

Lecture 6: Evaluation

CS 2281, Fall 2024

Julia Neagu & Deanna Emery

Quotient AI

Quick intro

Julia -

Deanna -

Quotient AI -

Today

- Introduction to LLM evals
- General model quality evaluation
- Downstream task evaluation
- Real-world use cases and examples
- Conclusion and Q&A

What is LLM evaluation?

- We will focus on **evaluation of LLM inference** (i.e. post-training or fine-tuning)
- **Two main purposes:**
 1. Understanding **general quality** of a trained model
 2. Assessing **performance on specific downstream tasks**
- **#1** is also called **benchmarking**
- **#2** adds addl. **degrees of freedom**: prompt engineering, RAG systems etc.

Generative AI Gone Wrong

B B C

Home News US Election Sport Business Innovation Culture Arts Travel Earth Video Live

Airline held liable for its chatbot giving passenger bad advice - what this means for travellers

DPD AI error causes chatbot to swear, calls itself the 'worst delivery service' to disgruntled user: report



A car dealership added an AI chatbot to its site. Then all hell broke loose.

businessinsider.com · 2023 ▾

On Sunday, Aharon Horwitz was listening to a podcast when he got an unusual Slack alert. Horwitz is the CEO of Fullpath, a tech company that sells marketing and sales software for [car dealerships](#). The automated Slack alert

General Quality Evaluation

Primary method: Benchmarks

- **Definition:** Sets of data with expected outputs
- **Strict performance metrics** (e.g. accuracy / exact match)
 - *Note: even this can be sometimes be implemented in a few ways*
Example: [GSM8K in Eleuther's eval harness.](#)

```
32 filter_list:
33   - name: "strict-match"
34     filter:
35       - function: "regex"
36         regex_pattern: "#### (\\-?[0-9\\.\\,\\,]+)"
37       - function: "take_first"
38   - name: "flexible-extract"
39     filter:
40       - function: "regex"
41         group_select: -1
42         regex_pattern: "(-?[$0-9.,]{2,})|(-?[0-9]+)"
43       - function: "take_first"
44 metadata:
45   version: 3.0
```

General Quality Evaluation

- Popular benchmarks:
 - [GS8MK](#) (Grade School Math 8K)

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

Solution: Beth bakes 4 2 dozen batches of cookies for a total of $4 \times 2 = 8$ dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12 \times 8 = 96$ cookies
She splits the 96 cookies equally amongst 16 people so they each eat $96 / 16 = 6$ cookies

Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs \$3.50?

Mrs. Lim got 68 gallons - 18 gallons = 50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = 200 gallons.
She was able to sell 200 gallons - 24 gallons = 176 gallons.
Thus, her total revenue for the milk is $\$3.50/\text{gallon} \times 176 \text{ gallons} = \616 .

Final Answer: 616

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?

Solution: Tina buys 3 12-packs of soda, for $3 \times 12 = 36$ sodas
6 people attend the party, so half of them is $6 / 2 = 3$ people
Each of those people drinks 3 sodas, so they drink $3 \times 3 = 9$ sodas
Two people drink 4 sodas, which means they drink $2 \times 4 = 8$ sodas
With one person drinking 5, that brings the total drank to $5 + 9 + 8 + 3 = 25$ sodas
As Tina started off with 36 sodas, that means there are $36 - 25 = 11$ sodas left

Final Answer: 11

General Quality Evaluation

- Popular benchmarks:
 - [HumanEval](#) (coding task generation)

```
def incr_list(l: list):  
    """Return list with elements incremented by 1.  
    >>> incr_list([1, 2, 3])  
    [2, 3, 4]  
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])  
    [6, 4, 6, 3, 4, 4, 10, 1, 124]  
    """  
    return [i + 1 for i in l]
```


General Quality Evaluation

- **Popular benchmarks:**

- **MMLU** (Massive Multitask Language Understanding): tasks ranging from simple math to legal reasoning

Professional Law

As Seller, an encyclopedia salesman, approached the grounds on which Hermit's house was situated, he saw a sign that said, "No salesmen. Trespassers will be prosecuted. Proceed at your own risk." Although Seller had not been invited to enter, he ignored the sign and drove up the driveway toward the house. As he rounded a curve, a powerful explosive charge buried in the driveway exploded, and Seller was injured. Can Seller recover damages from Hermit for his injuries?

