Learn by Hands on Labs with AJ

AWS BEDROCK
–LLM EVALUATION–

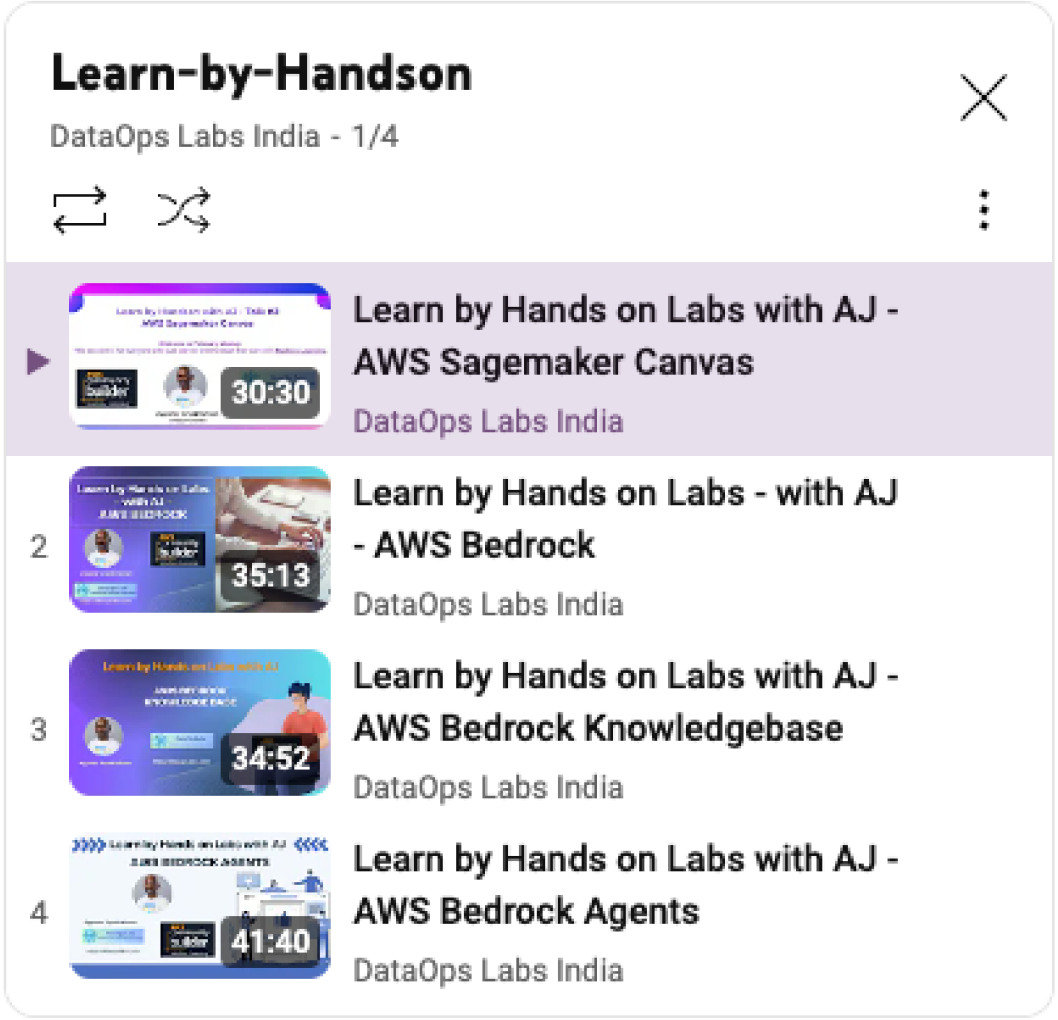
Ayyanar Jeyakrishnan

AGENDA

1. Why we need LLM Evaluation
2. Challenges in LLM Evaluation
3. Evaluation Approaches and Metrics
4. Demo - LLM Evaluation Amazon Bedrock

