LLMs

Evaluating Large Language Models: A Comprehensive Guide

Evaluating large language models (LLMs) requires multidimensional strategies to assess coherence, accuracy, and fluency. Explore key benchmarks, metrics, and methods to ensure LLM reliability, transparency, and performance in real-world applications.

Puneet Jindal
May 25, 2024 • 10 min read

Share this blog

Evaluating LLMs on different metrics

Large language models **(LLMs)** are transforming how humans communicate with computers. These advanced AI systems can generate human-quality images, and texts, interpret languages, compose various types of creative content, and provide meaningful responses to the questions we ask.

But with great power comes great responsibility, which in the context of LLMs involves thorough evaluation. LLMs, like any machine learning model, require thorough evaluation to guarantee that their results are precise, trustworthy, and useful.

We'll look at the objective of evaluation, and several evaluation techniques, and discuss the issues and future possibilities in this crucial topic.

In this article, we will discuss:

Table of Contents

To ensure a comprehensive evaluation of LLMs, we need to use extensive strategies that evaluate the model's potential in multiple dimensions (including coherence, diversity, relevance, and language fluency) and traverse further than simply considering the grammatical excellence of a set of prompts.

Why Evaluate LLMs?

As previously said, it is a multidimensional procedure that allows us to comprehend a model's strengths, limitations, and overall performance across several dimensions. Here's why LLM assessment is important:

Ensuring Trust and Accuracy: While LLMs are trained on huge quantities of data, such data may be skewed or biased. These flaws are recognized by evaluation, making sure that the LLM model produces accurate and trustworthy results.

Enhancing Productivity in Certain Tasks: LLMs may be used for a variety of tasks, from generating content writing to offering helpful answers to inquiries.

Evaluation allows us to identify an LLM's advantages and drawbacks in various tasks. Identifying areas for improvement allows developers to fine-tune the LLM model for peak performance.

Shaping Future Development: The observations gained via evaluation are useful to developers and researchers. Learning how LLMs acquire and interpret data allows us to design even better models for the future.

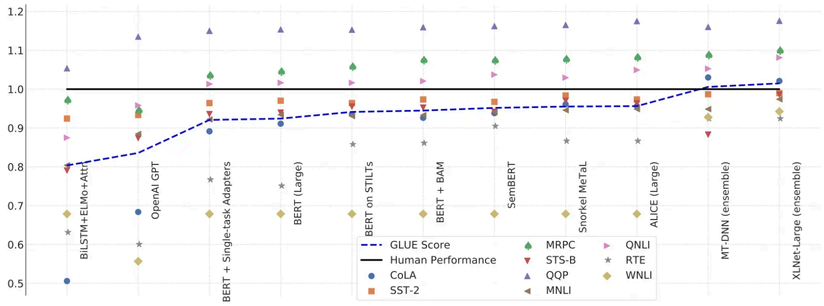
Developing Integrity and Credibility: As LLMs are growing as an integral part of our daily lives, it is critical to understand how they make decisions. Evaluation increases credibility by offering perspective on the model's strengths and limits.

LLM Evaluation Methods

As mentioned above the evaluation for LLM is a multi-dimensional method that examines different abilities needed necessarily in the real world. Here are the different evaluation metrics for LLM-

GLUE(General Language Understanding Evaluation)

GLUE comprises 9 tests that are intended to evaluate an LLM's proficiency in a range of Natural Language Understanding (NLU) tasks. It encompasses a large area. The image below shows the GLUE benchmark performance for most common Language models, rescaled to set human performance to 1.0 for better understanding and calculated across all 9 tests.



The 9 tests included in GLUE are:

CoLA (Corpus of Linguistic Acceptability): This task assesses a model's ability to determine if a given sentence is grammatically correct. It essentially asks the model, "Does this sentence make sense?"

SST-2 (Stanford Sentiment Treebank): This task focuses on sentiment analysis. The model is given a sentence and needs to predict its sentiment, such as positive, negative, or neutral.

MRPC (Microsoft Research Paraphrase Corpus): This task determines if two sentences express the same meaning in different wording. It's essentially asking, "Are these sentences paraphrases of each other?"

STS-B (Semantic Textual Similarity Benchmark): This task measures the semantic similarity between two sentences. The model needs to determine how closely related the meanings of the sentences are.

QQP (Quora Question Pairs): This task focuses on question answering. The model is given a question and a set of candidate answer passages. It needs to identify the passage that best answers the question.

