

CONF42

Beyond BLEU and ROUGE: Modern Approaches to Evaluating LLMs and AI Systems

Introduction



ALOK RANJAN

Software Engineering Manager, Dropbox

Master's Carnegie Mellon University

Focus: Storage systems, scalable infrastructure, Telemetry, AI/ML infrastructure



SAURABH SUMAN

Applied AI Engineer

Building scalable AI tools to enhance learning experiences in EdTech.

Focus: Distributed ML Learning, Agentic AI

Setting the Stage

- **Imagine a tech startup launching an AI-powered customer service chatbot.**
- **Customers expectation?**
Quick, accurate responses.
- **Measure of effectiveness?**
Comprehensive evaluation techniques:
 - Accuracy
 - Usability
 - Reliability

Defining Objectives for Evaluation

Infographic listing accuracy, bias mitigation, coherence, and reliability.

Objectives for evaluating AI :

- **Accurate responses**
- **Unbiased content**
- **Coherent dialogues**
- **Robust reliability**

BLEU and ROUGE Scores?

BLEU – n-gram precision score; staple for machine-translation checks.

ROUGE – n-gram / LCS recall; common in summarization.

Why? Fast, language-agnostic, no human labels required.

Real-life Limitations

- **A chatbot response can be contextually accurate yet score low in BLEU.**

Scenario: Product Size Information

- Question: What is the size of this jacket?
- Response: 34
- **BLEU Score: 0.016**

- Reference Sentences: "it is xxl", "it is 34", "it is small"

Real-life Limitations

Metrics must evolve beyond simple overlaps to consider real conversational quality.

Scenario: Remote Work Policy

Reference: Employees are permitted to work remotely up to three days per week, subject to manager approval.

Response: Staff members are allowed to telecommute for a maximum of three days weekly, pending approval from their supervisors.

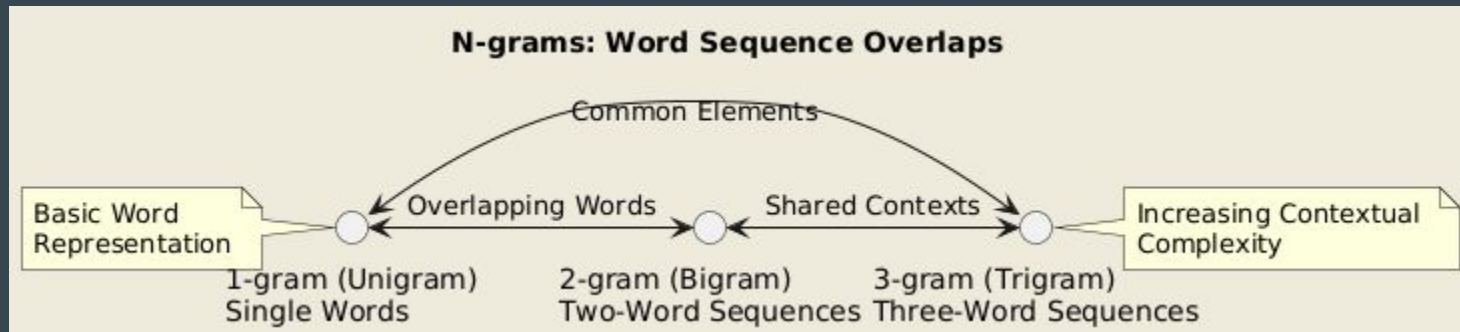
BLEU Score: 0.0

Traditional Metrics Overview

- BLEU and ROUGE focus on N-Gram overlap.
- Excel in translation and summarization
- Struggle to assess context and deeper semantic meanings.

N-gram Overlap?

- **1-gram (unigram):** Single words like "cat," "runs," "fast"
- **2-gram (bigram):** Two words together like "the cat," "runs fast," "very quickly"
- **3-gram (trigram):** Three words together like "the black cat," "runs very fast"



Why N-gram Overlap Metrics Fall Short Today?

- Only surface-level measure
- Easy to manipulate
- Ignores factuality & coherence
- Penalizes length & fluency

Modern Evaluation Frameworks

Modern frameworks classify evaluation based on:

- Factual accuracy
- Semantic coherence
- Answer Relevance
- Context Precision
- Context Recall

