Benchmarking and evaluation 1

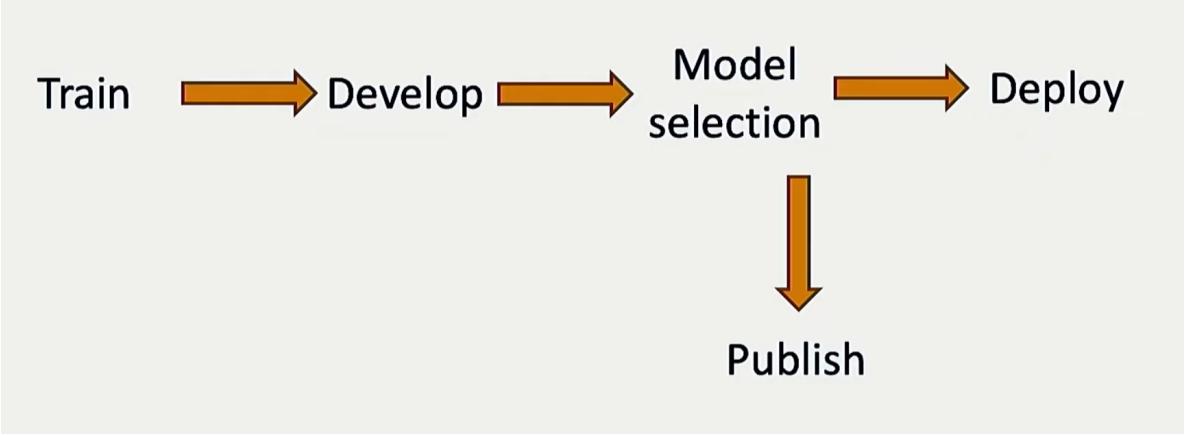
Apr 9, 2026

*Acknowledgment: Slides based on materials by CS224N @ Stanford University (Lecture by Yann Dubois).