(A) Yes, unless Hermit, when he planted the charge, intended only to deter, not harm, intruders. ❌

(B) Yes, if Hermit was responsible for the explosive charge under the driveway. ✅

(C) No, because Seller ignored the sign, which warned him against proceeding further. ❌

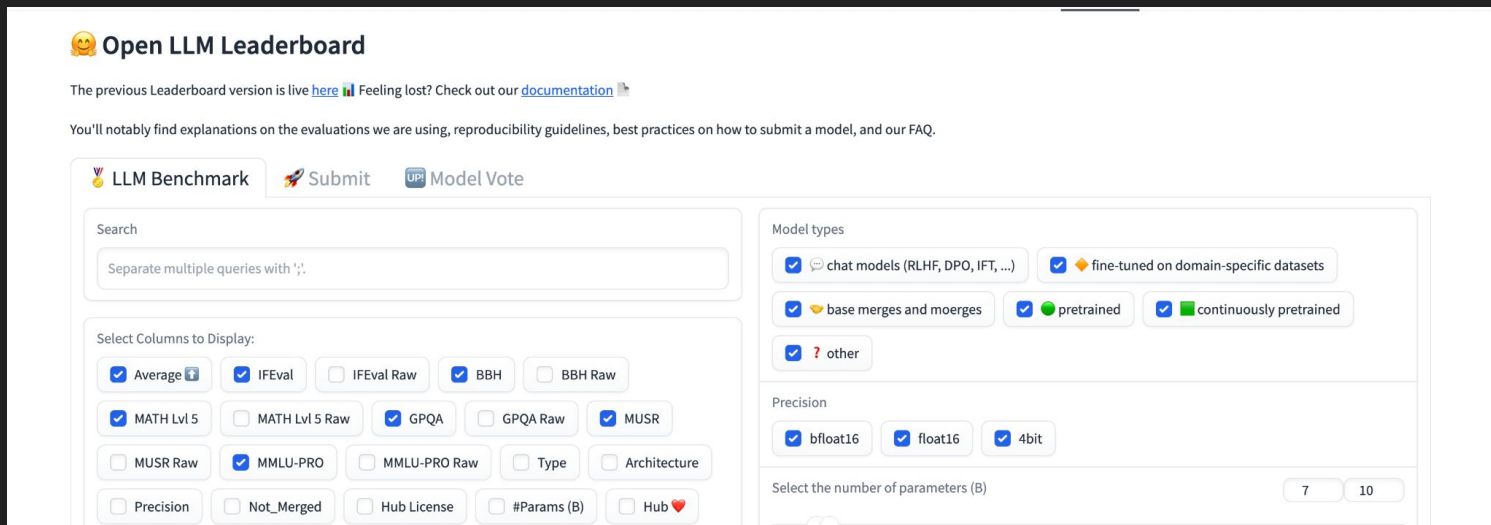
(D) No, if Hermit reasonably feared that intruders would come and harm him or his family. ❌

Figure 2: This task requires understanding detailed and dissonant scenarios, applying appropriate legal precedents, and choosing the correct explanation. The green checkmark is the ground truth.

Community-driven LLM benchmarking

Hugging Face Open LLM Leaderboard



- Community-maintained benchmark for open-source LLMs
- Covers various tasks: reasoning, math, coding, etc.
- Allows direct comparison of model performance



😊 Open LLM Leaderboard

The previous Leaderboard version is live [here](#) 🇺🇸 Feeling lost? Check out our [documentation](#) 📖

You'll notably find explanations on the evaluations we are using, reproducibility guidelines, best practices on how to submit a model, and our FAQ.

LLM Benchmark  Submit  Model Vote

Search

Separate multiple queries with ','

Select Columns to Display:

- Average 📊
- IFEval
- IFEval Raw
- BBH
- BBH Raw
- MATH Lvl 5
- MATH Lvl 5 Raw
- GPQA
- GPQA Raw
- MUSR
- MUSR Raw
- MMLU-PRO
- MMLU-PRO Raw
- Type
- Architecture
- Precision
- Not_Merged
- Hub License
- #Params (B)
- Hub ❤️

Model types

- chat models (RLHF, DPO, IFT, ...)
- fine-tuned on domain-specific datasets
- base merges and moerges
- pretrained
- continuously pretrained
- ? other

Precision

- bfloat16
- float16
- 4bit

Select the number of parameters (B)