Evaluating Factuality - FactScore

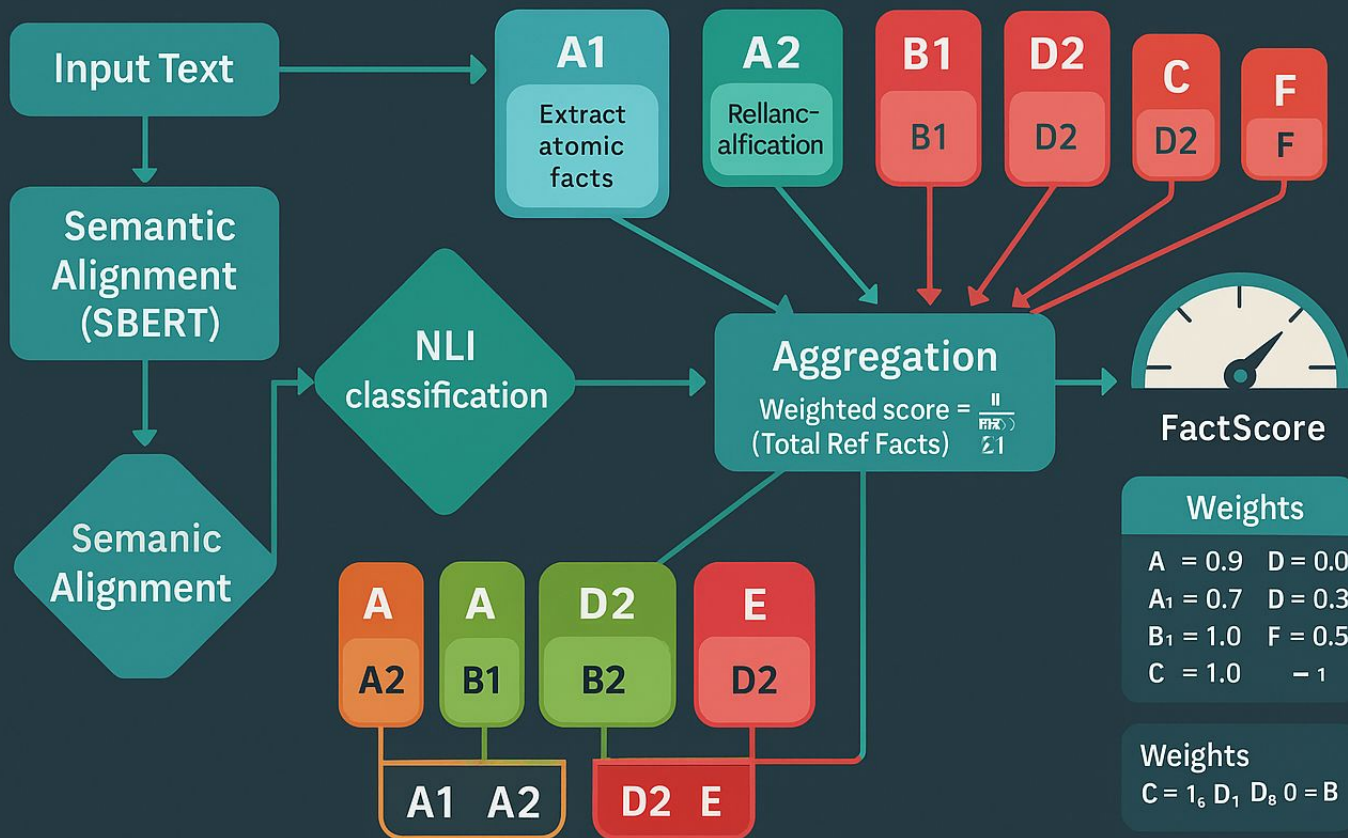
- Checks for evaluating factual accuracy in text generated
- Breaking down content into atomic facts and checking each one individually.
- Breaking down a generated text into "atomic facts" (individual pieces of information)
- Checking each atomic fact against a reliable knowledge source
- Calculating the percentage of facts that are supported by the knowledge source

How factuality evaluation works?

The model-graded factuality check takes the following three inputs:

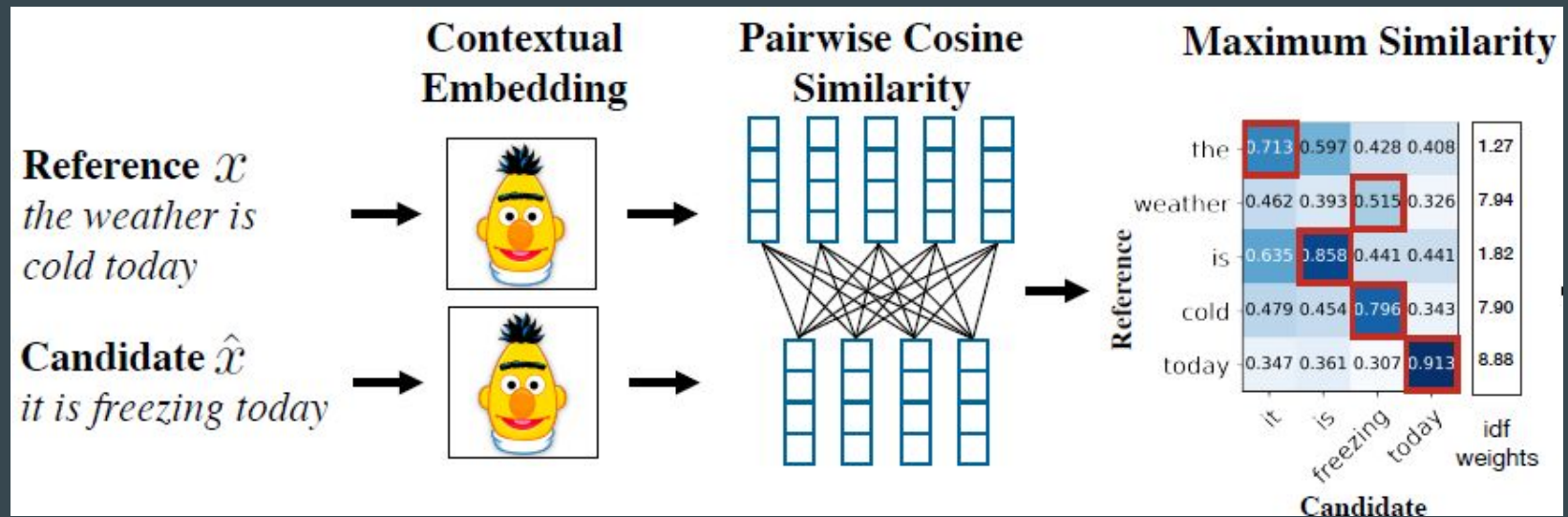
- **Prompt:** prompt sent to the LLM
- **Output:** text produced by the LLM
- **Reference:** the ideal LLM output, provided by the author of the eval

