Evaluation of LLMs and RAG Pipelines in a Smart Way

by Anurag Pathak

Why Evaluation Matters

1 **Manual Checks Insufficient**

Unscalable and prone to human error

2 **Need Scalable, Reliable Evaluation**

Ensure consistent performance

3 **Impact on Production**

Affects model alignment and user trust

Qualitative vs Quantitative Metrics

Qualitative Metrics

Human-centric assessment

- Interpretability
- Coherence
- Bias detection

Quantitative Metrics

Numeric, formula-based

- Accuracy
- F1 Score
- Perplexity

Confusion Matrix

True Positive (TP)

Correctly identified positive

False Positive (FP)

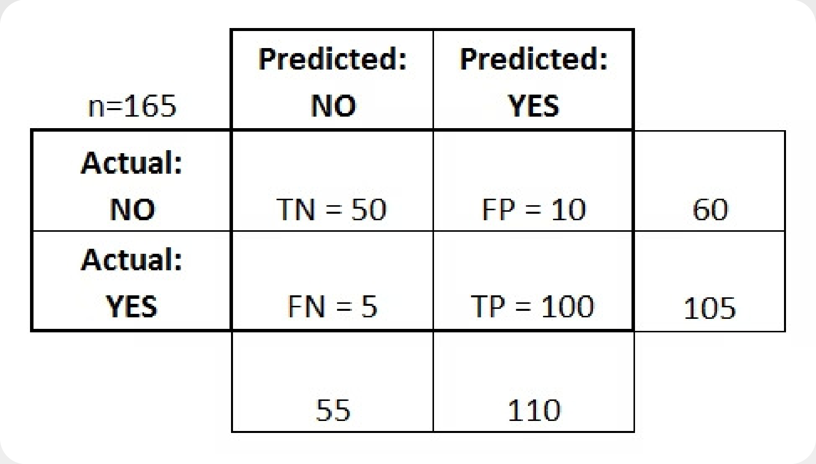
Incorrectly identified positive

False Negative (FN)