7 10

General Quality Evaluation

Example of foundational model benchmarks from the Llama-3

Paper: <https://arxiv.org/pdf/2407.21783>

Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 ⁰²²⁰	GPT-4o	Claude 3.5 Sonnet
General	MMLU ^(5-shot)	69.4	72.3	61.1	83.6	76.9	70.7	87.3	82.6	85.1	89.1	89.9
	MMLU ^(0-shot, CoT)	73.0	72.3 [△]	60.5	86.0	79.9	69.8	88.6	78.7 [◊]	85.4	88.7	88.3
	MMLU-Pro ^(5-shot, CoT)	48.3	–	36.9	66.4	56.3	49.2	73.3	62.7	64.8	74.0	77.0
	IFEval	80.4	73.6	57.6	87.5	72.7	69.9	88.6	85.1	84.3	85.6	88.0
Code	HumanEval ^(0-shot)	72.6	54.3	40.2	80.5	75.6	68.0	89.0	73.2	86.6	90.2	92.0
	MBPP EvalPlus ^(0-shot)	72.8	71.7	49.5	86.0	78.6	82.0	88.6	72.8	83.6	87.8	90.5
Math	GSM8K ^(8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6	96.8	92.3 [◇]	94.2	96.1	96.4 [◇]
	MATH ^(0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1	73.8	41.1	64.5	76.6	71.1
Reasoning	ARC Challenge ^(0-shot)	83.4	87.6	74.2	94.8	88.7	83.7	96.9	94.6	96.4	96.7	96.7
	GPQA ^(0-shot, CoT)	32.8	–	28.8	46.7	33.3	30.8	51.1	–	41.4	53.6	59.4
Tool use	BFCL	76.1	–	60.4	84.8	–	85.9	88.5	86.5	88.3	80.5	90.2
	Nexus	38.5	30.0	24.7	56.7	48.5	37.2	58.7	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	95.2	–	95.2	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	83.4	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	100.0	100.0	90.8
Multilingual	MGSM ^(0-shot, CoT)	68.9	53.2	29.9	86.9	71.1	51.4	91.6	–	85.9	90.5	91.6

Table 2 Performance of finetuned Llama 3 models on key benchmark evaluations. The table compares the performance of the 8B, 70B, and 405B versions of Llama 3 with that of competing models. We **boldface** the best-performing model in each of three model-size equivalence classes. [△]Results obtained using 5-shot prompting (no CoT). [◊]Results obtained without CoT. [◇]Results obtained using zero-shot prompting.

Benchmark leakage

- Benchmarks are often based on public data, so there's a risk that the benchmark data leaks into training data
- **Consequences:**
 - Inflated performance metrics
 - Overfitting to specific benchmarks
 - Unreliable evaluation of true model capabilities
- **Paper:** Xu et al, *Benchmarking Benchmark Leakage in Large Language Models* (2024)

