How Companies Like OpenAI Ensure Their LLMs Are of High Quality

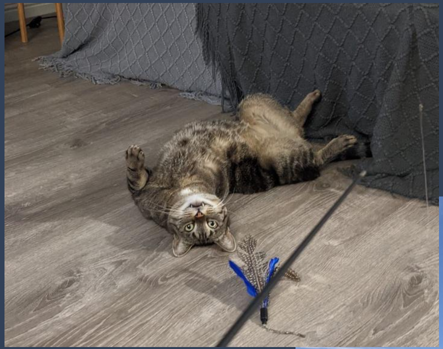
Evaluating Robustness, Accuracy, and Safety in Large Language Models

Introduction – Marina Wyss

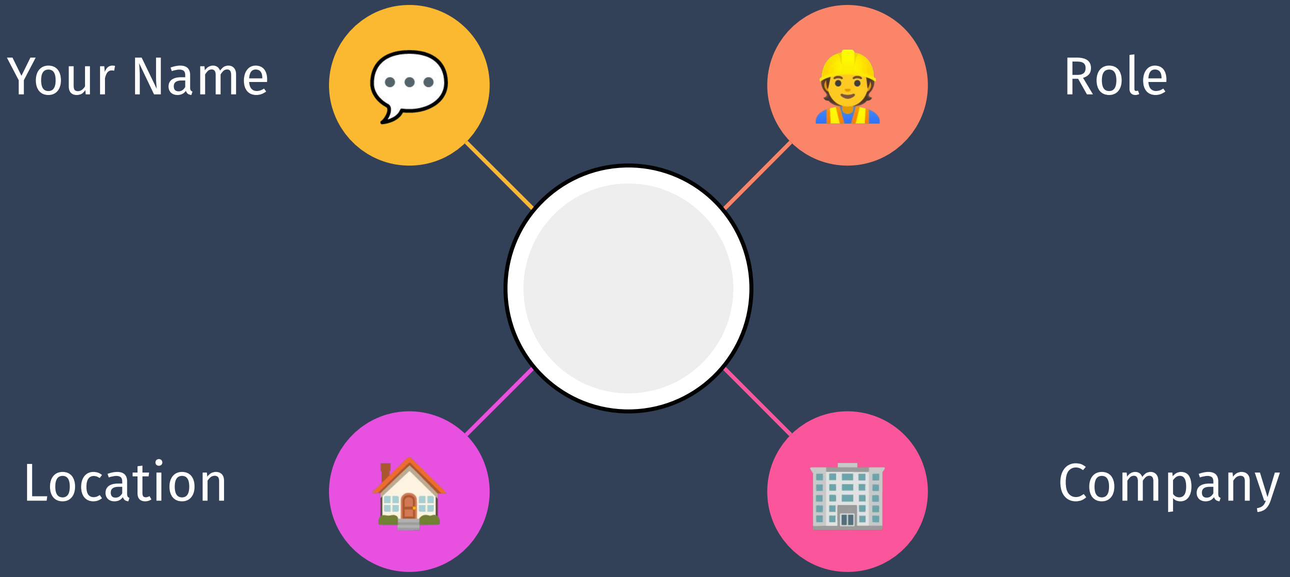
Applied Scientist at Twitch/Amazon

- Studied at U.C. Berkeley for undergrad, and did my Master's in social data science in Berlin.
- Started at a statistical consulting firm in Berlin, then worked on ML problems at Coursera and now Twitch.
- Most of my work has been focused on building production ML pipelines, ML Ops, and recently LLMs.
- I also work as a data science mentor and pro-bono consultant. I love teaching and helping people get started in this super fun and interesting field!

Fun fact: I have a three-legged cat from Poland named Arnold



About You...



Agenda

- Introduction to LLMs
- Introduction to LLM Performance Evaluation
- Supervised LLM Evaluation
- Interactive Demo: Evaluating an LLM Using Standard Metrics
- Advanced Evaluation Techniques
- Case Study: Real-World Application of Evaluation Techniques at OpenAI
- Future Directions and Challenges
- Q&A

Context: Significance & Expectations

- LLMs have rapidly become a major part of many user-facing applications.
- Developers need to know how to measure the quality of their models before and after deployment to avoid mistakes that could harm users or the business.
- Today's presentation is an introduction to this complex field. We'll talk about the different approaches at a high level, including a case study and demo.