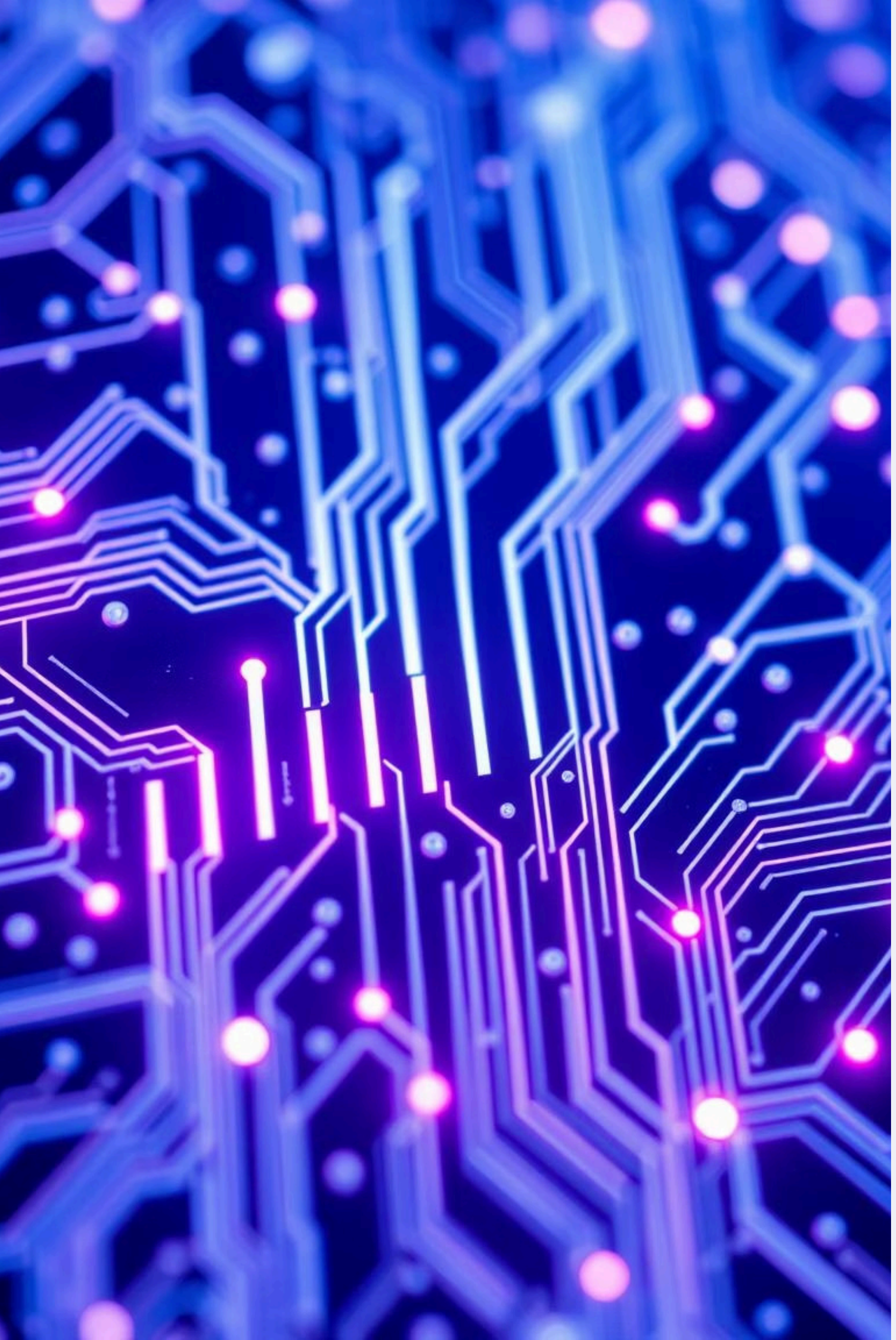
Incorrectly identified negative

True Negative (TN)