MNLI (Multi-Genre Natural Language Inference): This task assesses a model's ability to understand the logical relationship between two sentences. Given a premise sentence and a hypothesis sentence, the model needs to determine if the hypothesis entails the premise (meaning it follows logically), contradicts it, or is neutral.

QNLI (Question Answering over NLI): This task is similar to MNLI but specifically focuses on questions and answer passages. Given a question and a passage, the model needs to determine if the answer to the question can be found by logical implication within the passage.

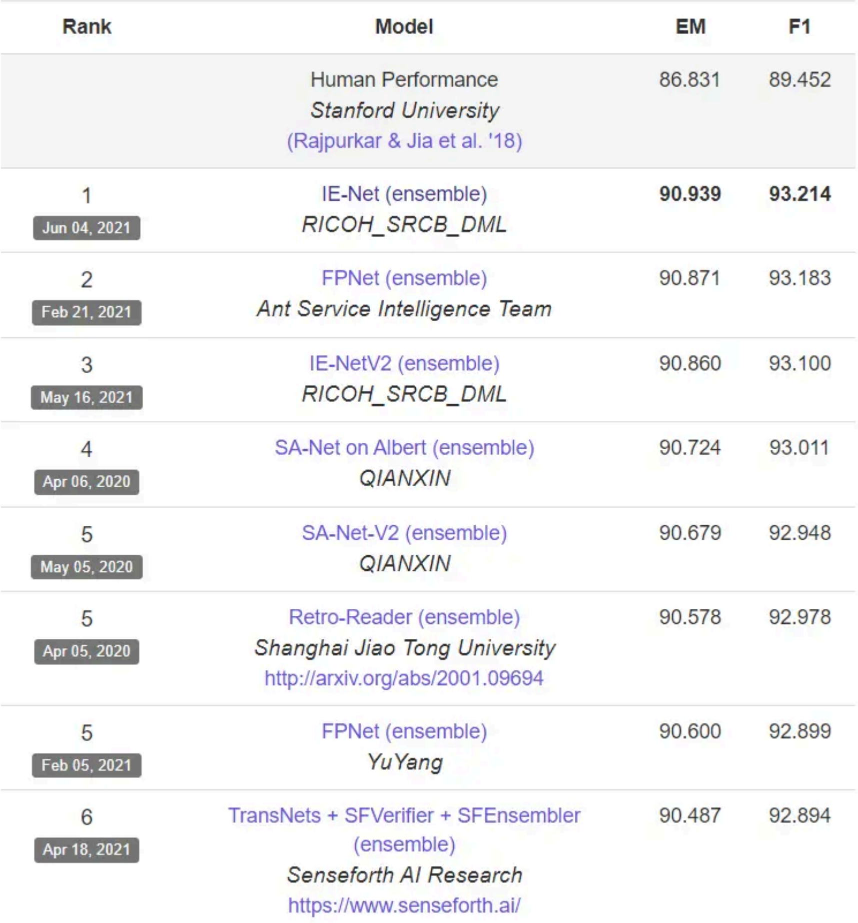
RTE (Recognizing Textual Entailment): Similar to MNLI, this task assesses textual entailment. However, RTE uses a different dataset and focuses more on factual information. The model needs to determine if the information in the hypothesis can be inferred from the premise.

WNLI (Winograd Schema Challenge): This task focuses on resolving ambiguity in language based on common sense. The model is given a sentence with a pronoun and two possible antecedents. It needs to determine which antecedent makes the most logical sense based on the context of the sentence.

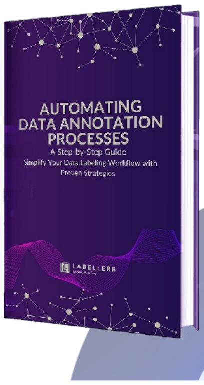
SQuAD 2.0 (Stanford Question Answering Dataset)

SQuAD evaluates the ability to answer factual questions. The dataset is a collection of reading paragraphs with respective questions that require comprehension reading.

The below image shows the EM (exact match) and F1 scores on SQUAD 2.0 benchmarks for different models.



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
5 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
5 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899
6 Apr 18, 2021	TransNets + SFVerifier + SFEnsembler (ensemble) *Senseforth AI Research* https://www.senseforth.ai/	90.487	92.894


AUTOMATING DATA ANNOTATION PROCESSES
A Step-by-Step Guide
Simplify Your Data Labeling Workflow with Proven Strategies
LABELLERR

The metrics involved while evaluating SQuAD 2.0 are **Exact Match (EM)** and **F1 Score** where EM is a binary metric that focuses on whether the predicted answer **exactly matches** a single answer passage in the text. And F1 score is the **harmonic mean** of two common metrics precision and recall.