Recap of Bedrock Learning Series

https://www.youtube.com/@DataOpsLabsIndia



Why We need LLM Evaluation

- **Quality Assurance:** Evaluation ensures that large language models (LLMs) meet desired standards of performance and reliability, enhancing trust in their outputs.
- **Model Improvement:** Evaluation identifies weaknesses and areas for improvement in LLMs, driving iterative development and refinement, Avoid Hallucination.
- **Ethical Considerations:** Evaluation helps assess the ethical implications of LLMs, including potential biases, misinformation propagation, and harmful content generation.
- **Application Suitability:** Evaluation determines the suitability of LLMs for specific tasks and contexts, guiding their deployment in real-world scenarios eg) Cost, Latency, Token Generation.
- **User Confidence:** Evaluation results provide users with insights into the capabilities and limitations of LLMs, fostering confidence in their usage and interpretation of generated content.

Challenges in LLM Evaluation

- **Scalability:** Evaluating large language models (LLMs) at scale poses significant challenges due to the computational resources required, especially when dealing with extensive datasets and complex tasks.

- **Subjectivity and Variability:** Assessing the quality of LLM outputs involves subjective judgments, leading to variability in human annotations and evaluation metrics, which can affect the reliability of evaluation results.

- **Interpretability:** Understanding and interpreting the outputs of LLMs, particularly in complex or nuanced tasks, can be challenging, making it difficult to discern whether generated content is accurate, relevant, or appropriate.

High Level Understanding of How LLM can be Evaluated

- **Model Evaluation Using Panels:** To mitigate the challenges posed by intra-model bias and high costs associated with using a single large model as an evaluator, the proposal suggests employing a Panel of LLM Evaluators (PoLL). This panel consists of multiple evaluator models drawn from diverse model families, offering a more comprehensive and unbiased assessment of LLM outputs.

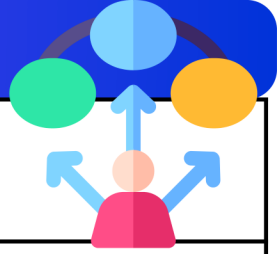
- **Reduced Intra-Model Bias:** By pooling evaluations from multiple models within the PoLL, the approach diminishes the impact of intra-model bias. Unlike relying on a single judge model, which may exhibit preferences for its own outputs, the PoLL composition ensures a broader perspective, enhancing the objectivity and fairness of evaluations.

- **Cost-Effective Evaluation:** Utilizing a PoLL of smaller models for evaluation proves to be over seven times less expensive compared to employing a single large model like GPT-4. This cost reduction makes evaluation resources more accessible and scalable, enabling researchers and developers to conduct comprehensive assessments of LLM performance without incurring prohibitive expenses.

Intrinsic vs Extrinsic Model Hallucination

Aspect	Intrinsic Model Hallucination	Extrinsic Model Hallucination
Evaluation Approach	Traditional metrics like BLEU or ROUGE are used to assess text quality.	Model outputs are compared against human-generated text or specific benchmarks.
Example	In intrinsic evaluation, a machine translation system is assessed using BLEU score, which measures how closely its output aligns with human translations without considering the context or coherence.	In extrinsic evaluation, a chatbot's responses are evaluated by comparing them to responses generated by humans or by another well-established chatbot system, considering factors like relevance, coherence, and conversational flow.

Let Us Focus on Text Summarization Usecase



Application	Description
News Summarization	Generating concise summaries of news articles for quick consumption and dissemination
Document Summarization	Condensing lengthy documents such as reports, contracts, and policies into key points
Social Media Summarization	Extracting essential information from social media feeds to identify trends and sentiments
Email Summarization	Summarizing long emails to quickly understand the main points and take necessary actions
Legal Summarization	Identifying critical information from legal documents, contracts, and court rulings
Educational Summarization	Creating summaries of educational texts and lectures to aid learning and revision

