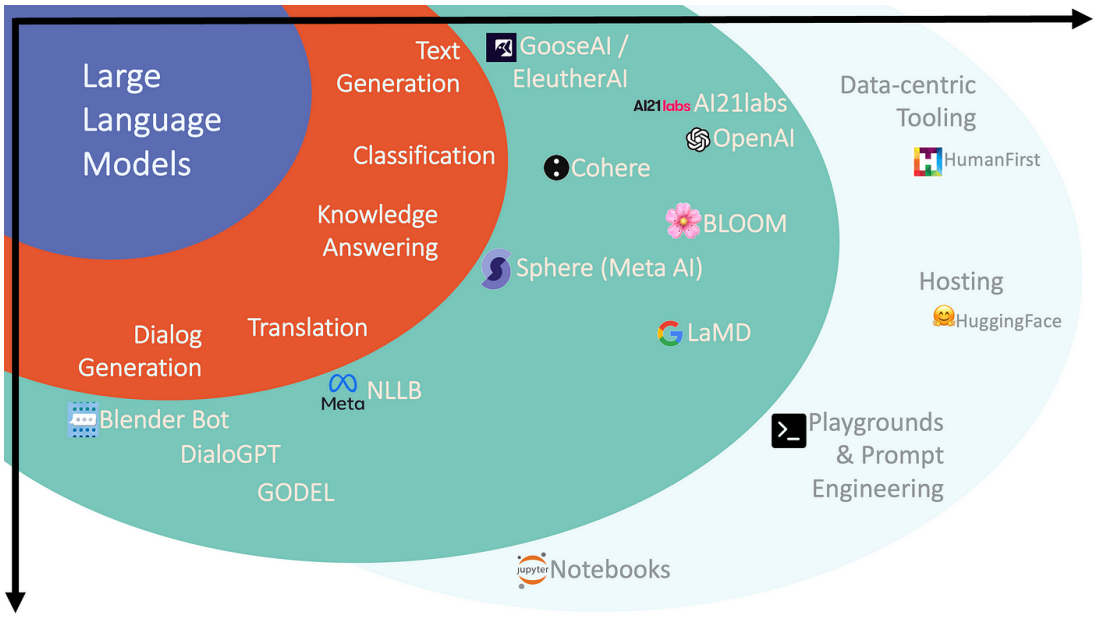
How to Evaluate a Large Language Model (LLM)?

BEGINNER CAREER CHATGPT CLASSIFICATION DATA SCIENCE GITHUB LARGE LANGUAGE MODELS

Introduction

The release of ChatGPT and other Large Language Models (LLMs) signifies a substantial surge in available models. New LLMs emerge frequently. However, a fixed, standardized approach for assessing the quality of these models still needs to be present. This article examines current evaluation frameworks for LLMs and LLM-based systems while analyzing the essential evaluation criteria for LLMs and that how to Evaluate LLMs Models?



Source: Cobus Greyling

Table of Contents

- Frequently Asked Questions

Why Do LLMs Need a Comprehensive Evaluation Framework?

During the early stages of technology development, it is easier to identify areas for improvement. However, as technology advances and new alternatives become available, it becomes increasingly difficult to determine which option is best. This makes it essential to have a reliable evaluation framework that can accurately judge the quality of LLMs.

In the case of LLMs, the immediate need for an authentic evaluation framework becomes even more important. You can employ such a framework to evaluate LLMs in the following three ways:

- A proper framework will help the authorities and concerned agencies to assess the safety, accuracy, reliability, or usability issues of a model.
- Currently, there seems to be a blind race among big tech companies to release LLMs, with many simply placing disclaimers on their products to absolve themselves of responsibility. Developing a comprehensive evaluation framework would help stakeholders to release these models more responsibly.
- A comprehensive evaluation framework will also help users of these LLMs determine where and how to fine-tune these models and with what additional data to enable practical deployment.

In the next section, we will review the current evaluation models.

What Are the Existing Evaluation Frameworks for LLMs?

It is essential to evaluate Large Language Models to determine their quality and usefulness in various applications. Several frameworks have been developed to evaluate LLMs, but none of them are comprehensive enough to cover all aspects of language understanding. Let's take a look at some major existing evaluation frameworks.