<https://arxiv.org/pdf/2404.18824>

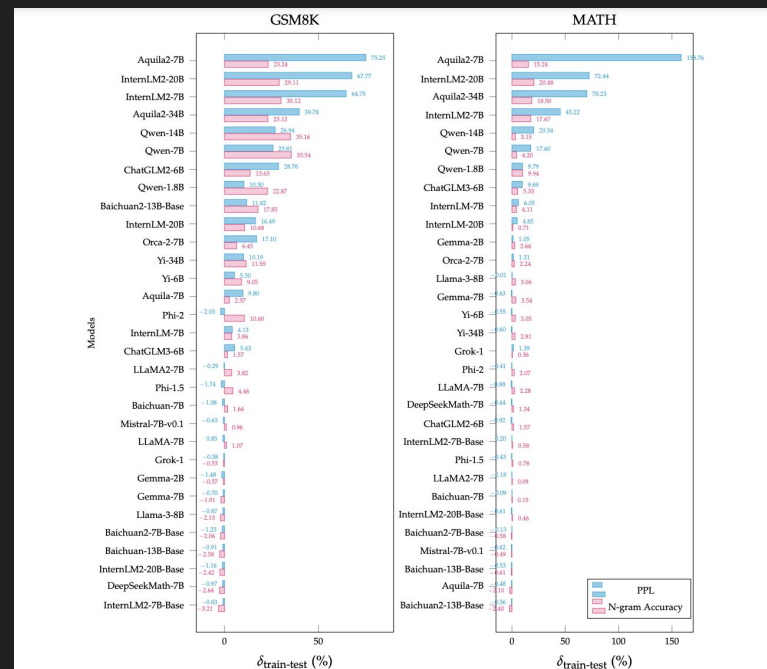


Figure 1: The relative possibility that various models conduct verbatim training on the training set of a benchmark over test set to enhance capabilities (measured based on PPL and N-gram Accuracy). Models exhibiting near-zero possibilities suggest either the absence of training and test split or the use of both splits in the training process. This metric does not imply cheating, but rather indicates the potential use of the benchmark data during the (pre-)training phase; while using benchmarks to enhance capabilities is acceptable, the lack of relevant documentation can reduce transparency, potentially resulting in unfair comparisons and hindering the field's healthy development.

General Quality Evaluation

Benchmarks are tricky to get right!

- **Prompt dependency:** results vary based on task phrasing
- **Model configuration effects:** performance affected by hyperparameters
- **Replication challenges:** difficult to reproduce results
- **Costs:** running comprehensive benchmarks can be expensive
- **Benchmark saturation:** top models approaching human performance

General Quality Evaluation

And even then, face significant limitations:

- **Quality issues:**
 - Poorly formulated questions or answers
 - Outdated or biased information
- **Metric limitations:**
 - Focus on accuracy or exact match
 - Often assume single correct answer
- **Real-world applicability:**
 - May not reflect diverse use cases
 - Limited correlation with task-specific performance

Benchmark quality concerns

"On summarization, we find that ... the reference summaries in standard summarization datasets (e.g., CNN/DM, XSUM) are actually worse (under the same human evaluations)."

<https://arxiv.org/abs/2211.09110>

HellaSwag Example

*Men are standing in a large green field playing lacrosse. People **is** around the field watching the game. men*

- *are holding tshirts watching **int** lacrosse playing.*
- *are being interviewed in a podium in front of a large group and a gymnast is holding a microphone for the announcers.*
- *are running side to side **of** the **ield** playing lacrosse trying to score.*
- *are in a field running around playing lacrosse.*

Crowdsourced LLM evaluations

Chatbot Arena (<https://lmarena.ai/>)

- Elo-based rating system for LLMs
- Users compare responses from different models
- Provides dynamic, user-centric evaluation
- Paper: [Chiang et al, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference"](#) (2024)

✂ Chatbot Arena: Benchmarking LLMs in the Wild

[Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#)

📖 Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can continue chatting until you identify a winner.
- Vote won't be counted if model identity is revealed during conversation.

🏆 Arena Elo [Leaderboard](#)

We collect 300K+ human votes to compute an Elo-based LLM leaderboard. Find out who is the 🏆 LLM Champion!

🗨 Chat now!

Expand to see the descriptions of 36 models

Model A

i found a property i would like to buy. please draft a letter to the owners expressing my interest to purchase and convincing them of completing a private sale with me. list the benefits of private sale and how it will be done with our lawyers.

Subject: Expression of Interest in Your Property for a Private Sale

Dear [Owner's Name],

I hope this message finds you well. My name is [Your Name], and I am writing to express my sincere interest in purchasing your property located at [Property Address]. I have recently had the privilege of viewing your property and am captivated by its unique features and the potential it holds.

I would like to propose a private sale as a method of transaction. This proposal aims to create a beneficial situation for both parties involved, and I am confident that this process will be both smooth and efficient.

The advantages of a private sale are numerous. Firstly, it offers a more personal approach, as it allows for direct negotiation between the buyer and seller, resulting in potentially better terms for both parties. Secondly, a private sale can generally be completed more quickly than traditional sales, as there are fewer parties involved in the decision-making process. Thirdly, the

Model B

i found a property i would like to buy. please draft a letter to the owners expressing my interest to purchase and convincing them of completing a private sale with me. list the benefits of private sale and how it will be done with our lawyers.

Subject: Expression of Interest to Purchase Your Property - Private Sale Proposal

Dear [Owner's Name],

I hope this letter finds you well. My name is [Your Name], and I recently had the opportunity to view your property at [Property Address]. I am writing to express my genuine interest in purchasing your property and to propose the possibility of completing this sale privately, directly between us, with the support of our respective lawyers.

Firstly, I want to commend you on the care and upkeep of your property. It's clear that a great deal of effort has been invested into making it a desirable and welcoming space. After careful consideration and viewing several properties, I am convinced that your property meets all the criteria I am looking for in a home.