Metrics to Evaluate

Ref: https://aws.amazon.com/blogs/machine-learning/evaluate-the-text-summarization-capabilities-of-llms-for-enhanced-decision-making-on-aws/

Metric	ROUGE Recall-Oriented Understudy for Gisting Evaluation	METEOR - Metric for Evaluation of Translation with Explicit Ordering	BERTScore
Evaluation Focus	Lexical overlap between summaries	Semantic similarity and flow of content	Semantic similarity beyond surface-level
Scoring Method	Counts overlapping n-grams	Compares stemmed forms, synonyms, and paraphrases	Measures similarity using BERT embeddings
Example Calculation	ROUGE-1: Count unigram overlaps	Calculate precision, recall, and F1 score	Compare contextual embeddings for each token
Limitations	Narrow focus on lexical overlap	Dependency on pre-trained models	Computational intensity, domain specificity
Use Cases	Baseline evaluation for content selection	Assessing order of ideas and fluency	Retaining semantic meaning in summaries
When to Use	Abstractive summarization tasks	Summaries where order of ideas matters	Semantic fidelity is critical

Bedrock - Model Evaluation

Build an evaluation

Automatic

Evaluates a single model using recommended metrics. Provides results based on the parameters that you specify when you create the evaluation, such as accuracy, toxicity, and robustness. Choose from built-in task types, text summarization, question and answer, text classification, and open-ended text generation, and scores will be calculated automatically. Model scores are calculated using various statistical methods such as BERTScore, F1, and more. You can bring your own prompt dataset or use built-in curated prompt datasets.

Create automatic evaluation

Human: Bring your own work team

Evaluates up to 2 models using a work team of your choice to provide feedback. Provides results based on the parameters that you specify when you create the evaluation. You can use recommended task types and their associated metrics, or customize the task types and metrics that are important to your needs. You provide your own prompt dataset to ensure the evaluation is relevant to you. This is a good option if you want feedback on subjective or complex evaluation metrics.

Create human-based evaluation

Human: AWS Managed work team

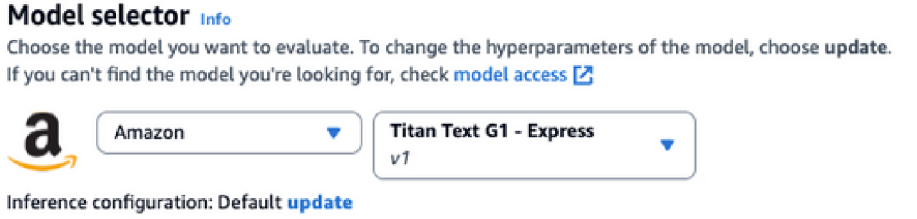
Customize the number of models to evaluate using a work-team designated by AWS. Provides results based on the parameters that you specify when you create the evaluation. You provide your own prompt dataset, define the task types and metrics that are important to your evaluation, and engage with an AWS team directly. The AWS team will ensure that your evaluation meets your needs. This is a good option if you want feedback on subjective or complex evaluation metrics, and want an expert AWS team to manage the whole evaluation workflow within your guidelines.

Create AWS managed evaluation

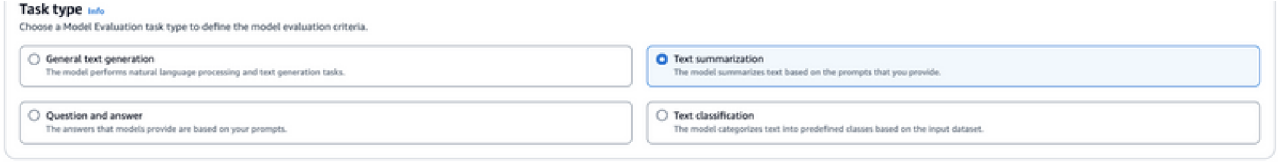
Bedrock - Automatic Model Evaluation - From Console