Table of the Major Existing Evaluation Frameworks

Framework Name	Factors Considered for Evaluation	Url Link
Big Bench	Generalization abilities	https://github.com/google/BIG-bench
GLUE Benchmark	Grammar, Paraphrasing, Text Similarity, Inference, Textual Entailment, Resolving Pronoun References	https://gluebenchmark.com/
SuperGLUE Benchmark	Natural Language Understanding, Reasoning, Understanding complex sentences beyond training data, Coherent and Well-Formed Natural Language Generation, Dialogue with Human Beings, Common Sense Reasoning (Everyday Scenarios and Social Norms and Conventions), Information Retrieval, Reading Comprehension	https://super.gluebenchmark.com/
OpenAI Moderation API	Filter out harmful or unsafe content	https://platform.openai.com/docs/api-reference/moderations
MMLU	Language understanding across various tasks and domains	https://github.com/hendrycks/test
EleutherAI LM Eval	few-shot evaluation and performance in a wide range of tasks with minimal fine-tuning	https://github.com/EleutherAI/lm-evaluation-harness
OpenAI Evals	Accuracy, Diversity, Consistency, Robustness, Transferability, Efficiency, Fairness of text generated	https://github.com/openai/evals
Adversarial NLI (ANLI)	Robustness, Generalization, Coherent explanations for inferences, Consistency of reasoning across similar examples, Efficiency in terms of resource usage (memory usage, inference time, and training time)	https://github.com/facebookresearch/anli
LIT (Language Interpretability Tool)	Platform to Evaluate on User Defined Metrics. Insights into their strengths, weaknesses, and potential biases	https://pair-code.github.io/lit/
ParlAI	Accuracy, F1 score, Perplexity (how well the model predicts the	https://github.com/facebookresearch/ParlAI

	next word in a sequence), Human evaluation on criteria like relevance, fluency, and coherence, Speed & resource utilization, Robustness (this evaluates how well the model performs under different conditions such as noisy inputs, adversarial attacks, or varying levels of data quality), Generalization	
CoQA	understand a text passage and answer a series of interconnected questions that appear in a conversation.	https://stanfordnlp.github.io/coqa/
LAMBADA	Long-term understanding using prediction of the last word of a passage.	https://zenodo.org/record/2630551#.ZFUKS-zML0p
HellaSwag	Reasoning abilities	https://rowanzellers.com/hellaswag/
LogiQA	Logical reasoning abilities	https://github.com/lgw863/LogiQA-dataset
MultiNLI	Understanding relationships between sentences across different genres	https://cims.nyu.edu/~sbowman/multinli/
SQUAD	Reading comprehension tasks	https://rajpurkar.github.io/SQuAD-explorer/

Also Read: 10 Exciting Projects on Large Language Models(LLM)

The Issue With the Existing Frameworks

Each of the above ways to evaluate the Large Language Models has its own advantages. However, there are a few important factors because of which none of the above seems to be sufficient-

- None of the above frameworks considers safety as a factor for evaluation. Although 'OpenAI Moderation API' addresses it to some extent, that is not sufficient.
- The above frameworks are scattered in terms of factors on which they evaluate the model. None of them is comprehensive enough to be self-sufficient.

In the next section, we will try to list down all the important factors which should be there in a comprehensive evaluation framework.

What Factors Should Be Considered While Evaluating LLMs?

After reviewing existing evaluation frameworks, the next step is determining which factors should be considered when evaluating the quality of Large Language Models (LLMs). We conducted a survey with a group of 12 data science professionals. These people had a fair understanding of how LLMs work and what they can do. They had also tried and tested multiple LLMs. The survey aimed to list down all the important factors, according to their understanding, on the basis of which they judge the quality of LLMs.

Finally, we found that there are several key factors that should be taken into account:

1. Authenticity

The accuracy of the results generated by LLMs is crucial. This includes the correctness of facts, as well as the accuracy of inferences and solutions.

2. Speed

The speed at which the model can produce results is important, especially when it needs to be deployed for critical use cases. While a slower model may be acceptable in some cases, rapid action teams require quicker models.

3. Grammar and Readability:

LLMs must generate language in a readable format. Ensuring proper grammar and sentence structure is essential.

4. Unbiased:

It's crucial that LLMs are free from social biases related to gender, race, and other factors.

5. Backtracking

Knowing the source of the model's inferences is necessary for humans to double-check its basis. Without this, the performance of LLMs remains a black box.

6. Safety & Responsibility

Guardrails for AI models are necessary. Although companies are trying to make these responses safe, there's still significant room for improvement.

7. Understanding the context

When humans consult AI chatbots for suggestions about their general and personal life, it's important that the model provides better solutions based on specific conditions. The same question asked in different contexts may have different answers.

8. Text Operations

LLMs should be able to perform basic text operations such as text classification, translation, summarization, and more.

9. IQ

Intelligence Quotient is a metric used to judge human intelligence and can also be applied to machines.

