Advanced Generative AI: Models, Tools and Applications

Benchmark and Evaluation of LLM Capabilities: Part 2

Quick Recap

Lorem ipsum dolor sit amet, adipiscing elit, sed do incididunt ut labore magna aliqua.?

- What constitutes the primary capabilities of LLMs?
- What constraints or drawbacks are associated with the capabilities of LLMs?

Engage and Think

Imagine being a part of a cutting-edge tech team tasked with developing an advanced virtual assistant designed to enhance everyday life. The team has access to powerful LLMs, but to ensure your virtual assistant truly understands and responds effectively to users, you need to master the art of benchmarking and evaluating LLM capabilities.

How would you design a benchmark that evaluates the LLMs’ capability?

Learning Objectives

By the end of this lesson, you will be able to:

- Analyze HELM for large language model evaluation
- Apply GLUE principles for practical language understanding enhancement
- Assess the SuperGLUE benchmark results to make informed decisions
- Utilize the BIG-bench benchmark to create new testing scenarios

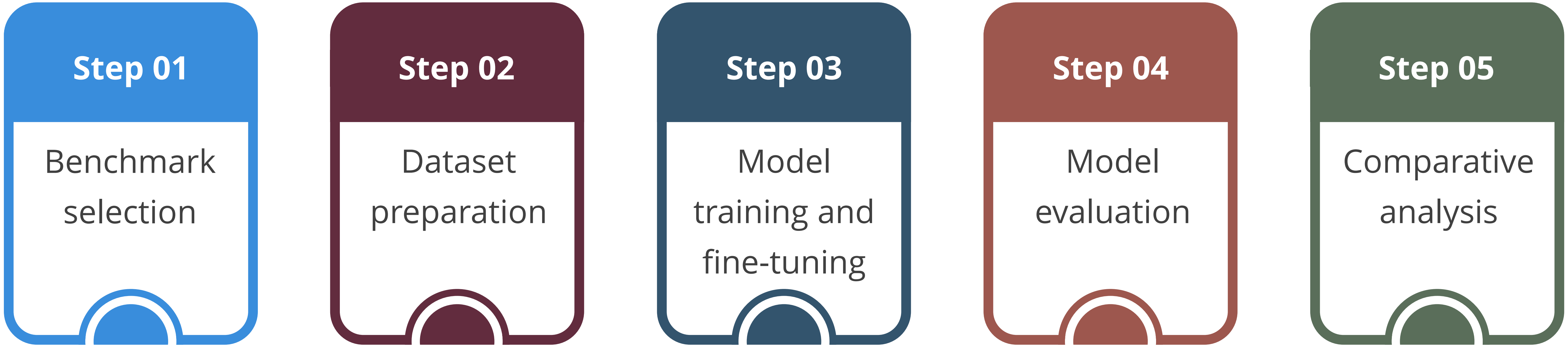
Introduction to Benchmarking

What Is Benchmarking?

Benchmarking in large language models (LLMs) is a process of evaluating and comparing the performance of various LLMs through standardized tests or tasks.

LLM Benchmarking: Steps

Benchmarking in LLMs involves the following steps:



Step 01

Benchmark selection

Step 02

Dataset preparation

Step 03

Model training and fine-tuning

Step 04

Model evaluation

Step 05

Comparative analysis

LLM Benchmarking: Steps

Benchmarking in LLMs involves the following steps:

Step 01

Benchmark selection

A benchmark is chosen to encompass a broad spectrum of challenges related to language.

Step 02

Dataset preparation

Specific datasets are meticulously curated for each benchmark task, comprising training, validating, and testing sets.

LLM Benchmarking: Steps

Benchmarking in LLMs involves the following steps:

Step 03

Model training and fine-tuning

The LLMs are trained or fine-tuned specifically for benchmarking tasks using separate datasets designed for this purpose, distinct from those used in the benchmarking evaluation phase.

Step 04

Model evaluation

The performance of the LLMs is assessed against the benchmark tasks.

LLM Benchmarking: Steps

Benchmarking in LLMs involves the following steps:

Step 05

Comparative analysis

The outcomes are scrutinized to compare the effectiveness of distinct LLMs.

Note

Benchmarking aims to thoroughly measure and compare LLMs, highlighting their strengths and weaknesses.

Quick Check

What is the purpose of the model evaluation step in the language model development process?

A. To choose benchmarks that encompass diverse language challenges.

B. To prepare datasets for training, validation, and testing.

C. To fine-tune the model for benchmarking tasks.

D. To assess the model’s performance against predefined benchmarks.

Benchmarks for Evaluating LLMs

Benchmarks for Language Models

Metrics such as ROUGE and BLEU scores are used to evaluate language models for summarization or text-generation tasks.

