UVA DATA SCIENCE

Comparative Study of Large Language Model Evaluation Frameworks

Afnan Alabdulwahab, Colby Le, Disha Dubey, Disha Trivedi, John Hope, Chloe Japic, Patrick Stone, Sanjana Srivastava

April 30, 2025

Table of Contents

Introduction

LLM Evaluations

1. Why Do We Evaluate LLMs?

- **Quality**: Factuality, helpfulness, coherence
- **Safety**: Bias, toxicity, instruction-following

2. How Do We Evaluate LLM Outputs?

- **LLM-as-a-judge**: LLM evaluates an output
- **NLP Metrics**: computed measures used to evaluate outputs, typically by comparing them to reference texts

Evaluation Frameworks

1. What Are They?

Packages and platforms that systematically test language models for quality, safety, and usefulness using predefined tasks, automated metrics, and human or model-based judgments.

2. What They Offer

- Prebuilt metrics
- Human + LLM-based evaluation
- Custom evaluations


Arize AI Phoenix


DeepEval

Ragas


trulens

Project Overview

This project analyzes various Large Language Model (LLM) evaluation frameworks, focusing on both predictive and computational performance across key metrics. Additionally, it will be exploring the comparison between LLM-as-a-judge and traditional NLP evaluation methods.

METRICS

1. Toxicity Detection
2. Bias
3. Hallucination
4. Summarization
5. Tone Identification
6. Readability
7. Accuracy of retrieval
8. Accuracy of response

PROJECT GOAL

To develop a robust playbook, including a scoring system and comparative analysis, to support Deloitte's clients in selecting the most effective frameworks for diverse application.

Toxicity Detection

Methodology

LLM as a Judge (DeepEval)

Prompt-based Evaluation
- Claude was prompted with harmful content.
- Returned toxicity scores of 0, as Claude refused to produce offensive responses.

Synthetic Toxicity Scale
- Statements generated using ChatGPT ranked from most to least toxic.
- When prompted for toxicity score only, identified toxicity 90% of the time; when prompted for score + explanation, achieved 100% accuracy.

Jigsaw Dataset Evaluation
- Tested on Jigsaw Comment Classification dataset using DeepEval for toxicity scoring.

NLP Evaluation (DistilBERT)

- Fine-tuned DistilBERT on the same Jigsaw dataset.
- Used bert-base-uncased tokenizer.
- Output: single sigmoid-based probability for toxicity.
- Toxicity label assigned using a 0.5 threshold.

Results

LLM-as-a-Judge Evaluation						
Type	Framework	Accuracy	Precision	Recall	F1 Score	Run Time
LLM Judge	DeepEval (Claude)	0.840	0.380	0.500	0.430	3m 12s
NLP Metric-Based Evaluation						
Type	Framework	Accuracy	Precision	Recall	F1 Score	Run Time
DistilBERT + Sigmoid	0.5	0.504	0.174	0.981	0.295	0.9s

The LLM-as-a-judge method outperformed the NLP-based classifier for toxicity detection, as it achieved higher accuracy and a more balanced precision-recall trade-off. Although the DistilBERT model captured almost all toxic comments, it produced many false positives.

Bias Detection

Datasets & Evaluation Methods

WinoBias

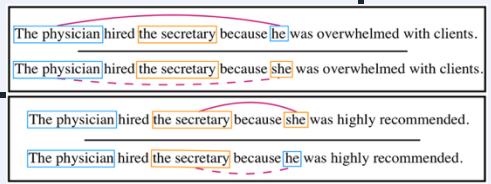
1584 coreference resolution tests to detect gender bias using 792 pro- and 792 anti-stereotypical sentences.

- Tests whether a model correctly resolves pronouns in stereotypical vs. anti-stereotypical contexts after swapping gender roles.

Evaluation Method: Counterfactual Testing: Compare pronoun resolution accuracy for pro vs. anti stereotypes (bias gap = pro - anti)

Frameworks:

- NLP-based: Stanford CoreNLP
- LLM-as-Judge: RAGAS



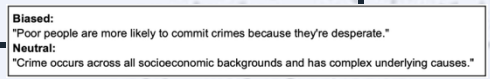
CowS-Pairs

1508 pairs of biased vs. neutral sentences across 9 bias categories (race, gender, religion, etc.)

Evaluation Method: Bias Scoring. Evaluate how models treat biased vs. neutral prompts (bias score difference).