10. EQ

The emotional Quotient is another aspect of human intelligence that can be applied to LLMs. Models with higher EQ will be safer to use.

11. Versatile

The number of domains and languages that the model can cover is another important factor to consider. It can be used to classify the model into General AI or AI specific to a given set of field(s).

12. Real-time update

A system that's updated with recent information can contribute more broadly and produce better results.

13. Cost

The cost of development and operation should also be considered.

14. Consistency

Same or similar prompts should generate identical or almost identical responses, else ensuring quality in commercial deployment will be difficult.

15. Extent of Prompt Engineering

The level of detailed and structured prompt engineering needed to get the optimal response can also be used to compare two models.

Common Challenges with Existing LLM Evaluation Methods

- **Data Contamination:** Ensuring the quality and integrity of the evaluation data is crucial. Contaminated data can lead to inaccurate assessments of LLM performance.
- **Over-reliance on Perplexity:** Overemphasis on perplexity as a metric may not capture the full range of language understanding and generation capabilities, potentially leading to biased evaluations.
- **Subjectivity in Human Evaluations:** Human evaluations introduce subjectivity, making it challenging to maintain consistency and objectivity in assessing LLM performance.
- **Limited Reference Data:** Limited availability of diverse and high-quality reference data can hinder comprehensive evaluation, especially for models dealing with specialized domains or languages.
- **Lack of Diversity Metrics:** Many evaluation methods lack metrics that specifically measure diversity in responses, which is crucial for assessing the creativity and adaptability of LLMs.
- **Generalization to Real-World Scenarios:** Evaluating LLMs in controlled settings might not reflect their real-world performance, where they must handle a wide range of unstructured and dynamic inputs.
- **Adversarial Attacks:** LLMs can be vulnerable to adversarial attacks, and evaluating their robustness in the face of such attacks is a significant challenge in the evaluation process.

Best Practices to Overcome Challenges

- Ensure transparency in training data sources and methodologies to enhance trustworthiness and accuracy.
- Utilize a spectrum of evaluation metrics, extending beyond perplexity, to comprehensively assess LLM performance.

- Combine automated metrics with human evaluation to capture subjective aspects and nuances in LLM responses.
- Access diverse and high-quality reference data for evaluation, especially for specialized domains and languages.
- Implement metrics that specifically measure diversity in LLM-generated responses to assess creativity and adaptability.
- Integrate real-world scenarios and complex inputs into the evaluation process to gauge the practical utility of LLMs.
- Subject LLMs to robustness evaluations, including adversarial testing, to assess their resistance to malicious inputs and potential vulnerabilities.

Conclusion

The development of Large Language Models (LLMs) has revolutionized the field of natural language processing. However, there is still a need for a comprehensive and standardized evaluation framework for Evaluate LLMs to assess the quality of these models. The existing frameworks provide valuable insights, but they lack comprehensiveness and standardization and do not consider safety as a factor for evaluation.

A reliable evaluation framework should consider factors such as authenticity, speed, grammar and readability, unbiasedness, backtracking, safety, understanding context, text operations, IQ, EQ, versatility, and real-time updates. Developing such a framework will help stakeholders release LLMs responsibly and ensure their quality, usability, and safety. Collaborating with relevant agencies and experts is necessary to build an authentic and comprehensive evaluation framework for LLMs.

Frequently Asked Questions

Q1. How do you evaluate an LLM performance?

A. Evaluating LLM performance involves appraising factors like language fluency, coherence, contextual understanding, factual accuracy, and the ability to generate relevant and meaningful responses. Metrics such as perplexity, BLEU score, and human evaluations can measure and compare LLM performance.

Q2. What are large language models LLM classified as?

A. Large Language Models (LLMs) are advanced natural language processing (NLP) models. They comprehend and generate human-like text by leveraging extensive pre-existing language data and complex machine learning algorithms.

Q3. What is an example of an LLM model?

A. GPT-3 by OpenAI is an example of a well known and influential LLM model. It can generate coherent and contextually relevant text responding to prompts, making it versatile for various NLP tasks.

Q4. What are examples of large language models?

A. Examples of large language models include GPT-3, GPT-2, BERT (Bidirectional Encoder Representations from Transformers), T5 (Text-to-Text Transfer Transformer), and XLNet. These models have undergone extensive training on massive datasets and exhibit strong language generation capabilities across domains and applications.

Article Url - https://www.analyticsvidhya.com/blog/2023/05/how-to-evaluate-a-large-language-model-llm/

Gyan Prakash Tripathi