online]. Available: https://arxiv.org/abs/1904.09675

[14] Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., & Liu, Y.. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. 2024. [Online]. Available: https://arxiv.org/abs/2412.05579

[15] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J.. A Survey on LLM-as-a-Judge. 2024. [Online]. Available: https://arxiv.org/abs/2411.15594