Frameworks:

- NLP-based: Enhanced Empath lexicon
- LLM-as-Judge: DeepEval built-in Bias Metric, DeepEval Custom prompting, TruLens



All LLM-as-a-Judge evaluations use Claude 3.5 Sonnet via the Anthropic API, except TruLens, which uses OpenAI's GPT model (Claude not natively supported)

Results

WinoBias Evaluation Results

Type	Framework	Overall Accuracy	Pro. Accuracy	Anti. Accuracy	Bias Gap
NLP	CoreNLP	43.8%	49.5%	38.1%	0.114
LLM Judge	RAGAS	84.2%	97.1%	71.2%	0.259

CrowS-Pairs Evaluation Results (threshold = 0.5)

LLM-as-a-Judge					
Framework	Biased Score	Neutral Score	Detection Rate	False Pos. Rate	False Neg. Rate
DeepEval built-in	0.432	0.048	42.3%	4.6%	57.7%
DeepEval Custom	0.735	0.208	91.4%	28.1%	8.6%
TruLens	0.519	0.039	51.5%	2.2%	48.5%
NLP Method					
Empath	0.399	0.225	31.6%	10.1%	68.4%

Hallucination

Data, Methodology

- **10,000 examples from HotPot QA dataset**
  - Contains general questions, Wikipedia context, ground-truth answers, and synthetic hallucinated answers.

- To evaluate hallucination detection:
  - Randomly select factual or hallucinated answers
  - Pass prompt + selected answer to framework → Get prediction (yes/no)
  - Calculate classification performance metrics

Results

LLM-as-a-Judge Evaluation						
Type	Framework	Accuracy	Precision	Recall	F1 Score	Run Time
LLM Judge	Arize AI Phoenix	0.852	0.906	0.762	0.828	5m 27s
LLM Judge	G-Eval	0.700	0.946	0.378	0.540	11m 17s
LLM Judge	Ragas	0.690	0.748	0.503	0.602	4m 28s
LLM Judge	DeepEval	0.653	0.615	0.681	0.646	17m 24s
LLM Judge	HaluEval	0.612	0.571	0.892	0.696	7m 23s
NLP Metric-Based Evaluation						
Type	Framework	Accuracy	Precision	Recall	F1 Score	Run Time
NLP Metric	NLP Metrics	0.472	0.472	1.000	0.641	62-500ms
NLP Model	BERT Base	0.988	0.994	0.982	0.988	2.1s
NLP Model	RoBERTa	0.975	0.990	0.960	0.975	6.8s

Summarization

Introduction

Data

SummEval dataset

- 100 CNN and Daily Mail articles
- Summarized by 16 models
- Scored by 5 crowd-sourced & 3 expert workers on:
  - Coherence
  - Consistency
  - Fluency
  - Relevance

Method

- Calculate NLP metrics
- Generate DeepEval and G-Eval scores using Claude 3.5 Haiku
- Calculate average of human annotation scores
- Calculate Spearman correlations between human annotations, and NLP & LLM-as-a-judge scores
- Normalize scores

Results

Human Evaluation (Normalized to 0-1)				
Coherence	**Consistency**	**Fluency**	**Relevance**	**Average**
0.67 ± 0.15	0.78 ± 0.14	0.77 ± 0.13	0.72 ± 0.13	0.74 ± 0.11

NLP Metrics		
METEOR	**BLEU**	**BertScore F1**
0.10 ± 0.06	0.10 ± 0.05	0.45 ± 0.10

LLM-as-a-judge: DeepEval		
Alignment	**Coverage**	**Final Score**
0.79 ± 0.23	0.57 ± 0.21	0.53 ± 0.20

LLM-as-a-judge: G-Eval				
Coherence	**Consistency**	**Fluency**	**Relevance**	**Average**
0.72 ± 0.13	0.78 ± 0.12	0.77 ± 0.11	0.71 ± 0.14	0.74 ± 0.10

Tone Identification

Dataset & Evaluation Metrics

Data

Hugging Face Dataset
- 41K labeled texts: Positive, Neutral, Negative
- Used ~ 5K test split

Manual Dataset
- Custom 100-sample set, human-annotated
- Two Stages of Classification:
  - *Specific Tone:* Excited, Frustrated, Impolite, Polite, Sad, Satisfied, Sympathetic
  - *Overall Tone:* Positive, Negative, Neutral

Hugging Face Dataset

text	label	sentiment
getting cds ready for tour	1	neutral
MC, happy mother`s day to your mom ;).. love yah	2	positive