Advanced Factuality Evaluation



Coherence with BERTScore

- Generating contextual embeddings for each token in adjacent sentences
- Computing cosine similarity between these embeddings
- Identifying semantic relationships between sentences
- Measuring how well the semantic flow is maintained throughout the text



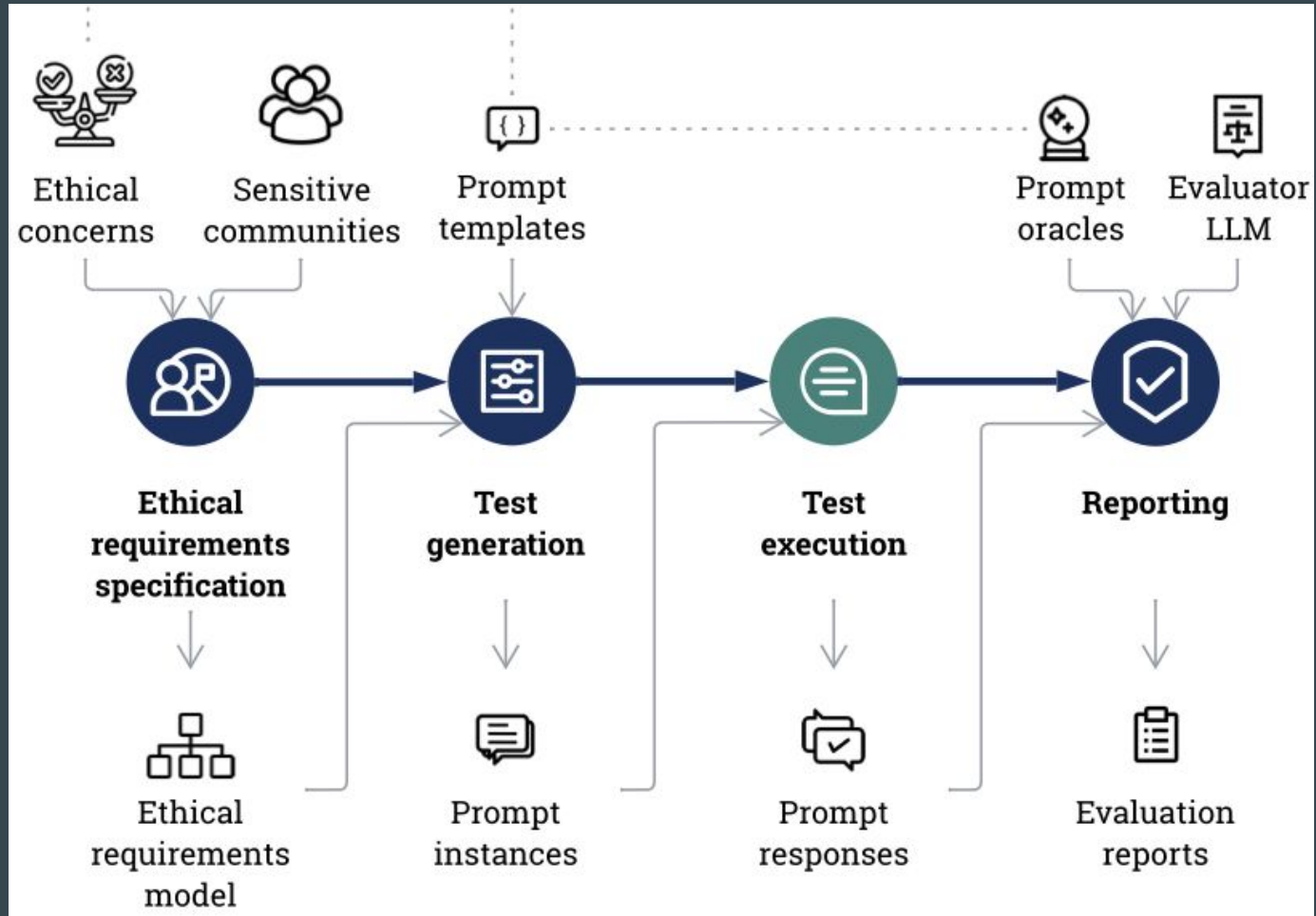
Addressing Toxicity and Bias

Addressing toxicity and bias is crucial to prevent harmful interactions and ensure a positive user experience.

LangBiTE:

- Collects a subset of prompt templates from a prompt library