Introduction to LLMs

Overview of LLMs

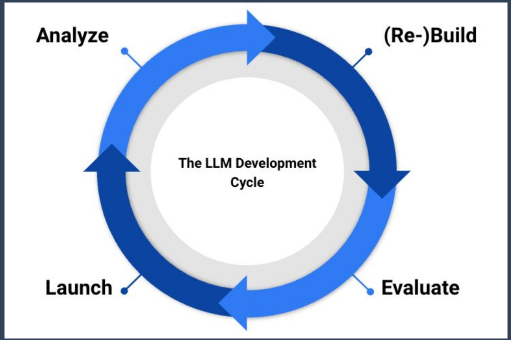
- LLMs are a type of ML model that is trained on vast amounts of text data to understand and generate **human-like text**.
- Trained using **deep learning**.
- LLMs excel at tasks like **translation, summarization, question-answering, and even creative writing**.

Key Players

- OpenAI
- Google
- Meta AI
- Microsoft
- Anthropic
- And more!

The LLM Lifecycle

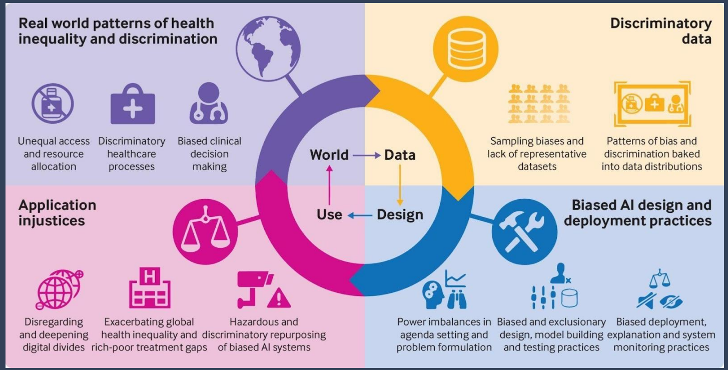
- Data Collection and Preprocessing
- Model Training
- Fine-tuning
- **Evaluation <- Our focus today!**
- Deployment and Monitoring
- Maintenance and Updates



LLM Performance Evaluation

Applications of LLM Performance Evaluation

- Model **comparison**
- **Bias detection** and mitigation
- **User satisfaction** and trust



Types of Evaluation for LLMs

- **System Evaluation**
  - Focus on the components we control, such as prompts and context.
  - Metrics like input-output determination efficiency, model perplexity, or retrieval relevance.
- **Model Evaluation**
  - Focus on the raw capability of the model, e.g. their ability to understand, generate, and manipulate language within the appropriate context.

Tools and Methods

- Automated Metrics
- Benchmarking
- Human Evaluation
- LLM-as-a-Judge
- Online Engagement Metrics
- Evaluation Platforms

Many Potential Things to Evaluate!

Evaluation criteria should be tailored to the specific application.

There are **many** potential things to consider!

- Task-specific (e.g. summarization, NER, RAG, Q&A)
- Responsible AI
- Fairness
- Robustness
- Factuality
- Speed/Cost
- Quality
- Consistency and generalizability

Key Characteristics of a Good Evaluation

- Focuses on the **most critical outcomes** of your LLM application.
- Uses a **small number of metrics** that are easy to interpret and understand.
- Should be **fast, reliable, and automatic** to compute.
- Tested on datasets that are **diverse and representative of real-world** scenarios.
- Metrics should be **highly correlated with human judgment** to ensure they reflect true performance.
- Enables **monitoring** score changes over time for continuous improvement.

Supervised LLM Evaluation

Specialized Metrics - Why?

- **Importance of Specialized Metrics:**
  - Capturing linguistic nuances
  - Assessing contextual understanding
  - Measuring generative quality
- **Unique Challenges in LLMs:**
  - Ambiguity and context sensitivity
  - Bias and ethical considerations
  - Managing large-scale and complex outputs

Gold Standard Data Set

- Like with any supervised learning task, we need a **labeled dataset.**
- Should be **diverse and representative**.

We can use LLMs to help with this part, too!

