Introduction to LLM Evaluation

Yogesh Haribhau Kulkarni

Outline

YHK

LLM Evaluation

Why Evaluate?

- Like Teacher grading your exam essay!!
- Was it really 6/10 or it should have been 8/10?
- Words for everyone are different, but still the Teacher has to assess, fairly?
- Subjectivity? Bias? Numerical output for Qualitative answers.

Why Evaluate?

- Identifying strengths and weaknesses
- User trust and reliability
- Resource Optimization
- Ethical and Bias considerations
- Regulatory Compliance
- Variability in models
- Model Improvements
- Usability

(Ref: The Science of LLM Benchmarks: Methods, Metrics, and Meanings — LLMOps - LLMOps Space)

YHK

How to Evaluate?

- Exact matching approach
- Similarity approach
- Functional Correctness
- Evaluation Benchmarks
- Human Evaluation
- Model based Approaches (cross val)

(Ref: Evaluating LLMs - Rajiv Shah)

Reliability of Leader-board

Model	Revision	Average	ARC (25-shot)	HellaSwag (10-shot)	MMLU (5-shot)	T
llama-65b	main	58.3	57.8	84.2	48.8	4
llama-30b	main	56.9	57.1	82.6	45.7	4
stable-vicuna-13b	main	52.4	48.1	76.4	38.8	4
llama-13b	main	51.8	50.8	78.9	37.7	3
alpaca-13b	main	51.7	51.9	77.6	37.6	3
llama-7b	main	47.6	46.6	75.6	34.2	3
EleutherAI/gpt-neox-20b	main	45.9	45.2	73.4	33.3	3
togethercomputer/RedPajama-INCITE-Base-7B-v0.1	main	45.7	44.4	71.3	34	3
togethercomputer/RedPajama-INCITE-Base-3B-v1	main	42.2	40.2	64.7	30.6	3
Salesforce/codegen-16B-multi	main	39.2	33.6	51.2	28.9	4
facebook/opt-1.3b	main	37.7	29.6	54.6	27.7	3
facebook/opt-350m	main	32.2	23.6	36.7	27.3	4
facebook/opt-125m	main	31.2	23.1	31.5	27.4	4
gpt2	main	30.4	21.9	31.6	27.5	4

(Ref: Evaluating LLMs - Rajiv Shah)

Metrics

What is a Metric

- Given "supervised data" how do we evaluate?
- Example: Summarizing news articles - metrics may include:
  - Run the model on the 'inputs' to get the 'predictions'.
  - Define the 'metric' (or score) that estimates how well the model 'predictions' reflect the 'gold' 'outputs'.
  - Compute the metric
- How to compute the score?
  - Compute it (Automatic Evaluation)
  - Let humans do it (Human Evaluation)

(Ref: LLM Evaluation Basics: Datasets & Metrics - Generative AI at MIT

Automatic Evaluation

Task	Metric	Automatic Scoring Function
Classification	Accuracy	Exact Match: Did the model predict the same output as the gold output?
Question Answering	F1 Score	How many words are in common between the prediction and gold output?
Translation	ROUGE/BLEU	How many words/phrases are in common between the prediction and gold output?
Program Synthesis	Accuracy	Does the predicted code produce the same result as the output when run?
...	...	...

(Ref: LLM Evaluation Basics: Datasets & Metrics - Generative AI at MIT

Human Evaluation

- Some tasks need more nuanced evaluation which cannot be done automatically
- Example: text generation
- Humans, Crowd Turker, compares model answers with the real answers, against:
  - Coherence, readability, fluency
  - Grammaticality
  - Extend to which the model follows instructions
- Can be done via preference judgment

- Example: Thinking about [insert assessed quality], rate the following passage on a scale of 1 to 5 with 1 being the worst and 5 being the best.
- Example: The generated story follows the instructions (e.g., includes all characters). How much do you agree with this statement?

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

(Ref: LLM Evaluation Basics: Datasets & Metrics - Generative AI at MIT

LLM: Difference in eval vs Classical ML

- LLMs are to be evaluated for
  - Knowledge
  - Reasoning
- Generative nature makes evaluation complex: semantic matching, format, length etc
- Subjectivity of the textual output
- Evaluation needs expertise, domain knowledge and reasoning
- Must be fast and cheap (even though LLMs are huge and take lots of time to infer)