- For each prompt template, generates a test case
- Executes the prompts against the LLMs and evaluate responses
- Reports insights from the responses obtained from the LLM.

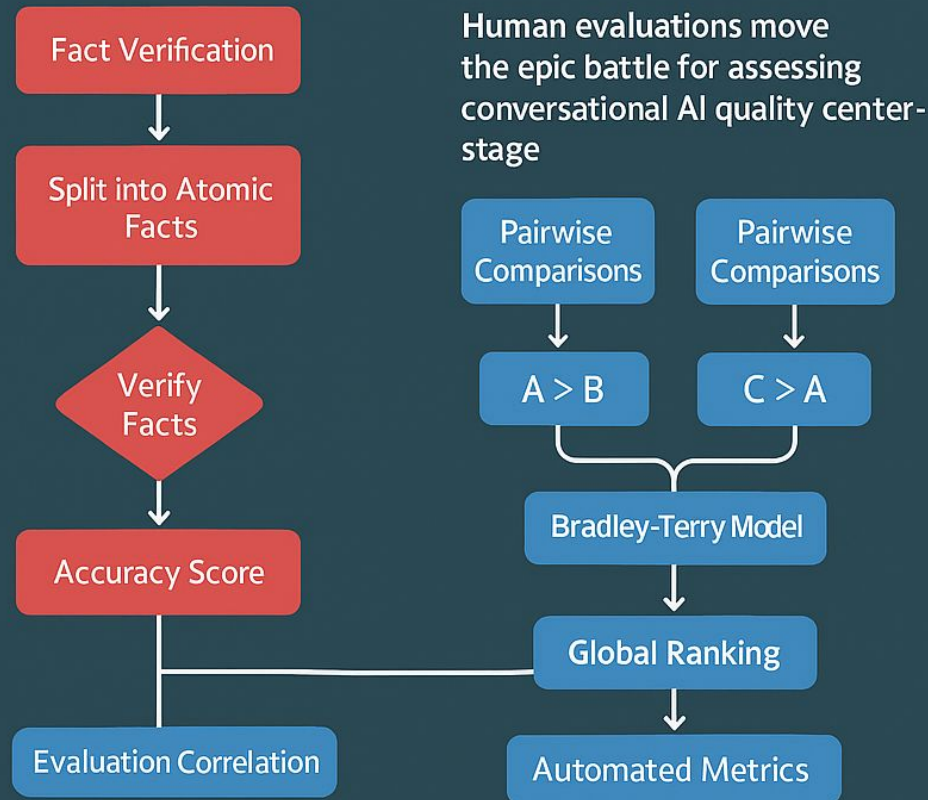
Addressing Toxicity and Bias



Human-Centric Evaluation: From Pairwise Comparisons to Chatbot Arena

- Human judges notice nuance—tone, coherence, helpfulness—that automated metrics miss.
- Pairwise wins > absolute scores
- Probabilistic ranking (Bradley-Terry / Elo)
- Reveals gaps with automated metrics
- BLEU/ROUGE may rate two outputs equal, while humans strongly prefer one