1. Create Assets - Provide Evaluation Details - Evaluation Name and Description

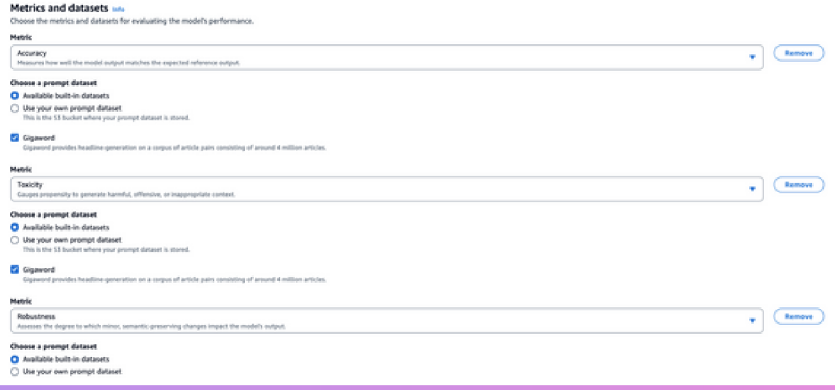
2. **Select the Model** - You want to Evaluate - Bedrock - has 5* Model available for Evaluation



3. **Select the Task Type**
   - General Text Genration
   - Text Summarization
   - Question and Answer
   - Text Classification



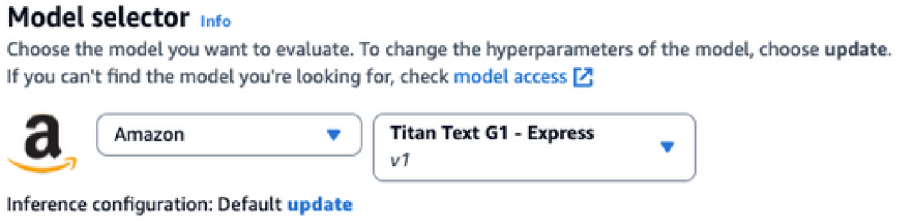
4. **Metrics and Dataset - Bring your Own Dataset or Default Gigaword Dataset**
   - Accuracy
   - Toxicity
   - Robustness



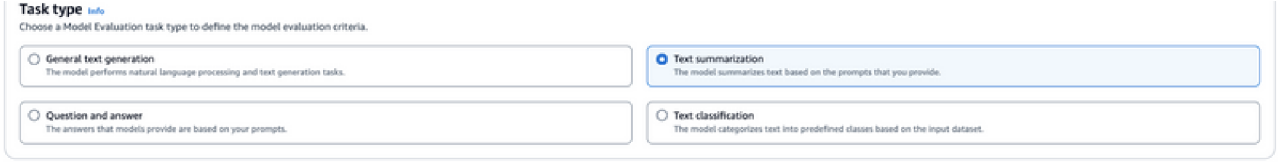
Bedrock - Automatic Model Evaluation - SDK - FMEval

1. Create Assets - Provide Evaluation Details - Evaluation Name and Description

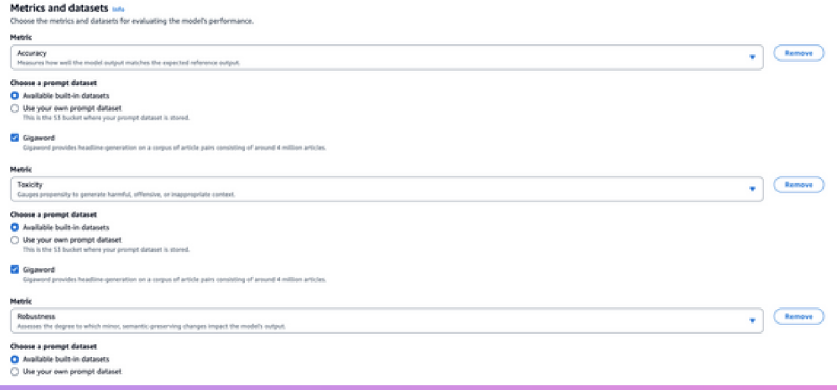
2. **Select the Model** - You want to Evaluate - Bedrock - has 5* Model available for Evaluation