I understand that selling a property is a significant decision, and I believe that a private sale could offer considerable benefits to both of us. Here are a few reasons why I think a private sale would be advantageous:

👉 A is better

👉 B is better

👉 Tie

👉 Both are bad

🗨 Enter your prompt and press ENTER

Send

Crowdsourced LLM evaluations

Advantages:

- Large-scale, diverse user input
- Captures real-world preferences and use cases
- Continually updated as new models emerge

Challenges:

- **Biased evaluator pool** (e.g. LLM hobbyists and researchers)
- **Significant data contamination** (e.g. toxicity, NSFW etc.)
- Difficulty in controlling evaluation conditions
- Potential for gaming or manipulation of results
- **Not reusable**
- **Cannot translate to private use-cases**

Downstream task evaluation

- The magic behind LLMs is the expectation that they can perform a very wide range of tasks

What are some tasks LLMs can perform?

Downstream task evaluation

- **The magic behind LLMs is the expectation that they can perform a very wide range of tasks:**
 - Translation, Question answering, Summarization, Code generation, Creative language generation...

Question(s) :

- How would you evaluate Q&A?
- How would you evaluate a summary?
- How would you evaluate code generation?

Reference-based v Reference-free evaluation

Reference-based evaluation

Deterministic metrics:

- **Syntactic similarity:**

- Accuracy
- Exact Match
- Bilingual Evaluation Understudy (BLEU)
<https://arxiv.org/pdf/1804.08771>
- Recall-Oriented Understudy for Gisting Evaluation (ROUGE)
<https://aclanthology.org/W04-1013/>

- **Semantic similarity**

- BERTScore <https://arxiv.org/abs/1904.09675>
- Semantic Textual Similarity
https://sbert.net/docs/sentence_transformer/usage/semantic_textual_similarity.html

Reference-based evaluation

Reference: "The cat is on the mat."

LLM Output: "The cat and the dog."

BLEU Score:

1-grams: ["The", "cat", "the"] --> 3/5 1-grams in output

2-grams: ["The cat"] --> 1/4 2-grams in output

$$\text{BLEU} = 1 * \exp(0.5 * (\log(3/6) + \log(1/4))) = 0.64$$

ROUGE Scores:

- ROUGE-2 Precision = 2/4

- ROUGE-2 Recall = 1/5

- ROUGE-2 F1 = 0.22

- ROUGE-L Precision = 3/5

- ROUGE-L Recall = 3/6

- ROUGE-L F1 = 0.54

Reference-based evaluation

```
references = ["the cat is on the mat"]
predictions = ["the cat and the dog"]
results = bertscore.compute(predictions:
```

results

```
{'precision': [0.8781610131263733],
 'recall': [0.8737168312072754],
 'f1': [0.8759333491325378],
 'hashcode': 'roberta-large_L17_no-idf_
```

```
references = ["cat is on mat"]
predictions = ["cat and dog"]
results = bertscore.compute(predictions:
```

results

```
{'precision': [0.8450708389282227],
 'recall': [0.8164956569671631],
 'f1': [0.8305374979972839],
 'hashcode': 'roberta-large_L17_no-idf_
```

```
reference = "the cat is on the mat"
output = "the cat and the dog"
```

```
output_embedding = model.encode(output.lower()).reshape(1, -1)
reference_embedding = model.encode(reference.lower()).reshape(1, -1)
```

```
similarity = cosine_similarity(output_embedding, reference_embedding)[0][0]
similarity
```

0.2915592

```
reference = "cat is on mat"
output = "cat and dog"
```

```
output_embedding = model.encode(output.lower()).reshape(1, -1)
reference_embedding = model.encode(reference.lower()).reshape(1, -1)
```

```
similarity = cosine_similarity(output_embedding, reference_embedding)[0][0]
similarity
```

0.3686059

Reference-based evaluation

Deterministic metrics:

Pros:

- Easy to compute, reproduce, and audit

Cons:

- May not capture nuanced aspects of language
- Sensitive to text preprocessing / cleaning
- Can be gamed or optimized for without true improvement
- Often not correlated with human judgements
- Require reference outputs

Still valuable, but limited for advanced LLMs

Beyond reference-based evaluation

Challenges with reference-based evaluation:

- Many tasks don't have single "correct" answers
- Creative tasks are particularly hard to evaluate

Alternative approaches:

- Human evaluation
- LLM-as-judge
- Task-specific automated metrics

Importance of multifaceted evaluation:

- Combine multiple metrics
- Consider both automated and human evaluation
- Align evaluation with specific use case requirements

Human Evaluation

ChatGPT 4o ▾

Draw me a puppy



Here is the drawing of a cute puppy you requested!