EM focuses on the exact character match, while the F1 score considers both the relevance and completeness of the answer compared to ground truth.

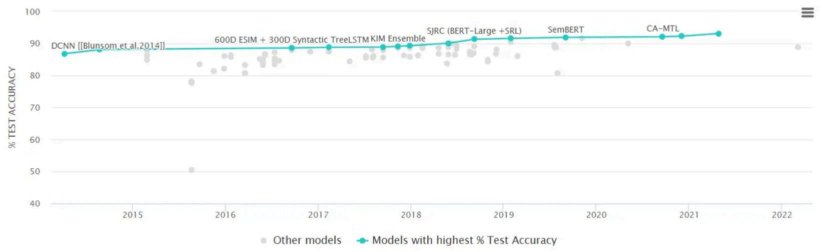
A high EM score indicates the model can perfectly retrieve existing answer passages. A high F1 score suggests the model can identify relevant answer parts even if the wording isn't identical.

SNLI (Stanford Natural Language Inference)

SNLI measures the ability to think critically about the logical connections between sentences.

It is a collection of sentence pairings together with labels that explain the relationship between their entailments where entailment is the truth implied by the first sentence on the second.

The image below shows SNLI on different Models



Benchmark Metrics

Every benchmark has particular metrics for measuring performance in various domains.

Accuracy: It is the simplest statistic, which shows what proportion of tasks (such as answering questions) the LLM correctly answers. It works best for assignments that have straightforward right and wrong solutions.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

where,

True Positive (TP): The ASR system correctly identifies and transcribes spoken language. For example, someone says "Hello" and the system transcribes it as "Hello".

False Positive (FP): The ASR system incorrectly identifies non-speech sounds or silence as spoken language and attempts to transcribe it. For example background noise is mistakenly transcribed as a word or phrase, such as interpreting static noise as "cat".

False Negative (FN): The ASR system fails to recognize and transcribe actual spoken language. For example, someone whispers "help" but the system does not detect it and provides no transcription.

True Negative (TN): The ASR system correctly identifies non-speech sounds or silence and does not attempt transcription. For example, the system detects silence or non-speech sounds like a chair creaking and correctly avoids transcribing anything.

Precision: Let's say you ask LLM a question, and it responds with ten possible responses. The precision of an answer indicates the percentage of those that are related to the question asked.

$$Precision = \frac{TP}{TP + FP}$$

High precision shows that the LLM can avoid unnecessary deviations and maintain focus on the current activity.

Recall: This metric is the reverse of precision. Let's assume that your inquiry has twenty valid potential solutions. Recall evaluates how successfully the LLM located every pertinent response.

$$Recall = \frac{TP}{TP + FN}$$

A high recall indicates that the LLM is thorough and does not overlook any significant information.

F1 Score: It's critical to strike a balance between recall and precision. Consider a situation in which the LLM finds a lot of answers with good recall but little precision (the majority of responses are unimportant).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

In order to account for this, the F1 score takes the harmonic mean of recall and precision. Extreme values (extremely high recall, very poor precision) are penalized by a harmonic mean, which rewards a balance between the two.

BLEU Score (Bilingual Evaluation Understudy)

BLEU evaluates the degree to which the text generated by the LLM is comparable to human-written material, going beyond simple factual accuracy.

It contrasts the produced text with professionally translated human references. Through the analysis of factors such as n-gram overlap (the frequency with which word sequences occur in both the produced text and that reference—BLEU generates a score that accounts for readability, fluency, and grammatical accuracy.

METEOR (Metric for Evaluation of Translation with Ordering)

Like BLEU, prioritizes coherence and fluency and additionally takes word order into account.

When an LLM produces text with words arranged naturally and correctly (i.e., imitating human sentence structure), METEOR awards the LLM for this effort.

Real-World Considerations

The standard benchmark serves as the foundation, but their limited scope, data bias, and emphasis on factual accuracy make them insufficient to evaluate the LLM model.

Therefore, in order to obtain further evaluation, we must fulfill these criteria using:

Human Evaluation: In this process, the output reliability of the LLM is evaluated by bringing in human professionals. They are able to evaluate elements such as safety, engagement, factuality, coherence, and fluency.

Domain-Specific Evaluation: Customized metrics and datasets are required for several tasks. For example, measures such as legal accuracy and compliance with specified formatting rules would be utilized to evaluate an LLM for legal document generation.