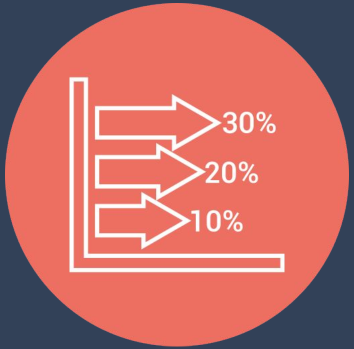
Fundamental Evaluation Metrics

- Classification Metrics (F1, Precision, Recall, etc.)
- Perplexity
- BLEU
- ROUGE

Classification Metrics

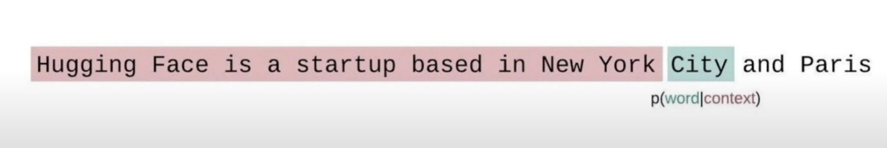
For classification tasks (e.g. sentiment analysis), **we can use typical classification metrics:**

- Precision
- Recall
- F1
- Accuracy



Perplexity

- Perplexity measures **how well a language model predicts a sample of text**.
- Lower perplexity indicates better predictive performance.
- **Advantages**:
  - Simple and widely used metric for language model evaluation.
  - Provides a quantitative measure of prediction accuracy.
- **Limitations**:
  - Does not capture context understanding, coherence, or relevance.
  - May not reflect real-world performance or user satisfaction.



BLEU (Bilingual Evaluation Understudy)

- Measures the quality of generated text by comparing it to one or more reference texts. Focuses on **precision**.
- It evaluates how many n-grams (contiguous sequences of n items) in the candidate text match the reference text.
  - BLEU calculates **precision** for n-grams of different lengths.
- Advantages:
  - Provides a quantitative measure of translation accuracy.
- Limitations:
  - Focuses on exact word matches, often missing context and semantic meaning.
  - Penalizes different word choices that might be correct, reducing the ability to capture paraphrased or rephrased content.
  - May not reflect human judgment of translation quality.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- Evaluates the quality of summaries by comparing them to reference summaries, focusing on **recall.**
  - ROUGE-N: Measures n-gram overlap (e.g., ROUGE-1 for unigrams, ROUGE-2 for bigrams).
  - ROUGE-L: Measures the longest common subsequence (LCS) between the candidate and reference summaries.
  - ROUGE-S: Measures the overlap of skip-bigrams (pairs of words in their sentence order, allowing for gaps).

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- **Advantages**
  - Focuses on how much of the reference content is captured in the generated text, making it effective for evaluating content preservation.
  - Versatile: Can be used across various text generation tasks beyond summarization.
- **Limitations**
  - Focuses on surface-level similarity, potentially overlooking deeper semantic meaning.
  - May penalize creative but valid paraphrasing.
  - Reference summaries are required, which may not capture all acceptable summaries for a given text.

Demo

Limitations

- Standard metrics are a good starting point, but **may miss nuance**.
- For example, BLEU might miss tone, style, or intended meaning of the original text.
- So, we also need some more advanced techniques we can use!

Advanced Evaluation Techniques

Advanced Techniques

- Benchmarking
- Human-in-the-Loop evaluations
- Automated tools and frameworks
- LLM-as-a-Judge
- Online Evaluation

Benchmarking

- Benchmarks **use known datasets** to evaluate LLMs by comparing generated outputs to correct answers.
    - Different benchmarks focus on various features like factual knowledge, math, reasoning, and language understanding.
- Examples include GLUE, SuperGLUE, HellaSwag, TruthfulQA, and MMLU.
- Tools like Big Bench, OpenAI Evals, and others assess general and specific tasks for broader evaluation.

Human-in-the-Loop

- **Qualitative** Assessment
- Alignment to the **Real World**
- **Bias Detection**

Evaluation Criteria

- Accuracy of the generated text.
- Relevance
- Fluency
- Transparency
- Safety
- Human Alignment

Automated Tools

- Offer **speed and scalability**.
- Ensures **consistent evaluations.**
- For example, Prompt Flow, Vertex AI Studio, Amazon Bedrock

LLM-as-a-Judge

- **One LLM (the evaluator) analyzes and evaluates the output of another LLM.**
    - May evaluate linguistic qualities, relevance, and adherence to prompts.
- Useful for preliminary assessments, continuous integration, and large-scale evaluations.
- Limitations:
    - Requires significant computational resources.
    - Sensitive to changes in response tokens, potentially missing subtleties like sarcasm or irony.

