Evaluating your RAG pipelines

1.

LLM-as-Judge

Why?

Increasingly Complex Tasks

Multi-turn Dialogue, Data Agents etc and n-grams, semantic similarity don’t fit

Cheaper and Faster than Human Evals

Not scalable but the gold standard (for now)

Less data wrt fine-tuned models

Lots of possibilities to scale up testing and different situations.

How should are you scoring

- **Direct Scoring** - Don’t need a Reference. Great for measuring hallucinations, policy violation etc
- **With Ground Truth** - Answer Correctness, Recall, Precision etc
- **Pairwise Comparison** - comparison metrics, great for style, policy etc. Easier to get human annotations

How should we evaluate the evaluators?

Classification Metrics

for intent classifications, routers, sentiment analysis

Correlation Metrics

Cohen's (kappa), Kendall's (tau), and Spearman's (rho)

Reproducibility Measurements

Reproducibility of Evaluators

2.

Ragas Metrics

Metrics we are going to cover

Faithfulness	**Answer correctness**	**Context Recall**
Context Precision	**Noise Sensitivity**	**Rubric Based**

Faithfulness

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

- Reference Free Metrics
- Measure Hallucinations in the response

Answer Correctness

$$\text{F1 Score} = \frac{|\text{TP}}{(|\text{TP}| + 0.5 \times (|\text{FP}| + |\text{FN}|))}$$

- Reference Based
- End 2 End score - hence my goto

Context Recall

$$\text{context recall} = \frac{|\text{GT claims that can be attributed to context}|}{|\text{Number of claims in GT}|}$$

- Reference Based
- Measures Recall of the Retriever
- For RAG this is more important

Context Entity Recall

$$\text{context entity recall} = \frac{|CE \cap GE|}{|GE|}$$

- Reference Based
- Usefullfull in fact based usecases like tourism QA, historical QA etc
- Eg of entities
- *“The Taj Mahal is an iconic monument in India. It is a UNESCO World Heritage Site and attracts millions of visitors annually. The intricate carvings and stunning architecture make it a must-visit destination.”*
- *[‘Taj Mahal’, ‘UNESCO’, ‘India’]*

Context Precision

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}}$$

$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})}$$

- Reference Based
- Rank Sensitive
- Helps reduce noise in your retriever

Noise Sensitivity

$$\text{noise sensitivity (relevant)} = \frac{|\text{Number of incorrect claims in answer}|}{|\text{Number of claims in the Answer}|}$$

- Metric with Reference
- How Sensitive is you're LLM to Noise, how susceptible is your LLM to extra context
- Use in conjunction with recall and precision to make trade-offs

Rubric Based

```
DEFAULT_WITH_REFERENCE_RUBRICS = {
    "score1_description": "The response is incorrect, irrelevant, or does not align with the ground truth.",
    "score2_description": "The response partially matches the ground truth but includes significant errors, omissions, or irrelevant information.",
    "score3_description": "The response generally aligns with the ground truth but may lack detail, clarity, or have minor inaccuracies.",
    "score4_description": "The response is mostly accurate and aligns well with the ground truth, with only minor issues or missing details.",
    "score5_description": "The response is fully accurate, aligns completely with the ground truth, and is clear and detailed.",
}
```

- Reference or without Reference
- Rubrics can be *per dataset* or *per row*
- Inspired from **Prometheus: Inducing Fine-grained Evaluation Capability in Language Models**

[Input Prompt]

Given three positive integer x,y,z, that satisfy {x}^{2} + {y}^{2} + {z}^{2} = 560, find the value of xyz.
You are not allowed to use your code functionality.

Coarse-grained Evaluation Criteria

Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses.

Domain-Specific Evaluation Criteria

Score 1 The logic of the model's response is completely incoherent.
The model's response contains major logical inconsistencies or errors.
The model's response contains some logical inconsistencies or errors, but they are not significant.
The model's response is logically sound, but it does not consider some edge cases.
Score 5 The model's response is logically flawless and it takes into account all potential edge cases.

Instance-Specific Evaluation Criteria

Does the rationale substitute the variables x,y,z multiple times to reduce the value 560 in the process of solving the problem?

Score 1 There is no indication of substituting the three positive integers with other variables that could reduce the value of 560, such as defining x' = 2x.

Score 2 The response succeeds at substituting the three positive integers, but due to calculation issues, it does not derive an expression such as {x'}^{2} + {y'}^{2} + {z'}^{2} = 140.

Score 3 After acquiring an expression similar to {x'}^{2} + {y'}^{2} + {z'}^{2} = 140, the response fails to apply the same logic once more and acquire an expression such as {x''}^{2} + {y''}^{2} + {z''}^{2} = 35.

Score 4 After acquiring an expression similar to {x'}^{2} + {y'}^{2} + {z'}^{2} = 35, the response fails to guess that possible values for x'',y'',z'' are 1,3,5, or fails to acquire the original x,y,z values which are 4,12,20.

Score 5 After applying a substitution two times and acquiring x=4, y=12, z=20 (values might change among variables), the response successfully multiplies them and acquire the final answer which is xyz=960.

3.

Tips and Tricks

Using Metrics as a Guiding Light

- Don’t over optimise for eval scores (0.78 → 0.79 could be noise)
- Instead use them to identify distributions that are not working
- Comparing how they change with different experiments

How to Choose the Judge LLM

- For Ragas, the model should have good support for structured output
- Consideration on metrics. Reasoning vs Style
- Cost and latency
- Case Study - How we chose the LLM for course leaderboard

The Alignment Problem

- Remember the entity issue?
- Align the judge to what you and your organisation want
- Tuning the instructions and few-short examples
- Can use DSPy techniques
- Active area of research
  - Who Validates the Validators: Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences
  - We Need Structured Output: Towards User-centered Constraints on Large Language Model Output


THANK YOU, SIR.
peacock