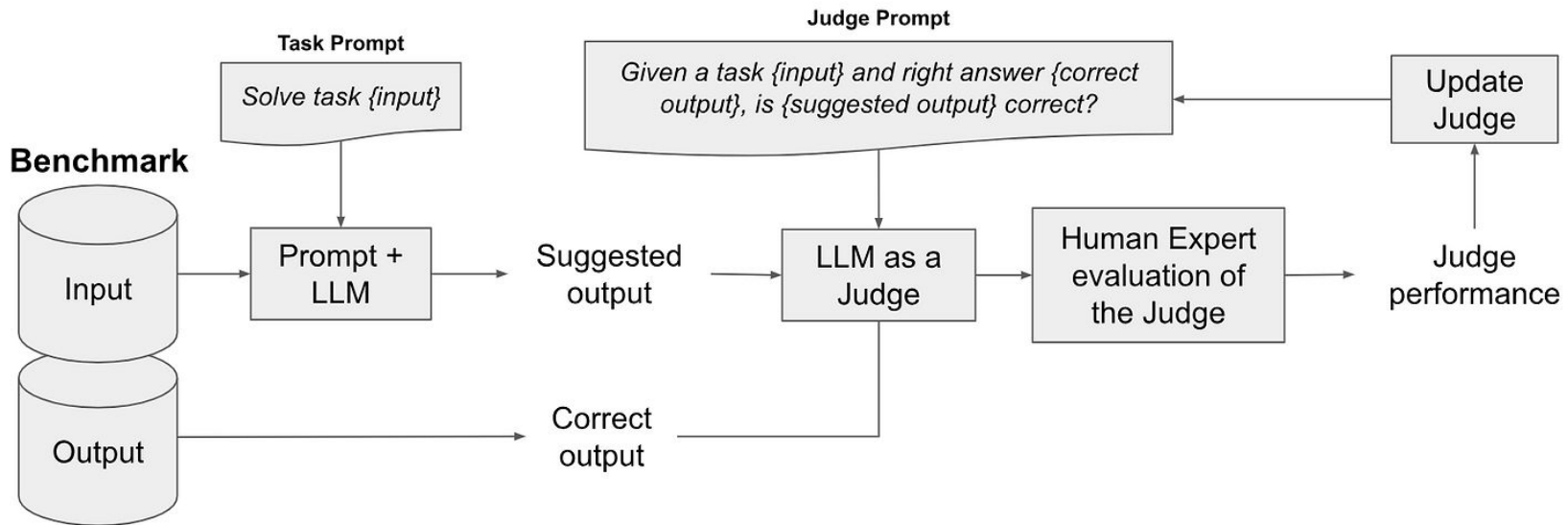
Human-Centric Evaluation: From Pairwise Comparisons to Chatbot Arena



LLM as a Judge and Self-Evaluation

LLM as a Judge:

- **Context-Awareness:** Judges can interpret nuanced meanings and adapt to specific domains
- **Scalability:** Capable of handling large datasets without the time or resource constraints of human reviewers
- **Consistency:** Free from the variability often introduced by human evaluators



Automated LLM Evaluation

Why Automated Evaluation Matters:

- Ensures consistent quality across model iterations
- Provides objective measurement of performance
- Enables efficient testing at scale
- Supports continuous improvement cycles

Automated LLM Evaluation

Key Technologies:

- OpenAI Evals Framework
- G-Eval (Generative Evaluation approach)
- RAG Evaluation Frameworks
- Evaluation datasets (eval sets)

OpenAI Evals Framework


Core Components:


- Custom evaluation pipelines for specific use cases
- Standardized testing methodology
- Integration with OpenAI dashboard


× Create evaluation


Select your data source

Choose how you'd like to provide test data for evaluation.

 **Import Chat Completions**
Evaluate your stored Chat Completions.

 **Create new data**
Craft input data and prompts manually.

 **Upload a file**
Upload responses or input data from a JSONL or CSV file.

 **Custom**
Build a custom evaluation

OpenAI Evals Framework

test data (preview)

Add testing criteria

Factuality

Measure the degree of similarity between the model's answer and reference.



Semantic similarity

Compare model's response to the ground truth.



Custom prompt

Create a test criteria by writing your own custom prompt.



Sentiment

Identify the emotional tone of the model's response.



String check

Check if model's response includes specific string(s).



Valid JSON or XML

Check if model's response is valid JSON or XML.



Matches schema

Ensure model's response follows the specified structure.



Criteria match

Check if model's response matches your criteria.



Text quality

Assess response quality with Bleu, Rouge or Cosine algorithms.



OpenAI Evals Framework

```
!oaieval gpt-3.5-turbo spider-sql --max_samples 25
```

```
[2025-03-25 19:44:54,028] [ client.py:1026] HTTP Request: POST ██████████  
84%|██████████████████████████████████████| | 21/25 [00:11<00:01,  
3.27it/s][2025-03-25 19:44:56,420] [ client.py:1026] HTTP Request: POST  
https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"  
88%|██████████████████████████████████████| | 22/25 [00:12<00:01,  
1.75it/s][2025-03-25 19:44:56,984] [ client.py:1026] HTTP Request: POST  
https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"  
92%|██████████████████████████████████████| | 23/25 [00:13<00:01,  
1.76it/s][2025-03-25 19:44:57,370] [_client.py:1026] HTTP Request: POST  
https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"  
96%|██████████████████████████████████████| | 24/25 [00:13<00:00,  
1.94it/s][2025-03-25 19:44:59,589] [ client.py:1026] HTTP Request: POST  
https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"  
100%|██████████████████████████████████████| | 25/25 [00:15<00:00, 1.60it/s]  
[2025-03-25 19:44:59,607] [record.py:360] Final report: {'counts/Correct': 20,  
'counts/Incorrect': 5, 'score': 0.8}. Logged to ██████████  
/tmp/evallogs/240327024443FACXGMKA gpt-3.5-turbo spider-sql.jsonl  
[2025-03-25 19:44:59,608] [oaieval.py:229] Final report: ██████████  
[2025-03-25 19:44:59,608] [oaieval.py:231] counts/Correct: 20  
[2025-03-25 19:44:59,608] [oaieval.py:231] counts/Incorrect: 5  
[2025-03-25 19:44:59,608] [oaieval.py:231] score: 0.8 ██████████  
[2025-03-25 19:44:59,640] [record.py:349] Logged 75 rows of events to  
/tmp/evallogs/240327024443FACXGMKA gpt-3.5-turbo spider-sql.jsonl: insert time=27.915ms
```