Correctly identified negative



n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	



Automated Evaluation Metrics

Text Overlap

BLEU, ROUGE, METEOR

Classification

Accuracy, Precision, Recall, F1

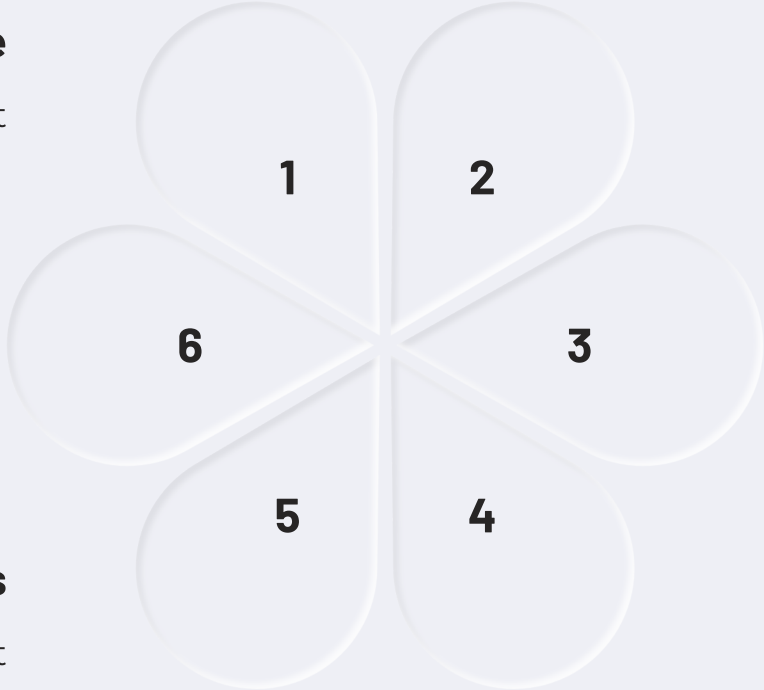
Generative Models

Perplexity, Exact Match

Ethics/Alignment

Toxicity, Bias detection

Human Evaluation Metrics



Coherence

Logical flow of text

Factual Accuracy

Correctness of information

Relevance

Pertinence to query

Helpfulness

Utility to user

Harmlessness

Absence of harmful content

Honesty

Truthfulness of claims

Testing LLM Inference

Prompt Engineering

Crafting effective inputs

Zero/Few-shot

Evaluating adaptation

Benchmarks

HELM, MMLU, TruthfulQA

Performance

Latency and throughput

RAG Pipeline Evaluation

Retriever Metrics

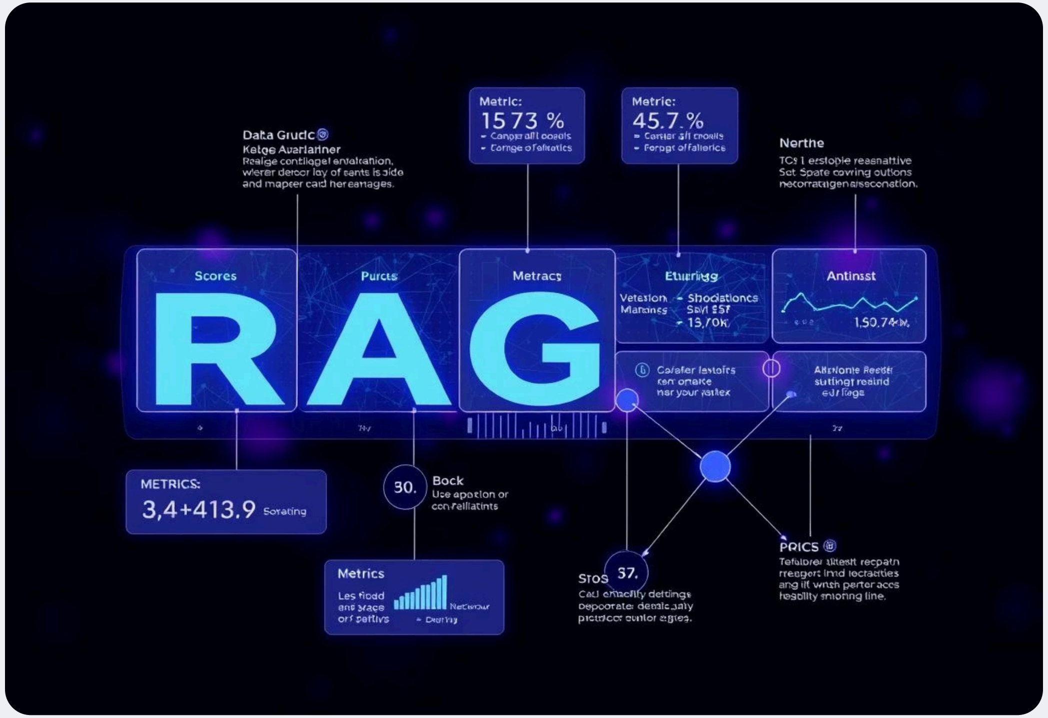
Recall@k, Context Relevancy, Context Precision

Generator Metrics

Factual grounding, hallucination rate, bias, answer correctness

End-to-End Testing

Crucial for overall quality



Evaluation Frameworks & Tooling

1

Giskard Introduction

Leading evaluation framework

Bias & Robustness

Automated checks

Human Feedback

Seamless integration

CI/CD Integration

For LLM pipelines

Best Practices & Pitfalls

Hybrid Evaluation

Combine automated and human insights

Beyond BLEU/ROUGE

Use diverse metrics for depth

Update Datasets

Keep evaluation data fresh

Align with Goals

Match eval to user needs