Human Evaluation (*vibes*)

Relying on human intuition and judgment

Pros:

- Can capture subtle qualities missed by automatic metrics
- Useful for assessing overall quality and coherence
- Accounts for specialized subject matter expertise
- Do not require reference or ground truth data

Cons:

- Subjective and potentially inconsistent
- Doesn't scale well
- Best used in conjunction with other evaluation methods

Human Evaluation

If you are going to read one paper out of this entire session read this one:

Shankar et. al (2024) - "Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences"
<https://arxiv.org/pdf/2404.12272>

Our findings find overall support for EVALGEN, with one important caveat. We observed a “catch-22” situation: to grade outputs, people need to externalize and define their evaluation criteria; however, the process of grading outputs helps them to define that very criteria. We dub this phenomenon *criteria drift*, and it implies that *it is impossible to completely determine evaluation criteria prior to human judging of LLM outputs*. Even when participants graded first, we observed that they still refined their criteria upon further grading, even going back to change previous grades. Thus, our findings suggest that users need evaluation assistants to support rapid iteration over criteria and implementations *simultaneously*. Since criteria are *dependent* upon LLM outputs (and not independent from them), this raises questions about how to contend with criteria drift in the context of other “drifts”—e.g., model drift [4], prompt edits, or upstream changes in a chain. Our findings also (i) underscore the necessity of *mixed-initiative* approaches to the alignment of LLM-assisted evaluations that also embrace messiness and iteration, and (ii) raise broader questions about what “alignment with user preferences” means for evaluation assistants.

LLM-as-a-Judge

Using one LLM to evaluate outputs of another

Examples:

- **Bigger models (GPT-4) evaluating outputs of smaller models**
- **Panel of judges**
 - Verga et al (2024) - "Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models"
<https://arxiv.org/abs/2404.18796>
- **Specialized models trained for evaluation tasks**
 - **Natural Language Inference (NLI) models** - a classification model that measures whether a statement follows from a given premise. Typically used to detect hallucinations
 - Zhu et al (2023) - "JudgeLM: Fine-tuned Large Language Models are Scalable Judges" <https://arxiv.org/abs/2310.17631>

LLM-as-a-Judge

Pros:

- Can potentially capture more nuanced aspects of quality
- Opens up the kinds of qualities to test for - you can ask a model to do anything!
- Scalable compared to human evaluation

Cons:

- Raises the question: "Who validates the validators?"
- May inherit biases or limitations of the evaluating LLM
 - Chen et al (2024) - "Humans or LLMs as the Judge? A Study on Judgement Bias" <https://arxiv.org/pdf/2402.10669>

LLM-as-a-Judge

A common implementation of judge LLMs is rubrics – structured prompts that guide their evaluation process. Rubrics can include:

- **Evaluation criteria:** The specific aspects of the output to be assessed.
- **Scoring guidelines:** How to rate or score each criterion.
- **Examples:** Illustrations of high-quality and low-quality responses.

A Prompt templates

We list the prompt templates for LLM judges. Please refer to our github repository³ for full details.

```
[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[User Question]
{question}

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]
```

Figure 5: The default prompt for pairwise comparison.

```
[System]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

[Question]
{question}

[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]
```

Figure 6: The default prompt for single answer grading.

Human-in-the-Loop Evaluation

Combining automated metrics with human judgment

Approaches:

- **Generating assertions** (SPADE)

Shankar et al - "SPADE: Synthesizing Data Quality Assertions for Large Language Model Pipelines" (2024)

<https://arxiv.org/abs/2401.03038>

- **Generating LLM judges** (SMELL)

Emery & Liounis - "Subject-Matter Expert Language Liaison (SMELL): A framework for aligning LLM evaluators to human feedback" (2024) [link](#)

Human-in-the-Loop Evaluation

Approaches *(continued)*:

- **Training reward models on human preferences**
 - Rafailov et al (2023) - "*Direct Preference Optimization: Your Language Model is Secretly a Reward Model*"
<https://arxiv.org/abs/2305.18290>
 - Ethayarajh et al (2024) - "*KTO: Model Alignment as Prospect Theoretic Optimization*" <https://arxiv.org/abs/2402.01306>
 - Jung et al (2024) - "*Binary Classifier Optimization for Large Language Model Alignment*" <https://arxiv.org/abs/2404.04656>

Human-in-the-Loop Evaluation

Pros :

- Balances efficiency of automated methods with human insight
- Can capture complex aspects of quality and appropriateness

Cons :

- More resource-intensive than purely automated methods
- Potential for human bias or inconsistency
- Crucial for sensitive applications and safety considerations

Domain-Specific Evaluation Datasets

Importance of user-created datasets:

- Tailored to specific use cases and domains
- Better reflect real-world application needs
- Allow for more meaningful evaluation in context

Characteristics:

- Highly relevant to the target domain
- Often smaller but more focused than general benchmarks
- May include proprietary or sensitive information

Creation process:

- Curated by domain experts
- Often involves labeling or annotating real user data

Domain-Specific Evaluation Datasets

Advantages:

- More accurate assessment of model performance for specific tasks
- Helps identify domain-specific biases or errors
- Enables continuous improvement in targeted applications

Challenges:

- Ensuring dataset quality, consistency and coverage
- Balancing specificity with generalizability
- Keeping datasets updated as domain knowledge or scenarios evolve

Best practices:

- Regular updates to reflect changing domain knowledge
- Collaboration between domain experts and ML practitioners
- Careful consideration of privacy and ethical concerns

Future Directions in Evaluation

- Improving human-in-the-loop evaluation methodologies
- Evaluating CoT / Reasoning (think: o1)
- Expanding evaluation frameworks beyond text

Evals for LLMs in the wild

Evals for Policy Compliance: Wayfair Case Study

Goals

- Ensure LLM outputs align with company policies and regulations
- Unique to each company's specific use case and industry

Key aspects:

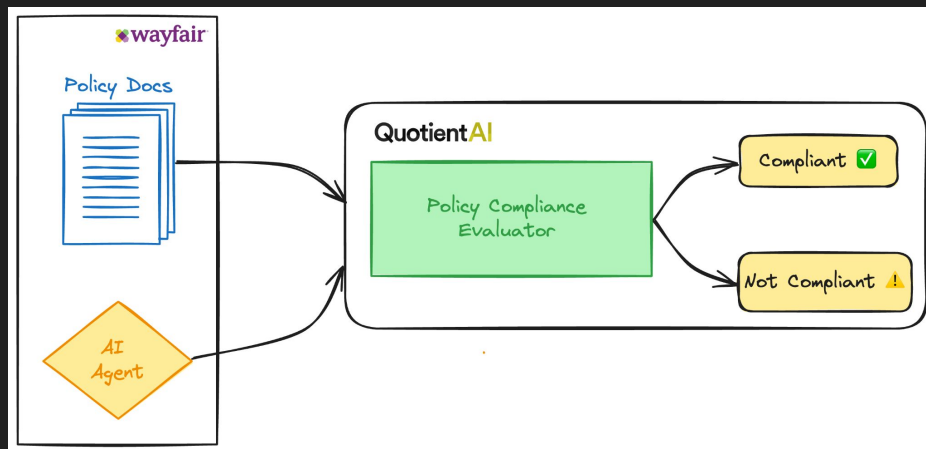
1. **Customization:** Tailored to company-specific policies
2. **Dynamic:** Regular updates to reflect policy changes

Case Study: Wayfair's Customer Support AI

Evals for Policy Compliance: Wayfair Case Study

We tested the Evaluator w/ the following methodology:

- We identified the relevant Wayfair policies for each real conversation
- For each of those conversations and relevant policies, we generated multiple policy-violating LLM responses using the Evaluation Benchmark Generator
- 70% of data points out of a 100 datapoint sub-sample were confirmed by Subject Matter Expert (SME) reviewers to be indeed policy-violating



Lessons from Building with LLMs

- **Real-world insights from deploying and evaluating LLMs**
- **Key takeaways:**
 - Importance of continuous evaluation in production
 - Balancing automated metrics with user feedback
 - Adapting evaluation strategies as models evolve
- **Reference:** Yan et al - "What We Learned from a Year of Building with LLMs" [O'Reilly article](#)

Questions?