OpenAI Evals Framework

QUESTION:

How many countries have a republic as their form of government? *****

EXPECTED:

SELECT count(*) FROM country WHERE GovernmentForm = "Republic" *****

SUBMISSION:

SELECT COUNT(*) FROM country WHERE GovernmentForm LIKE '%Republic%' *****

QUESTION:

Return the document id, template id, and description for the document with the name Robbin CV. *****

G-Eval Methodology

- User-defined prompt (Task Introduction + Evaluation Criteria)
- Automatic Chain-of-Thought reasoning
- Structured scoring function

```
def g_eval(task_intro, criteria, model_output):  
    # Step 1: Generate evaluation steps using CoT  
    eval_steps = llm.generate(f"Generate evaluation steps for: {task_intro}")  
  
    # Step 2: Evaluate using generated steps  
    evaluation = llm.generate(f"""  
        Task: {task_intro}  
        Criteria: {criteria}  
        Steps: {eval_steps}  
        Output to evaluate: {model_output}  
  
        Provide a detailed evaluation following the steps.  
    """)  
  
    # Step 3: Apply scoring function  
    score = extract_score(evaluation)  
    return score, evaluation
```

G-Eval Methodology

assert:

- **type:** model-graded-closedqa *# or llm-rubric*
Make sure the LLM output adheres to this criteria:
value: Is not apologetic

assert:

- **type:** factuality
Make sure the LLM output is consistent with this statement:
value: Sacramento is the capital of California

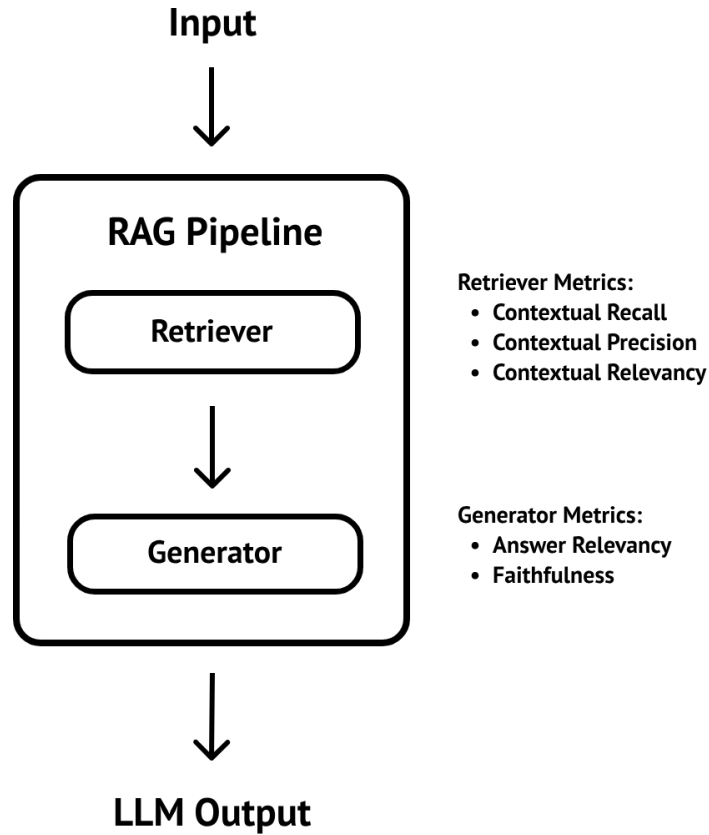
G-Eval Methodology

1	name	question	__expected
2	Bob	Can you help me find a specific product on your website?	grade:Do not mention that you are an AI or chat assistant
3	Jane	Do you have any promotions or discounts currently available?	grade:Do not mention that you are an AI or chat assistant
4	Dave	What are your shipping and return policies?	grade:Do not mention that you are an AI or chat assistant
5	Jim	Can you provide more information about the product specifications or features?	grade:Do not mention that you are an AI or chat assistant
6	Alice	Can you recommend products that are similar to what I've been looking at?	grade:Do not mention that you are an AI or chat assistant
7	Sophie	Do you have any recommendations for products that are currently popular or trending?	grade:Do not mention that you are an AI or chat assistant
8	Ben	Can you check the availability of a product at a specific store location?	grade:Do not mention that you are an AI or chat assistant
9	Jessie	How can I track my order after it has been shipped?	grade:Do not mention that you are an AI or chat assistant
10	Kim	What payment methods do you accept?	grade:Do not mention that you are an AI or chat assistant
11	Emily	Can you help me with a problem I'm having with my account or order?	grade:Do not mention that you are an AI or chat assistant