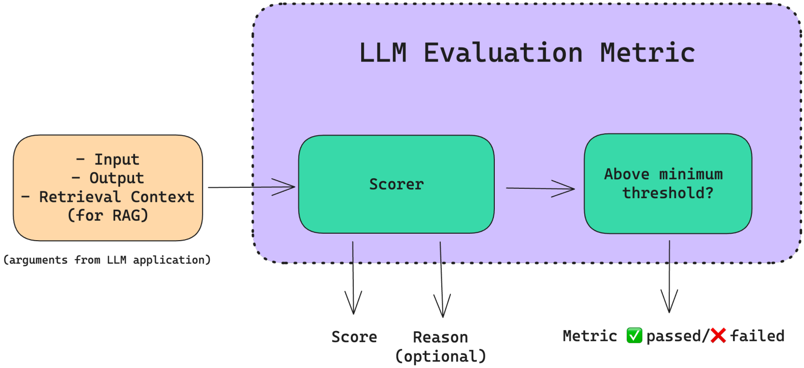
(Ref: Evaluating LLM Models for Production Systems: Methods and Practices - Andrei Lopatenko

YHK

Evaluating Large Language Models

- **Understanding Needs:** Crucial to evaluate whether LLMs meet specific requirements.
- **Clear Metrics:** Establish clear metrics to gauge the value added by LLM applications.
- **Comprehensive Evaluation:** Encompasses assessing the entire pipeline, including prompts, retrieved documents, and processed content.
- **Pipeline Evaluation:**
  - Assess the effectiveness of individual components within the LLM pipeline.
  - Includes evaluation of prompts and quality of retrieved documents.
- **Model Evaluation:**
  - Evaluate the performance of the LLM model itself.
  - Focus on the quality and relevance of its generated output.
- **Prompt Quality:** Assess the appropriateness and effectiveness of prompts used for LLMs.
- **Document Retrieval Quality:** In RAG use-cases, evaluate the quality of retrieved documents.
- **Output Quality:** Evaluate the quality of the generated output by the LLM.

What are LLM Evaluation Metrics?



(Ref: LLM Evaluation Metrics: Everything You Need for LLM Evaluation - Jeffrey Ip

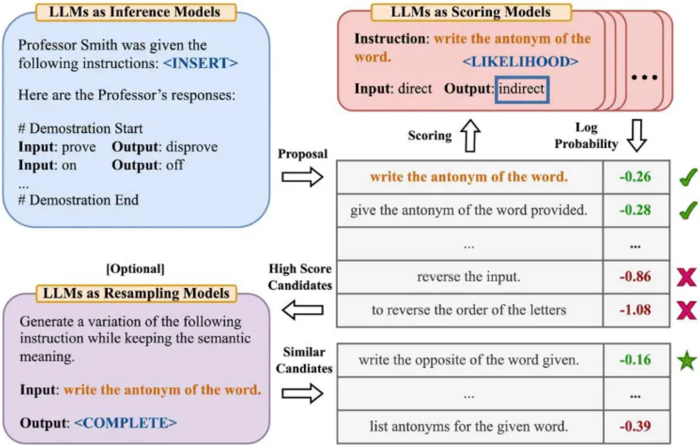
LLM Evaluation Metrics

- Metrics for scoring an LLM's output based on specific criteria.
- Example: Summarizing news articles - metrics may include:
  - Sufficient information in the summary.
  - Absence of contradictions or hallucinations.
- For RAG-based architecture, assess the quality of the retrieval context.
- LLM evaluation metrics align with the tasks designed for the application.
- Note: LLM application can be the LLM itself.

LLM Pipeline Evaluation

- **Types of Evaluation:**
  - Evaluating Prompts.
  - Evaluating the Retrieval Pipeline.
- **Evaluating Prompts:**
  - Evaluate prompts' impact on LLM output.
  - Utilize prompt testing frameworks.
  - Tools like Promptfoo, PromptLayer, etc., are commonly used.
- **Automatic Prompt Generation:**
  - Recent methods automate prompt optimization.
  - Example: Automatic Prompt EngineerAPE.
- **Evaluating Retrieval Pipeline:**
  - Essential for LLM pipelines, especially RAG use-cases.
  - Assessing top-k retrieved documents' quality.

LLM Pipeline Evaluation



(Ref: Applied LLMs Mastery 2024 - Aishwarya Reganti)

YHK

Example

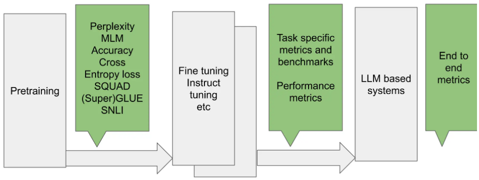
```
Hint

Question: What is the capital of France?

High context relevancy: France, in Western Europe, encompasses medieval
      cities, alpine villages and Mediterranean beaches. Paris, its capital,
      is famed for its fashion houses, classical art museums including the
      Louvre and monuments like the Eiffel Tower.

Low context relevancy: France, in Western Europe, encompasses medieval cities,
      alpine villages and Mediterranean beaches. Paris, its capital, is famed
      for its fashion houses, classical art museums including the Louvre and
      monuments like the Eiffel Tower. The country is also renowned for its
      wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's
      Roman theater and the vast Palace of Versailles attest to its rich
      history.
```