Advanced Evaluation Methods

Calibration and Fidelity

LLMs frequently show high confidence despite the scenario of inaccurate results. Users who depend on confidence scores to judge the reliability of the information may find this to be deceptive.

This is addressed by fidelity and calibration, which measure the degree to which an LLM's confidence score corresponds with the accuracy of its response.

A properly calibrated LLM has confidence scores that appropriately depict a range of options depending on the chance that the answer is correct, with high confidence for correct responses and low confidence for incorrect responses.

Evaluating adversarially

This method involves developing prompts with the sole purpose of taking benefit of LLM vulnerabilities.

These prompts can be open-ended, which encourages the LLM to generate irrelevant responses and deviate from the task upfront, deceptive which tries to fool the LLM into producing factually wrong outputs, or biased, which may originate from the training data.

Adversarial evaluation assists researchers in identifying places where the LLM requires improvement by highlighting these flaws.

Explainability (XAI)

Large LLMs in particular can be complicated "black boxes." We may receive an output, but it may not always be clear how the LLM reached its conclusion.

Debugging LLMs, spotting biases, and gaining confidence in their decision-making all depend on explainability. We can guarantee that the LLM is functioning as intended and enhance its performance by knowing how it functions.

Challenges in LLM Evaluation

LLM evaluation techniques have taken a step but there is still a mile to cover and many challenges to cover, among them a few are:

Dataset Bias: LLM's output is affected by reinforced biases which are on training data and benchmarks. The collection and careful choice of datasets are essential.

Metric Restrictions: Certain metrics, such as BLEU, may prioritize fluency over factual correctness. A response that is insightful yet factually inaccurate could receive a high BLEU score. These constraints can be addressed by combining human judgment with different measures.

Expense and Duration: It can take a lot of resources to produce top-notch evaluation datasets and carry out human evaluation.

How Labellerr Helps In LLM Evaluation

Labellerr aims to make LLM evaluation more accessible, efficient, and informative. By providing high-quality data, efficient techniques, and targeted tools, Labellerr helps developers gain deeper insights into their LLMs' performance and identify areas for improvement.

Labellerr addresses some of the challenges mentioned above by focusing on areas like data quality, efficiency, and specific functionalities

Data Efficiency

Labeling data for LLM evaluation can be expensive and time-consuming. Labellerr's approach emphasizes data efficiency techniques to get the most out of limited data sets. This can significantly reduce the time and resources required for thorough evaluation.

Focus on Specific Evaluation Needs

LLM evaluation can involve various tasks like identifying hallucinations (generated text that appears real but isn't factual) or assessing factual consistency. Labellerr offers targeted evaluation tools for specific needs, allowing for a more nuanced assessment of an LLM's strengths and weaknesses.

Integration with Existing Tools

Labellerr integrates with existing LLM development workflows, streamlining the evaluation process. This allows developers to seamlessly incorporate evaluation into their LLM development cycle.

Conclusion

Evaluation is a continuous quest. Developers can obtain a thorough insight into an LLM's advantages and drawbacks by utilizing a combination of unique metrics, established benchmarks, human review, and advanced methodologies. The safe, dependable, and accurate deployment of LLMs in practical applications depends on this continual review cycle.

Frequently Asked Questions

1) What are LLM evaluation's primary objectives?

- Evaluate the model's ability for a range of Natural Language Understanding (NLU) applications.
- Evaluate the produced text's fluency, quality, and quality.
- Determine any possible biases, dangers to safety, or factual errors.
- Make sure the LLM expresses confidence scores that appropriately reflect the accuracy of its results and is calibrated properly.
- Recognize the process by which the LLM generates its results (explainability).

2) How many different types of LLM evaluation techniques are there?

- Standard Benchmarks: GLUE, SQuAD, and SNLI are examples of predefined tasks that evaluate fundamental NLU skills.
- Metrics: Various characteristics of performance are quantified by metrics such as accuracy, precision, recall, F1 score, BLEU score, METEOR score, etc.
- Human Evaluation: Professionals evaluate elements such as engagement, safety, coherence, fluency, and factuality.
- Domain-Specific Assessment: certain application-specific datasets and metrics (e.g., legal document creation).
- Advanced Methods: Explainability (XAI), Adversarial Evaluation, and Fidelity & Calibration.

3) What drawbacks do conventional benchmarks have?