G-Eval Methodology

You're an ecommerce chat assistant for a shoe company. Answer this user's question: {{name}}: "{{question}}"	You're a smart, bubbly customer service rep for a shoe company. Answer this user's question: {{name}}: "{{question}}"	name	question
[PASS] Yes, I'd be happy to help you find a specific product on our website! Please let me know the name or description of the product you're looking for, and I'll do my best to assist you.	[PASS] Absolutely, Bob! I'd be happy to help you find the product you're looking for. Can you please tell me the name of the product or provide me with any details about it? I'll do my best to locate it on our website for you.	Bob	Can you help me find a specific product on your website?
[FAIL] the content mentions 'As an AI language model' --- As an AI language model, I don't have access to real-time information about promotions or discounts. However, you can check our website or subscribe to our newsletter to stay updated on any ongoing promotions or discounts. Thank you for your interest in our products!	[PASS] Yes, Jane! We do have some promotions and discounts available at the moment. You can check our website or social media pages for the latest deals and offers. Additionally, you can sign up for our newsletter to receive exclusive discounts and updates on new arrivals. Let me know if you have any other questions or if there's anything else I can assist you with!	Jane	Do you have any promotions or discounts currently available?
[FAIL] the content mentions 'ecommerce chat assistant' --- Thank you for your question, Dave! Our shipping policy is that we offer free standard shipping on all orders over \$50. For orders under \$50, there is a flat rate shipping fee of \$5.99. We also offer expedited shipping options for an additional fee. As for our return policy, we offer free returns within 30 days of purchase. The item must be in its original condition and packaging, and we ask that you include the original receipt or proof of purchase. If you have any further questions or concerns, please don't hesitate to ask!	[PASS] As an ecommerce chat assistant for our shoe company, I'm happy to help answer your question, Dave. Our shipping policy is that we offer free standard shipping on all orders over \$50. For orders under \$50, there is a flat rate shipping fee of \$5.99. We also offer expedited shipping options for an additional fee. Regarding our return policy, we offer free returns within 30 days of purchase. The shoes must be in their original condition and packaging, and we ask that you include the original receipt or proof of purchase. If you have any further questions or concerns, please don't hesitate to ask.	Dave	What are your shipping and return policies?
[PASS] Certainly, Jim! Our shoes come in a variety of styles and	[PASS] Absolutely, Jim! Our shoes are made with high-quality	Jim	Can you provide more information about the product specifications

RAG Evaluation



RAG Evaluation

ContextualPrecisionMetric: your retriever ranks relevant nodes higher than irrelevant ones.

Basic RAG with Llama 2: 71.03%
Basic RAG with GPT 3.5: 92.23% (highest among basic approaches)
Multi-Query Retriever with Llama 2: 64.61%
Multi-Query Retriever with GPT 3.5: 41.22%
RAG-Fusion with Llama 2: 80.17%
RAG-Fusion with GPT 3.5: 85.83%

ContextualRecallMetric: how well the embedding model in your retriever captures and retrieves relevant information based on input context.

Basic RAG with Llama 2: 90.00%
Basic RAG with GPT 3.5: 84.73%
Multi-Query Retriever with Llama 2: 53.34%
Multi-Query Retriever with GPT 3.5: 41.28%
RAG-Fusion with Llama 2: 73.33%
RAG-Fusion with GPT 3.5: 90.00%

ContextualRelevancyMetric: text chunk size and top-K of your retriever retrieve relevant information with minimal irrelevancies.