Dimensions of LLM Evaluation

- **Relevance Metrics:** Assess pertinence of response to user's query and context.
- **Alignment Metrics:** Evaluate alignment with human preferences. Consider fairness, robustness, and privacy.
- **Task-Specific Metrics:** Gauge LLM performance across various tasks. Examples: multihop reasoning, mathematical reasoning, etc.

YHK

Pipeline



(Ref: EEvaluating LLM Models for Production Systems: Methods and Practices - Andrei Lopatenko

Evaluation of RAG - RAGAs

Faithfulness consistency of the answer with the context (but no query!) Two LLM calls, get context that was used to derive to answer, check if the statement supported by the context	**Context Relevance** how is the retrieved context "focused" on the answer , the amount of relevant info vs noise info , uses LLM to compute relevance of sentences / total number of retrieved sentences
Answer Relevancy is the answer relevant to the query , LLM call, get queries that may generate answer, verify if they are similar to the original query	**Context Recall** (ext, optional) if all relevant sentences are retrieved , assuming existence of ground_truth answer

(Ref: Evaluating LLM Models for Production Systems: Methods and Practices - Andrei Lopatenko

YHK

Relevance Metrics

- **Introduction:**
  - Evaluation metrics focusing on response relevance.
- **Common Metrics:**
  - **Perplexity:** Measures text prediction quality. Lower values indicate better performance.
  - **Human Evaluation:** Human assessors judge relevance, fluency, coherence, and overall quality.
  - **BLEU (Bilingual Evaluation Understudy):** Compares generated output with a reference answer. Higher scores indicate better performance.
- **Diversity Metric:**
  - Measures variety and uniqueness of LLM responses.
  - Includes n-gram diversity or semantic similarity metrics.
  - Higher scores indicate more diverse and unique outputs.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):**
  - Evaluates LLM-generated text quality by comparing it with reference text.
  - Assesses precision, recall, and F1-score.
  - Provides insights into similarity between generated and reference texts.

YHK

RAG Specific Relevance Metrics

- **Introduction:**
  - RAG pipelines employ specific relevance metrics beyond generic ones.
- **Faithfulness (From RAGas Documentation):**
  - Measures factual consistency of the generated answer against the provided context.
  - Calculated from answer and retrieved context, scaled to (0,1) range.
  - Higher score indicates better faithfulness.
- **Faithfulness Calculation:**
  - Identify claims in the generated answer.
  - Cross-check each claim with the given context for inference.
  - Faithfulness score is based on the ability to infer claims from context.

$$\frac{|\text{Number of claims in the generated answer that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

YHK

Answer Relevance (From RAGas Documentation)

- Focuses on assessing how pertinent the generated answer is to the given prompt.
- Scores between 0 and 1, where higher scores indicate better relevancy.
- Emphasizes completeness and avoids redundancy in answers.

```
Hint

Question: Where and when was Einstein born?

Context: Albert Einstein (born 14 March 1879) was a German-born theoretical
     physicist, widely held to be one of the greatest and most influential
     scientists of all time

High faithfulness answer: Einstein was born in Germany on 14th March 1879.

Low faithfulness answer: Einstein was born in Germany on 20th March 1879.
```

RAG Specific Relevance Metrics

- **Assessment Criteria:**
  - Relevance based on how well the answer addresses the original question.
  - Importance given to completeness, penalizing incomplete or redundant answers.
  - Evaluation does not directly consider factuality.
- **Scoring Process:**
  - LLM prompted to generate an appropriate question for the answer multiple times.
  - Mean cosine similarity between generated questions and the original question is measured.
  - Higher scores indicate better alignment between generated answer and the original question.
- **Answer Semantic Similarity (From RAGas Documentation):**
  - Assesses semantic resemblance between the generated answer and the ground truth.
  - Values range from 0 to 1, with higher scores indicating better alignment.
  - Utilizes a cross-encoder model for calculating semantic similarity.