3. **Select the Task Type**
   - General Text Genration
   - Text Summarization
   - Question and Answer
   - Text Classification

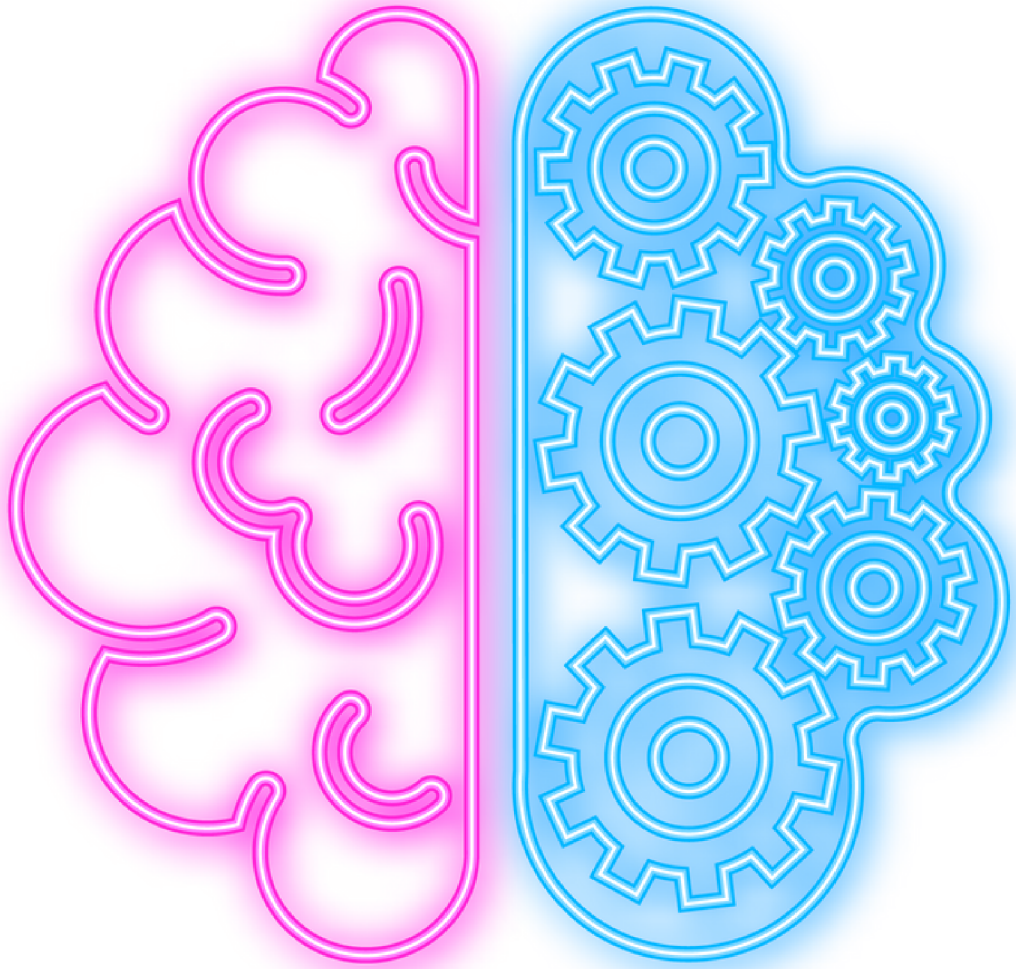


4. **Metrics and Dataset - Bring your Own Dataset or Default Gigaword Dataset**
   - Accuracy
   - Toxicity
   - Robustness



LLM EVALUATION DEMO

FMEVAL



Feedback

Thank you!

QUESTIONS WELCOME

Connect with me

https://www.dataopslabs.com/

https://www.youtube.com/@DataOpsLabsIndia