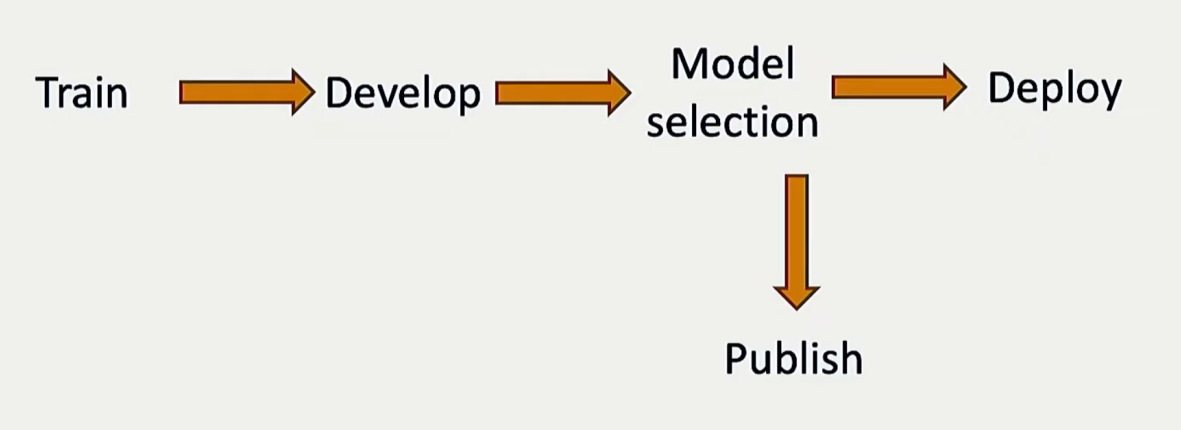
Outline

Outline

Different desiderata for measuring performance

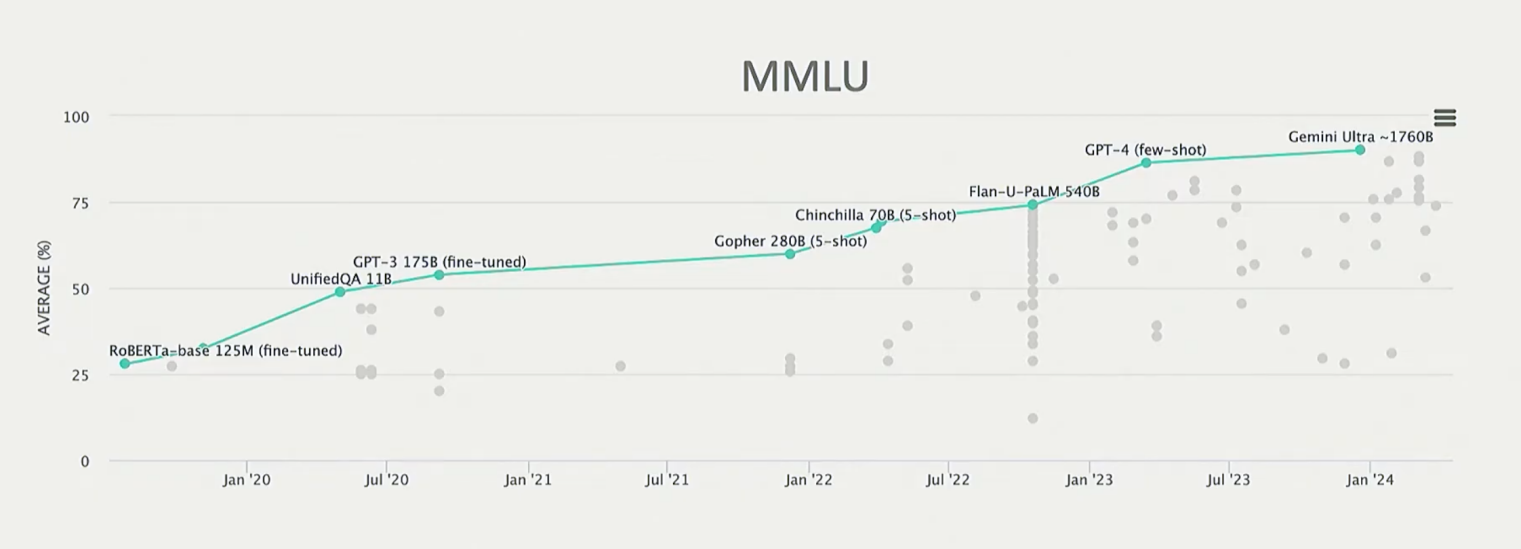


Different desiderata for measuring performance



- Deployment / product-oriented research: task-specific, reliability-critical, requiring high trust
- Publication-oriented research: prioritizes reproducibility; simpler or approximate metrics may be acceptable

Benchmarks and evaluations drive progress



- **MMLU** provide a standardized way to track model progress; MMLU?
- Recall Hendrycks et al. (2021)
- Widely used in academia as a reference point for comparison; Small performance differences are often less meaningful in isolation
- The key focus is on overall progress and relative improvement over time
- `https://artificialanalysis.ai/evaluations/mmlu-pro`

Two major types of evaluations

- Close-ended evaluations
- Open-ended evaluations

Two major types of evaluations

- Close-ended evaluations
- Open-ended evaluations

Example

Text: Read the book, forget the movie!

Label: Negative

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Outline

Close-ended tasks

- Limited number of potential answers
- Often one or just a few correct answers
- Enables automatic evaluation as in ML

Closed-ended tasks (Benchmarks)

- Sentiment analysis (SST, IMDB, Yelp ...)
  - *Text: Read the book, forget the movie!*
  - *Label: Negative*
- Entailment (SNLI)
  - *Text: A soccer game with multiple males playing*
  - *Hypothesis: Some men are playing sport.*
  - *Label: Entailment*
- NER (CoNLL-2003)
- Part-of-Speech (PTB)

Close-ended tasks (more examples)

Example

Text: Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.

Coreference: False

Example

Endangered Species Act Paragraph: "... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a **1937 treaty** prohibiting the hunting ofright and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little **opposition** was raised."

Question 1: "Which laws faced significant **opposition**?"
Plausible Answer: later laws

- Don't forget the metrics that we use in the standard ML classes: accuracy, recall, precision, F1, ROC!
- Look at multi-task benchmarks concurrently.

