Evaluation of LLMs Is All You Need

Introduction

🔍 **Unlocking the Unseen:** Delve into the world of LLM evaluation, a facet often underestimated yet holding the power to transform your model's prowess! 💎

🔢 **Crucial Queries Explored:**

- **Why Evaluate LLMs?** Discover the critical need for evaluation in maximizing model potential. Uncover why this step is a cornerstone of success, and how it can take your model from good to extraordinary.

- **What to Evaluate?** Explore the multifaceted approach to evaluation. Understand the dimensions that demand attention – from coherence to context handling – and how they collectively shape model effectiveness.

- **Where to Evaluate From?** Learn about the datasets that fuel evaluation. Explore the diverse sources required to gauge the model's performance across various aspects addressed in the 'What' section.

- **How to Execute Evaluation?** Unveil the strategies and techniques to conduct effective evaluation. From methodologies to metrics, this section provides the roadmap to assessing and enhancing model powers

🚀 **Elevate Your Model:** Elevating your model's performance is just a click away! Through comprehensive evaluation, you can turn setbacks into triumphs, failures into stepping stones towards success. 📈 This video is your complete A-to-Z guide, providing insights that can reshape your approach to LLMs.

Evaluation

- Model evaluation is the process of analyzing the performance of the model with the help of some metrics

- Evaluating an LLM performance involves assessing factors such as language fluency, coherence, contextual understanding, factual accuracy, and ability to generate relevant and meaningful responses.

What is a good evaluation?

- **Correlated with outcomes:** Appropriate metrics used for appropriate models

- **Correlated with outcomes:** Appropriate metrics used for appropriate models
- **Very less number of metrics, in an ideal world single metric:** Easy to track and monitor and make a judgement accordingly

- **Fast and automatic as possible to compute:** We can't completely automate the evaluation. It is important to have a human intervention but yet the evaluation should be as automated and fast as possible

Why the conventional methods of evaluation doesn't work for LLMs?

- The data used while training and production are always not the same. It can be as different as possible

- Another key bottleneck is that in LLMs we wont have definitive results. It has a complex generation behavior which is hard to understand. Though the sentence generated would be different from the ground truth the generated sentence will provide the same contextual meaning.

- **For eg:**

```
In Traditional ML, lets consider a scenario of sentiment analysis

    pred = [P, N, P, P]
    label = [P, N, P, N]

For the above set to be evaluated we can use metrices like accuracy which here
will be 0.75 but that cannot be the case for LLMS

For LLMs, lets consider a case of summarization of a context given

    pred = Usually LLMs works very well with wide variety  of NLP tasks
because they are great generalists by nature
    label = LLMs are great generalists, so they usually work pretty good with
variety of NLP tasks

Both convey the same meaning if we see it in a contextual way then the model
can be given 100% but usually traditional methods are not qualitative but
quantitative.
```

Thus it is hard to have a conventional metric to quantify the evaluation.

Critical Questions of Evaluation

There are four main questions to consider in evaluation. They are:

- Why to evaluate?
- What to evaluate?

- Where to evaluate?

- How to evaluate?

Now lets answer each question in detail one by one right now

Why to evaluate?

Here are some of the reasons why evaluation is very necessary and why it is said to be one of the most underrated aspect of the LLM pipeline

Increase model performance

By evaluating the model one can understand the strengths and weaknesses of a model. Once a models weaknesses are known one can then move onto the next steps to increase the performance working on the weaknesses of the model.

For eg: PromptBench indicates a fact that the current LLMs are sensitive to adversarial prompts which implies that you can gain better performance with careful prompt engineering

Better Human-LLM interaction

With better evaluations it can provide better guidance for human-LLMs interaction which can lead to some better experience of the users

For eg: Once you know if your model is exhibiting an emotion for a specific way of interaction then a work around could be made to make the interaction better to get the desired output from the model

Safety and Reliability

LLMs have a broad applicability and are used in various sectors even in some sectors which may require safety and reliability like some financial or healthcare institutions. So it is important to ensure the safety and reliability of the model

Thus it is important to have evaluation as one of the most important discipline in the LLM building pipeline

What to evaluate

Once we find the answer to the question we can claim the strengths and weaknesses of LLMs. the answer to the question is the different tasks which are there to evaluate against the model.