YHK

Alignment Metrics in LLMs

- **Importance of Alignment Metrics:**
  - Crucial, especially in applications directly interacting with people.
  - Ensures conformity to acceptable human standards.
  - Difficulty in mathematical quantification; relies on specific tests and benchmarks.
- **Evaluation Challenge:**
  - Difficult to quantify alignment metrics mathematically.
  - Adoption of indirect measures through tests on specialized benchmarks.
  - No universally correct method for evaluation.
- **Dimensions for Alignment Evaluation:**
  - Truthfulness: Accurate representation of information.
  - Safety: Avoidance of unsafe or illegal outputs, promotion of healthy conversations.
  - Fairness: Prevention of biased outcomes, assessment of stereotypes and biases.
  - Robustness: Stability and performance across various input conditions.
  - Privacy: Preservation of human and data autonomy, evaluation of privacy awareness.

Alignment Dimensions

- **More Alignment Dimensions:**
  - Machine Ethics: Challenges in defining machine ethics, divided into implicit ethics, explicit ethics, and emotional awareness.
  - Transparency: Concerns the availability of information about LLMs and their outputs.
  - Accountability: Ability to autonomously provide explanations for behavior.
  - Regulations and Laws: Abiding by rules and regulations posed by nations and organizations.
- **Detailed Analysis:**
  - Each dimension further dissected into specific categories.
  - Example: Truthfulness segmented into misinformation, hallucination, sycophancy, and adversarial factuality.
  - Corresponding datasets and metrics designed for quantification.

Task-Specific Metrics

- **Introduction:**
  - Tailored benchmarks are essential for task-specific LLM evaluation.
  - Custom datasets and metrics for specific performance assessment.
- **GLUE (General Language Understanding Evaluation):**
  - Collection of nine tasks measuring English text understanding.
  - Includes sentiment analysis, question answering, and textual entailment.
- **SuperGLUE:**
  - Extension of GLUE with more challenging comprehension tasks.
  - Involves word sense disambiguation, complex question answering, and reasoning.
- **SQuAD (Stanford Question Answering Dataset):**
  - Evaluates models on reading comprehension.
  - Requires predicting answers based on given passages.
- **Commonsense Reasoning Benchmarks:**
  - Winograd Schema Challenge: Tests models on commonsense reasoning.
  - SWAG (Situations With Adversarial Generations): Assesses predicting likely sentence endings.

YHK

Task-Specific Benchmarks

- **Natural Language Inference (NLI) Benchmarks:**
  - MultiNLI: Predicting entailment, contradiction, or neutrality.
  - SNLI (Stanford Natural Language Inference): Similar to MultiNLI.
- **Machine Translation Benchmarks:**
  - WMT (Workshop on Machine Translation): Annual competition across language pairs.
- **Task-Oriented Dialogue Benchmarks:**
  - MultiWOZ: Evaluates dialogue systems in task-oriented conversations.
- **Code Generation and Understanding Benchmarks:**
  - MBPP Dataset: Includes around 1,000 Python programming problems.
- **Chart Understanding Benchmarks:**
  - ChartQA: Focuses on complex reasoning tasks using machine-generated questions.

Popular Benchmarks

For text based large language models

- MT-Bench
- MMLU
- ARC
- HELLASWAG
- TRUTHFULQA
- WINOGRADE
- GSM8K

YHK

MT-Bench

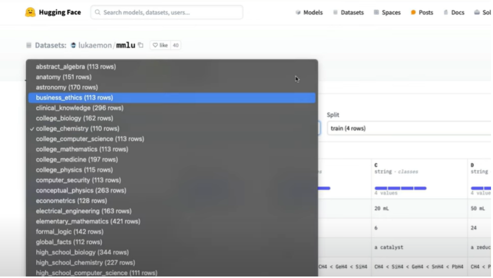
For chatbots

- LLM as judge to evaluate conversational and instruction following abilities
- 8 primary categories such as Writing, Role-play, Coding etc
- 10 multi-tern questions in each category

MMLU

Massive Multitask Language Understanding

- Evaluates how well the ILM can multitask, multi-modal capabilities
- Multi-choice, variety ot tasks, different domains such as STEM, humanities, etc
- 15k hand collected questions dataset across 57 tasks
- Averages scores per category then averages them



(Ref: Everything WRONG with LLM Benchmarks (ft. MMLU)!!! - 1littlecoder)

ARC

AI2 (Allen INstitute also) Reasoning Challenge

- Evaluates how well a model can reason.
- Easy set, Challenge set for reasoning and understanding
- Hand collected multi-choice set from standardized tests
- Difficulty level 3 to 9th grade

HELLASWAG

Harder Ending Longer-Context Low-shot Activities Situations With Adversarial Generations

- Evaluates Common sense
- Presents scenarios with multi-choice endings
- Data has actions in videos, and there is only one write answer
- Dataset of 70k sentence completions (10 shot)
- Humans are at 95%, GPT-4 at 95%, Palm 87%

TRUTHFULQA

The Winograd Schema Challenge

- Evaluates Truthfulness
- Common sense reasoning benchmark with 44k fill-in-the-banks with binary options only.
- "The doctor diagnosed Justin with bipolar and Robert with anxiety. —- had terrible nerves recently." Chose between Justin and Robert.