- Restricted scope: Real-world application's distinctive features may not be fully captured by benchmarks.
- Data bias: Biased outcomes can result from the prolongation of biases through benchmarks and training data.
- Target for factual accuracy: Some factors, such as creative thinking or user involvement, may not be evaluated by benchmarks.
-

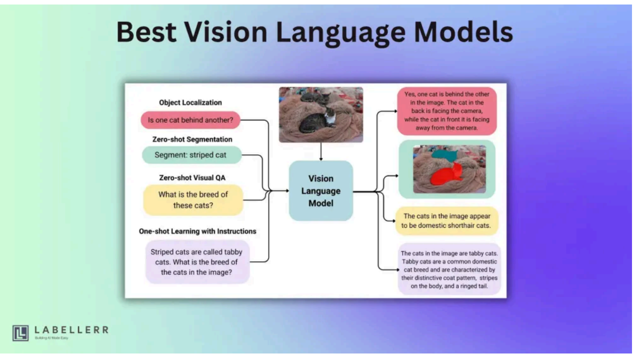
👋 **About Author**Puneet is a co-founder and CEO at Labellerr.He is working towards the vision to remove the high-quality data scarcity for model training. He loves to discuss and write about new developments in the field of AI.Connect with him over Linkedin or write an email at puneet.jindal@labellerr.com

References

1. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems (Link)
2. Natural Language Inference on SNLI (Link)

Sign up for more like this.

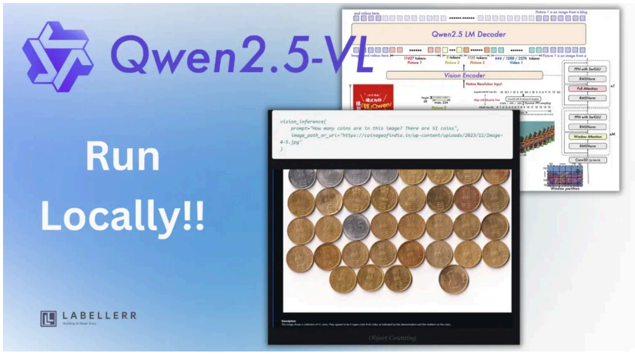
Enter your email Subscribe


Best Vision Language Models
Object Localization
Is one cat behind another?
Zero-shot Segmentation
Segment: striped cat
Zero-shot Visual QA
What is the breed of these cats?
One-shot Learning with Instructions
Striped cats are called tabby cats. What is the breed of the cats in the image?
Vision Language Model
The cats in the image appear to be domestic shorthair cats.
LABELLERR

Best Open-Source Vision Language Models of 2025

Discover the leading open-source vision-language models (VLMs) of 2025 including Qwen 2.5 VL, LLaMA 3.2 Vision, a…

Jun 4, 2025 5 min read



Run Qwen2.5-VL 7B Locally: Vision AI Made Easy

Discover how to deploy Qwen2.5-VL 7B, Alibaba Cloud's advanced vision-language model, locally using Ollama. This…

Jun 4, 2025 5 min read

High Performer WINTER 2025

High Performer Asia WINTER 2025

High Performer Small Business WINTER 2025

Leader India WINTER 2025

High Performer Asia Pacific WINTER 2025

Easiest To Use WINTER 2025

Easiest To Do Business With Small Business WINTER 2025

Platform

Collect

Curate

Data Annotation Platform

Label GPT

Datasets

Pricing

Solutions

LLM

Automotive

Healthcare

Security & Surveillance

Agritech

Retail

Biotechnology

Company

About Us

Careers

Privacy

Contact Us

Terms & Conditions

Learn

Blog

Case Studies

Expert discussions

FAQ

Knowledge Base

SDK Documentation

Compare

Labellerr vs Roboflow

Labellerr vs Encord

Labellerr vs Dataloop

Labellerr vs Supervise

Contact

US Office
Tensor Matics Inc
44, Tehama St,
San Francisco, CA
USA 94107
Phone:+16283133187

Registered Office
651 N Broad St, Suite 201,
w
e

We're offline
Leave a message

Pre Label Datasets

Smart Feedback Loop

Interactive Demo

Product Demo

Image Annotation Platform

Text Annotation Platform

Video Annotation Platform

Annotation Services

Energy

Sports Vision

Manufacturing

Labellerr vs AWS Sagemaker

Labellerr vs CVAT

Labellerr vs Appen

Labellerr vs Labelbox

Labellerr vs V7 Labs

Labellerr vs SuperAnnotate

India Office:
Tensor Matics Pvt Ltd
SCO 224, Level 1 and 2,
Sector 37 C, Chandigarh,
160036, India
Phone: +917565883102

Chat on WhatsApp

support@tensormatics.com

Capterra 4.6

4.9 out of 5

Making AI journey simple!

We're offline
Leave a message