Online Evaluation

- Once we are confident in our LLM's performance offline, we can **run A/B tests online to gather user data.**
- Leverages authentic user data to assess live performance and user satisfaction.
- Measures both direct and indirect user feedback.
- Ideal for continuous performance monitoring.

Case Study: OpenAI

GPT-4: Overview

- GPT-4 is a large **multimodal model**
- Accepts image and text inputs, emits text outputs
- Exhibits **human-level performance** on various benchmarks

Model paper

GPT-4: Qualitative Evaluations

- **External experts** recruited in August 2022
- Stress testing, boundary testing, and red teaming
  - Structured effort to find flaws and vulnerabilities
  - Iterative process: hypothesis, testing, adjusting
- Experts **from diverse fields** (fairness, cybersecurity, law, etc.)
- **Internal testing**

GPT-4: Quantitative Evaluations

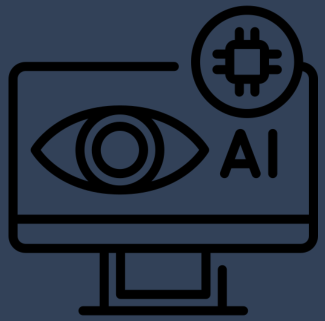
- **Internal evaluations** for categories against content policy
  - Measures likelihood of generating harmful content
- **Automated evaluations** for different model checkpoints

GPT-4: Benchmarking - Text

- **Exams:**
  - Evaluated on professional and academic benchmarks
  - Used public exams and practice tests without specific training
- **Traditional ML:**
  - MMLU, HellaSwag, HumanEval, and TruthfulQA
- **Multi-lingual:**
  - Translated MMLU benchmark into 24 languages

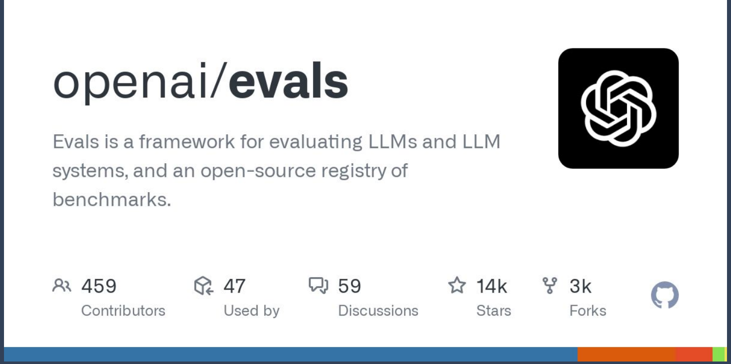
GPT-4: Benchmarking - Vision

- Evaluated on standard academic **vision benchmarks**
- Benchmarks include VQAv2, TextVQA, ChartQA, AI2 Diagram, DocVQA
- Constantly **discovering new tasks** the model can tackle!



GPT-4: OpenAI Evals

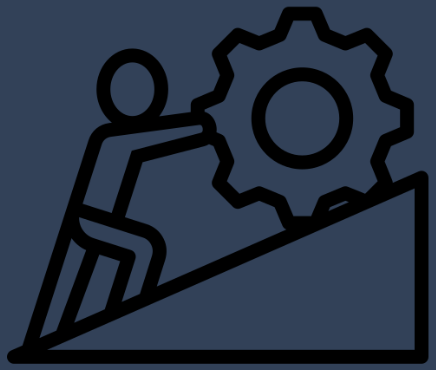
- **Open-source framework** for creating and running benchmarks
- Used for tracking performance and preventing regressions
- Compatible with existing benchmarks



Future Directions and Challenges

Challenges

- **Ethical and bias concerns.**
- **Computational resource** demands.
- **Overfitting** and data contamination.
- Limited **diversity metrics**.
- **Balancing innovation with regulation.**



Overcoming These Challenges

- Leverage **multiple evaluation metrics**.
- Enhance **human evaluation**.
- Incorporate **diverse reference data.**
- Implement **real-world evaluation**.
- Assess **robustness and security**.

Future Directions

- Enhanced **evaluation metrics**.
- Real-time and **adaptive evaluation**.
- **Cross-domain** generalization.

Q&A