Close-end multi-task benchmark - superGLUE

Leaderboard Version: **2.0**

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	Inspur Cloud	Hairuo		91.4	92.5	96.5/97.6	100.0	90.5/67.9	94.1/93.2	92.8	76.1	100.0	64.6	96.1/94.7
2	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
3	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
4	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
5	ERNIE Team - Baidu	ERNIE 3.0	ERNIE 3.0	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
6	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
7	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
8	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
9	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
10	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
11	SPoT Team - Google	Frozen T5 1.1 + SPoT		89.2	91.1	95.8/97.6	95.6	87.9/61.9	93.3/92.4	92.9	75.8	93.8	66.9	83.1/82.6
12	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4

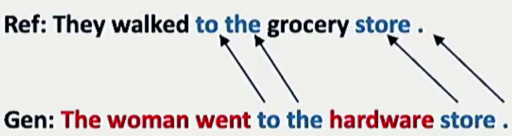
- `https://super.gluebenchmark.com/leaderboard`
- Attempt to measure "general language capabilities" (what GLUE stands for)
- Cover a number of different tasks

Outline

Open-ended tasks

- Long generations with too many possible correct answers to enumerate
  - $\rightarrow$ cannot use standard ML metrics
- There are now better and worse answers (not just right and wrong)
- Example tasks (Benchmarks):
  - Summarization (CNN-DM, Gigaword)
  - Translation (WMT)
  - Instruction-following (Chatbot Arena, AlpacaEval, MT-bench)
    - evaluates how well models follow user instructions and produce helpful, aligned responses, often using human or LLM-based preference judgments
    - how do we evaluate ChatGPT?

Types of evaluation methods for text generation

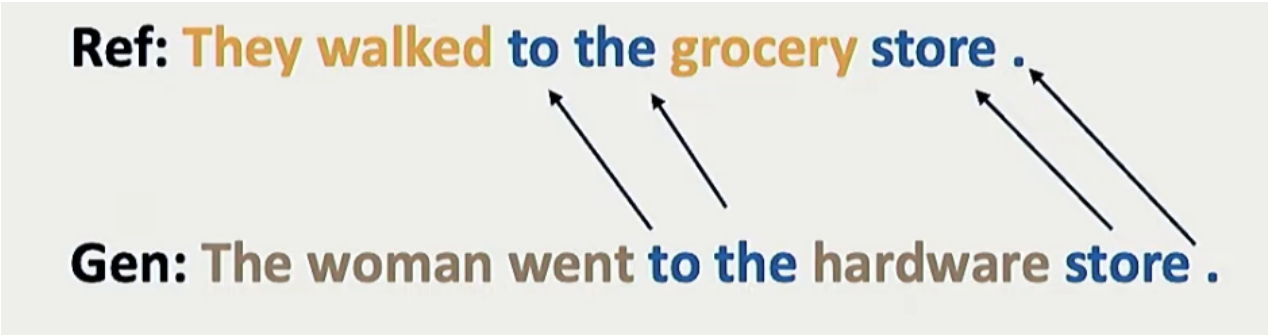


Content Overlap Metrics

Model-based Metrics

Human Evaluations

Content overlap metrics



- Compute a score that indicates the lexical similarity between *generated* and *gold-standard (human-written) text*
- Fast, efficient
- N-gram overlap metrics (e.g., BLEU [precision], ROUGE [recall], METEOR, CIDEr, etc.)
- Not ideal but often still reported for translation and summarization
- Semantic representation? BERTscore

Reference free evals

- Reference-based evaluation:
  - Compare human written reference to model outputs
  - Used to be "standard" evaluation for most NLP tasks
  - e.g., BLUE, ROUGE, BERTscore
- Reference free evaluation:
  - Have a model give a score
  - No human reference
  - was nonstandard - now becoming popular with GPT4
  - AlpacaEval, MT-Bench

Human evaluations

- Automatic metrics fall short of matching human decisions
- Human evaluation is most important form of evaluation for text generation
- Gold standard in developing new automatic metrics
  - New automated metrics must correlate well with human evaluations

Human evaluations

- Ask humans to evaluate the quality of generated text
- Overall or along some specific dimension:
  - fluency
  - coherence
  - consistency
  - factuality, correctness
  - commonsense
  - style, formality
  - grammaticality
  - redundancy

Human evaluations: Issues

- Human judgments are regarded as the gold standard
- But it also has issues:
  - slow
  - expensive
  - inter-annotator disagreement (subjective)
  - intra-annotator disagreement across time
  - not reproducible