YHK

WINOGRADE

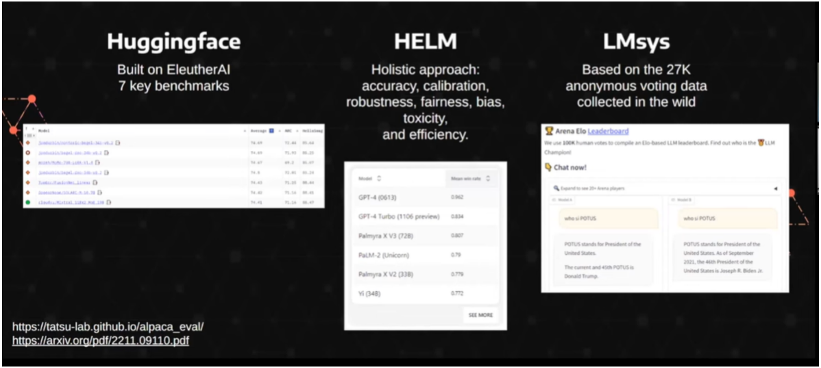
- Check against facts, say "Is Earth flat?"
- Random 800 questions, generally misleading

GSM8K

Grade School Math 8K

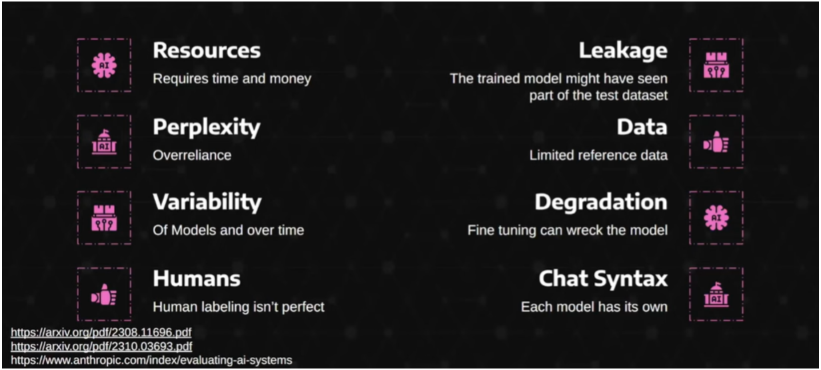
- 8.5K basic math problems, needing step-by-step reasoning (2-8 steps).
- Tests logic and mathematical abilities

Leader-board



(Ref: The Science of LLM Benchmarks: Methods, Metrics, and Meanings — LLMOps - LLMOps Space

Challenges



(Ref: The Science of LLM Benchmarks: Methods, Metrics, and Meanings — LLMOps - LLMOps Space

Conclusions

Key Characteristics

- Quantitative: Metrics should provide a numerical score for task evaluation.
- Set a minimum passing threshold for determining LLM application adequacy.
- Monitor score changes over time to iterate and improve implementation.
- Reliable: Ensure consistency in metric performance, especially with unpredictable LLM outputs.
- Beware of inconsistency in LLM-Evals like G-Eval; traditional scoring methods may be more stable.
- Accurate: Align metrics with human expectations for meaningful evaluation.
- Reliable scores are meaningless if they do not truly reflect LLM application performance.

LLM Takeaways

- Larger Model → Richer Knowledge
- Prompting → Need to model to provide explanations
- Experiment with prompting!
- Consider KNN/Few shot approach
- In Domain → Can't expect explanations outside of the training data

References

Many publicly available resources have been refereed for making this presentation. Some of the notable ones are:

- Evaluation of LLMs and RAGs - AI Anytime https://www.youtube.com/playlist?list=PLrLEqwuz-mRI5ubqVJ7DpbHheCflJDDXk
- Evaluation of LLM is All You Need — Why, What, Where and How to Evaluate - Neural Hacks with Vasanth https://www.youtube.com/watch?v=hxwa8aPmpow
- Ragas : evaluation framework https://github.com/explodinggradients/ragas

Thanks . . .

- Search **"Yogesh Haribhau Kulkarni"** on Google and follow me on LinkedIn and Medium
- Office Hours: Saturdays, 2 to 5pm (IST); Free-Open to all; email for appointment.
- Email: yogeshkulkarni at yahoo dot com

(Generated by Hugging Face QR-code-AI-art-generator, with prompt as "Follow me")