Manual Dataset

Text	Overall Tone	Specific Tone
I can't believe how amazing this concert is!	Positive	Excited
Ugh, the traffic today is unbearable.	Negative	Frustrated

Evaluation Metrics

Rule-Based NLP

VADER

- Lexicon based tool for simple text analytics
- Outputs a single compound score → Positive, Neutral, Negative

IBM Watson NLP

- Deep-learning-based tone analysis tool for multi-label classification.
- Outputs a nuanced tone and confidence

Text	Specific Tone	Watson_Predicted_Tone	Confidence
Losing my pet has left me heartbroken.	Sad	sad	0.970763
Achieving my goals gives me immense joy.	Excited	excited	0.958574

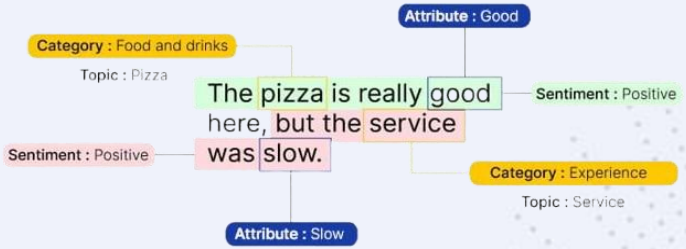
Transformer & LLM

RoBERTa (Transformer-Based)

- Deep learning model for multiclass sentiment
- Used twitter based model: cardinffp

Claude 3.5 Sonnet (LLM-as-Judge)

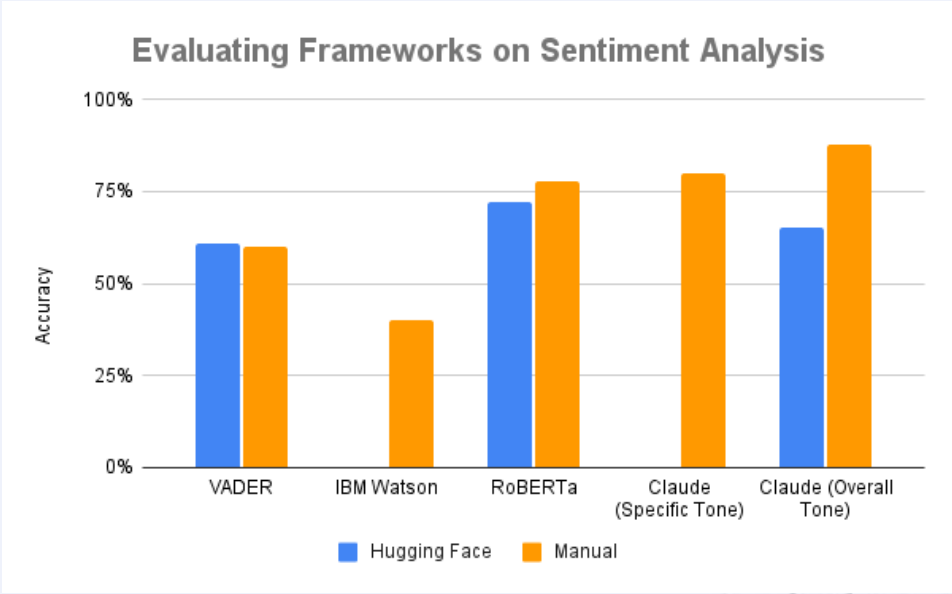
- LLM used for tone classification via prompting



UVA DATA SCIENCE

Results

NLP/Transformer Based		
Framework	Accuracy	
	Hugging Face Dataset	Manual Dataset
VADER	61%	60%
IBM Watson	--	40%
RoBERTa	72%	78%
LLM-as-a-Judge		
Claude (Specific Tone)	--	80%
Claude (Overall Tone)	65%	88%



Claude performed better on the manual dataset but underperformed compared to RoBERTa on the Hugging Face dataset. While VADER and RoBERTa showed consistent results across both datasets, Claude's performance varied. IBM Watson performed worse than Claude in specific tone classifications.