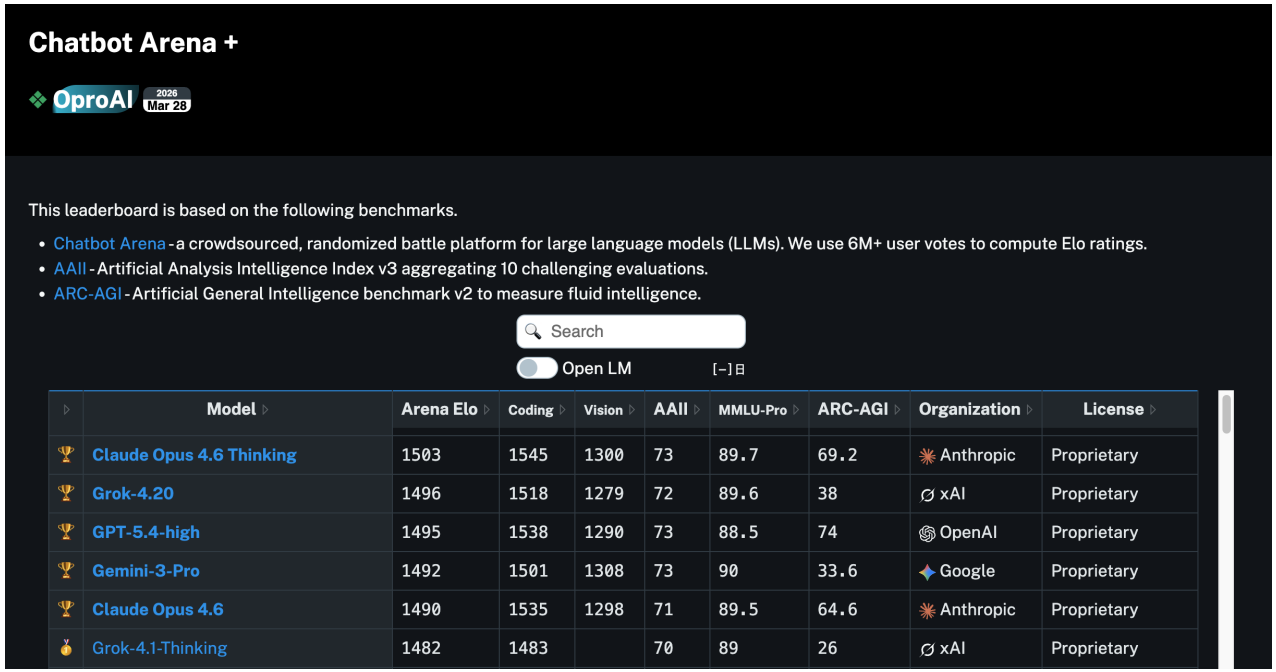
Human evaluations: Issues

- Challenges with human evaluation
  - how to describe the task?
  - how to show the task to the humans?
  - selecting the annotators?
  - monitor the annotators: time, accuracy, ...

Reference-free eval: chatbots

- How do we evaluate something like ChatGPT?
- So many different use cases - hard to evaluate (e.g., generation, Open/closed QA, Brainstorming, Chat, rewrite, summarization, classification, extract)
- The responses are also long-form text, which is even harder to evaluate

Chatbot Arena+



Chatbot Arena +

OproAI 2026 Mar 28

This leaderboard is based on the following benchmarks.

- Chatbot Arena - a crowdsourced, randomized battle platform for large language models (LLMs). We use 6M+ user votes to compute Elo ratings.
- AAII - Artificial Analysis Intelligence Index v3 aggregating 10 challenging evaluations.
- ARC-AGI - Artificial General Intelligence benchmark v2 to measure fluid intelligence.

Search

Open LM

	Model	Arena Elo	Coding	Vision	AAII	MMLU-Pro	ARC-AGI	Organization	License
🏆	Claude Opus 4.6 Thinking	1503	1545	1300	73	89.7	69.2	Anthropic	Proprietary
🏆	Grok-4.20	1496	1518	1279	72	89.6	38	xAI	Proprietary
🏆	GPT-5.4-high	1495	1538	1290	73	88.5	74	OpenAI	Proprietary
🏆	Gemini-3-Pro	1492	1501	1308	73	90	33.6	Google	Proprietary
🏆	Claude Opus 4.6	1490	1535	1298	71	89.5	64.6	Anthropic	Proprietary
🥇	Grok-4.1-Thinking	1482	1483		70	89	26	xAI	Proprietary

https://openlm.ai/chatbot-arena/

Outline

Presentations

- Sanjeev: Yao et al. (2023). ReAct: Reasoning and Acting in LLMs.
- Dmitrii: Shick et al. (2023). Toolformer: Teaching LLMs to Use Tools.