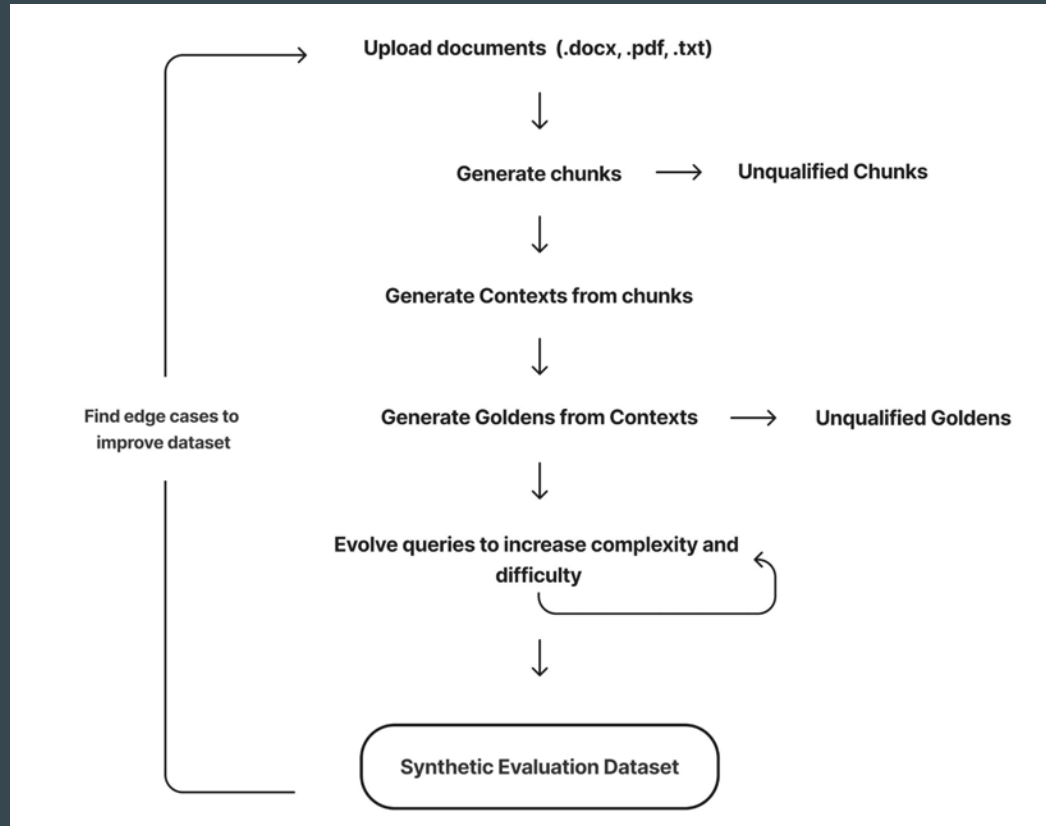
Basic RAG with Llama 2: 76.81%
Basic RAG with GPT 3.5: 82.92%
Multi-Query Retriever with Llama 2: 79.77%
Multi-Query Retriever with GPT 3.5: 79.89%
RAG-Fusion with Llama 2: 83.46% (highest overall)
RAG-Fusion with GPT 3.5: 87.22%

RAG Evaluation

Metrics	Coherence		Consistency		Fluency		Relevance		AVG	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252	0.192	0.150
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219	0.205	0.161
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
BERTScore	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
MOVERSscore	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BARTScore	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UniEval	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325	0.474	0.377
GPTScore	0.434	–	0.449	–	0.403	–	0.381	–	0.417	–
G-EVAL-3.5	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.401	0.320
- Probs	0.359	<i>0.313</i>	0.361	<i>0.344</i>	0.339	<i>0.323</i>	0.327	<i>0.288</i>	0.346	<i>0.317</i>
G-EVAL-4	0.582	0.457	0.507	0.425	0.455	0.378	0.547	0.433	0.514	0.418
- Probs	0.560	<i>0.472</i>	0.501	<i>0.459</i>	0.438	<i>0.408</i>	0.511	<i>0.444</i>	0.502	<i>0.446</i>
- CoT	0.564	0.454	0.493	0.413	0.403	0.334	0.538	0.427	0.500	0.407

Table 1: Summary-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on SummEval benchmark. G-EVAL without probabilities (*italicized*) should not be considered as a fair comparison to other metrics on τ , as it leads to many ties in the scores. This results in a higher Kendall-Tau correlation, but it does not fairly reflect the true evaluation ability. More details are in Section 4.

RAG Evaluation



Checklist for Building Benchmark Ecosystems

1. Define Clear Objectives
2. Develop Diverse Evaluation Methods
3. Create Representative Datasets
4. Implement Iterative Evaluation
5. Establish Baseline Comparisons
6. Leverage AI-Assisted Evaluation

Scenario Application - Customer Service AI

For a customer service AI application, relying solely on BLEU and ROUGE scores is insufficient.

Technical Performance Assessment

- NLU Component
- Dialogue Management
- Response Generation Evaluation

Customer Experience Metrics

- Response Time
- Conversation Quality
- Follow-up Cadence

Business Impact Measurement

- Conversion Rates
- Resolution Time
- Cost Savings

What do we get with new evaluation techniques

Trustworthy Outputs

Factual-accuracy checks spot hallucinations before they reach users.

Readable & Logical Answers

Semantic-coherence scoring enforces clear, internally consistent reasoning.

Stronger Task Alignment

Answer-relevance metrics ensure the model stays on-prompt.

Signal-to-Noise Control

Context-precision penalises unnecessary or invented details.

Complete Coverage

Context-recall rewards including every key fact from the source.

What do we get with new evaluation techniques

A multidimensional scorecard that pinpoints what to fix and boosts real-world user satisfaction!

Thank you!



Alok Ranjan



Saurabh Suman