Demo: ROUGE Benchmark

Duration: 40 minutes

Overview:

This demo is designed to read a PDF file and a summary of that file, and then compute the ROUGE scores for the summary by comparing it with the original document. The ROUGE scores provide a measure of the quality of the summary.

Note

Please download the solution document from the Reference Material Section and follow the Jupyter Notebook for step-by-step execution.

Need for New Benchmarks

LLMs transcend traditional language models, showcasing unparalleled proficiency in tasks beyond linguistic understanding.

- Existing benchmarks for language models are inadequate for evaluating these advanced LLMs.
- A compelling need exists for the creation of new benchmarks specifically tailored to assess the enhanced functionalities of LLMs.

Benchmarks for LLMs

There are some metrics to measure these models with billions of parameters:

- HELM
- GLUE
- SuperGLUE
- BIG-bench

Holistic Evaluation of Language Models (HELM)

It was developed by the Center of Research on Foundation Models (CRFM) at Stanford University.

The HELM benchmark tasks include:

Text generation

Translation

Question answering

Code generation

Commonsense reasoning

General Language Understanding Evaluation (GLUE)

It is a collection of nine different natural language understanding tasks, introduced by researchers at Google AI.

The GLUE benchmark tasks include:

- Natural language interface
- Sentiment analysis
- Corpus of linguistic acceptability
- Sentence similarity
- Textual entailment
- Multi-genre natural language interface
- Paraphrasing
- WordNet hypernymy
- Question natural language interface

GLUE Benchmark Tasks

Natural language interface (NLI)

This task involves determining whether a given hypothesis matches, opposes, or doesn't relate to a given statement.

Sentence similarity (STS)

This task involves assessing the degree of similarity in meaning between two given sentences.

Paraphrasing (QQP)

This task involves establishing whether two sentences serve as paraphrases of one another.

GLUE Benchmark Tasks

Sentiment analysis (SST-2)

This task involves determining if a given sentence carries a positive, negative, or neutral sentiment.

Textual entailment (RTE)

This task involves determining if a provided hypothesis is implied by a given premise.

WordNet hypernymy (WNLI)

This task involves determining if a given word is a broader or more general term (hypernym) in comparison to another provided word.

GLUE Benchmark Tasks

Corpus of linguistic acceptability (CoLA)

This task involves determining the grammatical correctness of a provided sentence.

Multi-genre natural language interface (MNLI)

This task is like NLI, but the hypothesis might consist of multiple sentences.

Question natural language interface (QNLI)

This task is like NLI, but the hypothesis is consistently framed as a question.

SuperGLUE

It is a collection of eight different natural language understanding tasks, introduced by researchers at Facebook AI, Google AI, the University of Washington, and New York University.

The SuperGLUE benchmark tasks include:

- BoolQ
- CB
- COPA
- MultiRC
- Record
- SNLI
- WIC
- WiQA

SuperGLUE Benchmark Tasks

BoolQ

This task involves responding to factual yes-or-no questions.

CB

This task involves determining whether a given sentence is common sense or not.

COPA

This task involves deducing an accurate response to a question from the context provided in two sentences.

MultiRC

This task involves responding to multiple-choice questions related to a series of reading comprehension passages.

SuperGLUE Benchmark Tasks

Record

This task involves completing the blanks in a brief passage of text.

SNLI

This task involves determining if a provided sentence is entailed, contradictory, or neutral to a given premise.

WIC

This task involves selecting the most appropriate word or phrase to complete a sentence with a blank.

WiQA

This task involves responding to questions related to articles on Wikipedia.

Beyond the Imitation Game Benchmark (BIG-bench)

It serves as a collaborative benchmark designed to examine large language models and anticipate their future capabilities.

The BIG-bench benchmark tasks include:

- Text generation
- Translation
- Question answering
- Code generation
- Commonsense reasoning
- Social reasoning
- Logical reasoning

Quick Check

Which SuperGLUE task entails responding to multiple-choice questions?

A. COPA

B. MultiRC

C. BoolQ

D. Record

Guided Practice

Overview

Duration: 45 minutes

In this activity, you will create a PDF summarizer and will benchmark the summary using ROUGE. This guided practice aims to reinforce your skills and familiarity with these technologies, ensuring a more comprehensive understanding of future practical applications in the respective field.

Key Takeaways

- The HELM benchmark is a comprehensive evaluation of LLMs.
- GLUE is a collection of nine different natural language understanding tasks.
- SuperGLUE is a collection of eight different natural language understanding tasks.
- BIG-bench acts as a collaborative benchmark designed to evaluate LLMs.

Q&A