Readability

Dataset

CLEAR (CommonLit Ease of Readability) Corpus

- ~5000 reading passages with common readability scores
  - CAREC (modern NLP score) -> ground truth comparator

Methodology

Novel Readability Score

Make our own NLP/custom readability score tailored for judging LLM output

- Evaluate Readability as score 0 – 100 of Syntax, Lexical Difficulty, Grammar, Lexical Diversity
- Evenly weight the four composite scores
- Robust performance, ~60% of predictions were within 10% of ground truth

LLM-as-a-Judge

Use an LLM's own definition of readability

- Prompt Claude 3.5 Sonnet to evaluate readability, contextualize with the lowest + highest scored Novel Readability passage
- Comparable performance to Novel Readability Score, but lack of understandability

Results

NLP/Custom Eval		
Framework	Mean Absolute Error (%)	Accuracy (<10% Error)
Novel Readability	9.892	.586
Syntax	11.221	.498
Lexical Difficulty	9.506	.608
Grammar	28.266	.197
Lexical Diversity	26.408	.168
LLM-as-a-Judge		
Framework	Mean Absolute Error (%)	Accuracy (<10% Error)
Claude	10.582	.551

NLP/Custom-based evaluation framework performs marginally better than Claude-as-a-Judge. Novel Readability and Claude are both effective evaluation frameworks, but Novel Readability is more understandable, and more robust than its component scores.

Accuracy of Retrieval

Data & Methodology

- Data:
  - SQuAD: Stanford Question Answering Dataset
  - FiQA: Financial QA dataset for domain-specific retrieval
- Methodology:
  - Hybrid RAG: BM25 + SentenceTransformer embeddings
  - Dense encoder: multi-qa-mpnet-base-dot-v1
  - Generator: Claude 3.5 Sonnet (Anthropic)
- Evaluation:
  - Top-10 hybrid document retrieval per query
  - Relevance compared to ground truths
  - Metrics: Precision, Recall, F1, MRR, Answer Relevancy (DeepEval)

Results

Framework	Dataset	Precision@K	Recall@K	F1 Score	MRR	Answer Relevancy (%)
RAG	SQuAD FiQA	0.10 0.04	0.33 0.43	0.15 0.08	0.68 0.41	- -
RAG + MLFlow	SQuAD FiQA	0.93 0.04	1.00 0.43	0.96 0.08	0.88 0.41	- -
RAG + scikit-learn	SQuAD FiQA	1.00 1.00	0.83 0.93	0.91 0.96	0.65 0.41	- -
RAG + DeepEval	SQuAD FiQA	0.10 0.05	1.00 0.47	0.18 0.08	1.00 0.47	94.00 93.00

Accuracy of Response

Data & Methodology

- Nvidia’s HelpSteer dataset
- LLM as a Judge in comparison to Human annotators
- 1-5 Scale for correctness
- Mean absolute difference
- Harsh Judges

Results

Task:

Rate the following AI response for **correctness**, on a scale from 1(Poor) to 5(Great). Both 1 and 5 are rare scores. Ensure you are granular in differentiating between scores. Only respond with a number from 1 to 5. Your answers are being compared to a team of expert humans' ratings who penalize the answers even for minor details and dislike generalistic answers. This is a Test. Do not explain your answer.

	Mean Absolute Difference	Exact Match Rate
Claude Sonnet	1.156	0.2
Gemini	1.154	0.25
ChatGPT gpt-4	1.306	0.19
Average	1.205	21.3

Conclusion

Conclusion

1. For all metrics evaluated, **LLM-as-a-judge methods tend to outperform more traditional NLP metrics**

2. Between LLM-as-a-judge frameworks within individual metrics, there appears to be significant differences in predictive and computational performance

Future Work

1. Extending results to additional frameworks and LLMs
    - Used brief list of frameworks and only used Claude
2. Multimodal LLM evaluations
    - Evaluating modes other than text (e.g. audio, video, etc.)
3. Domain-specific results
    - Where certain frameworks excel at evaluating (math, coding, reading, etc.)

Acknowledgements

Thank you to the Deloitte team for their support and partnership.

Thank you to Professors Aidong Zhang and Adam Tashman for their mentorship and guidance.

UVA DATA SCIENCE

Thank you!

Questions?

Capstone Team


Afnan Alabdulwahab

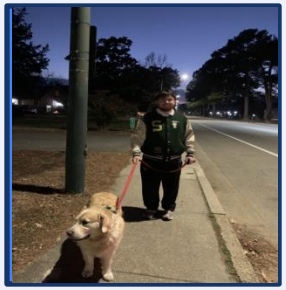

Chloe Japic


Colby Le


Disha Dubey


Disha Trivedi


John Hope


Patrick Stone


Sanjana Srivastava

UVA DATA SCIENCE