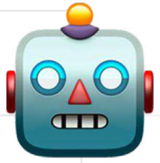


Evaluating LLMs

100x

Why Evaluation?

- Risk Factor
- Consistency in performance
  - There's a new model in town everyday
  - Even the best LLMs hallucinate and make mistakes (bad formatting, too open or too cautious)
  - Just because your LLM works well in a few cases doesn't mean it can be trusted
- Follow similar approaches we take with code (writing tests). Build a testing suite to make sure you cover common failure cases and iterate over it as you go.

Challenges with Evaluation

How do we evaluate Traditional ML models?

- Oriented towards outcomes (what your users care about)
- Has a set of metrics (one or more)
- Automated setup

Traditional ML

pred=[“cat”, “dog”, “dog”, “cat”, “dog”, “cat”, “dog”, “dog”, “cat”, “dog”]
label=[“dog”, “dog”, “dog”, “cat”, “dog”, “cat”, “dog”, “dog”, “cat”, “dog”]

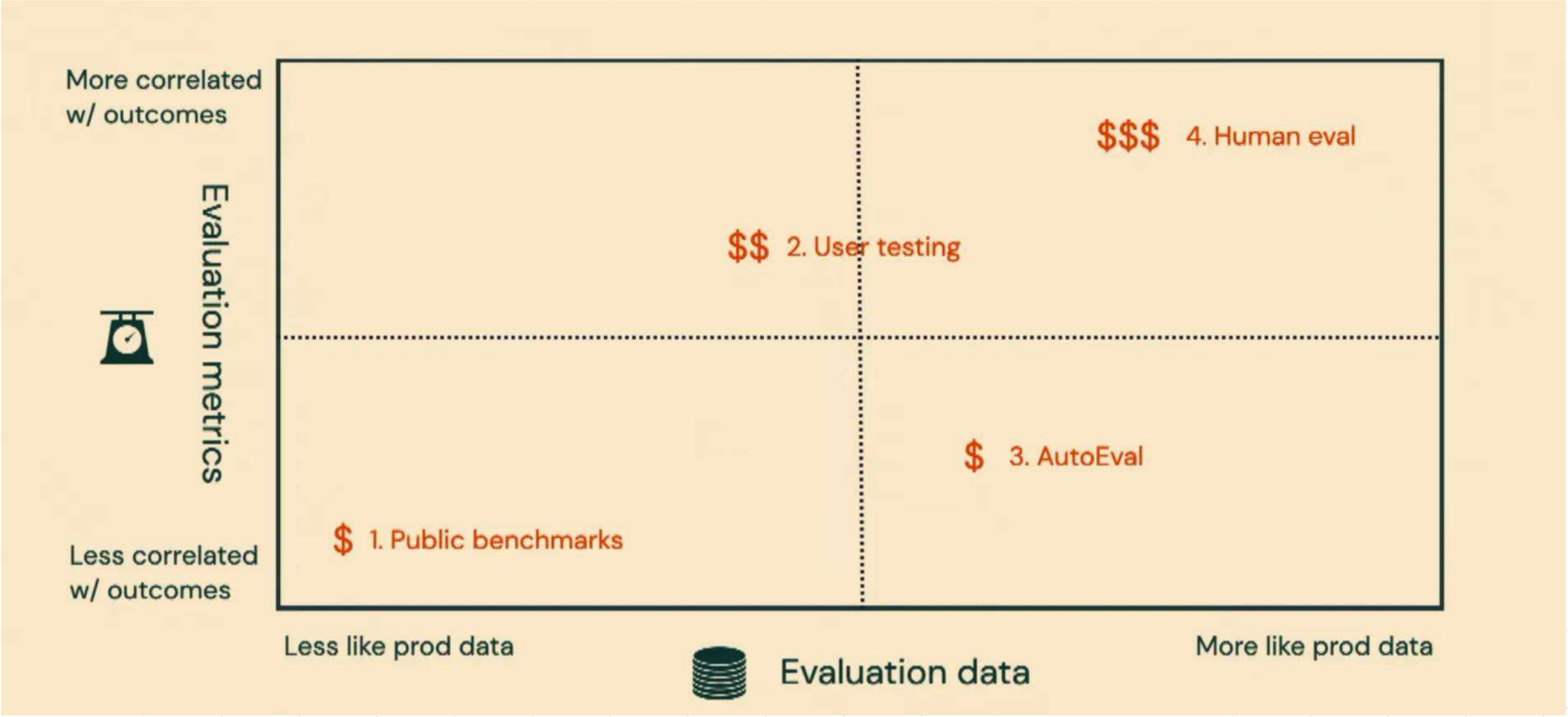
Generative

pred=[“this is an image of a tabby cat”]
label=[“photograph of a cat”]

Evaluation Methods

100x

Evaluation Methods



Public Benchmarks

- HumanEval - Coding HellaSwag - Common sense and reasoning MMLU - Multiple subjects, problem solving TruthfulQA - Q & A across multiple categories ARC (AI2 Reasoning Challenge) - MCQs on scientific reasoning
- Chatbot Arena - Evaluating through human preference
- Example: Meta Llama 3 Benchmarks

Why benchmarks don’t work for you?

- ❌ They don’t evaluate the model on your use case
- ❌ They don’t take into account your way of prompting
- ❌ They are not updated very often

Problems with using LLMs for evaluation

- They are biased, prefer their own outputs
- If asked to score, they might pick a favourite number
- Might prefer longer answers

Problems with using Humans for evaluation

- Cost and Speed
- GPT4/Claude is more accurate than random humans in many cases
- Human evaluation might not give attention to detail, might prefer style over factuality
- People are not consistent when asked to score, different meaning of score 5

Recipe for Evaluating LLMs

1. Build a handful of examples to start with
2. Use LLM to generate test cases
    a. auto-evaluator
    b. lm-evaluation-harness
3. Add more data based on feedback and for edge cases

Metrics for LLM Eval

- Semantic similarity (accurate & factually consistent)
- Structural consistency
- Give your criteria & grade it

Human-verified LLM Evaluation