UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

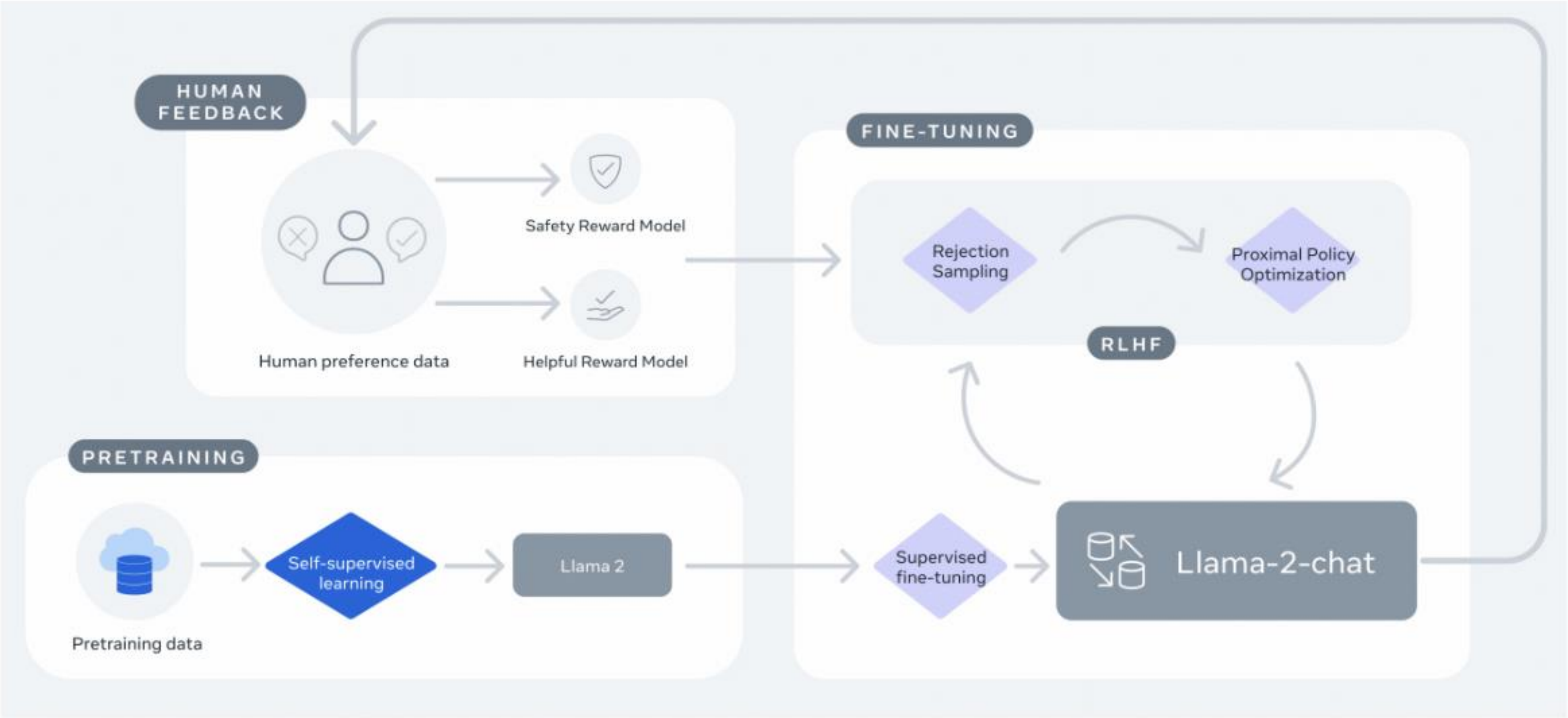
Artificial Intelligence
Lecture12 – Evaluation

Agenda

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

1. Evaluation of LLMs

Overview of LLMs Training

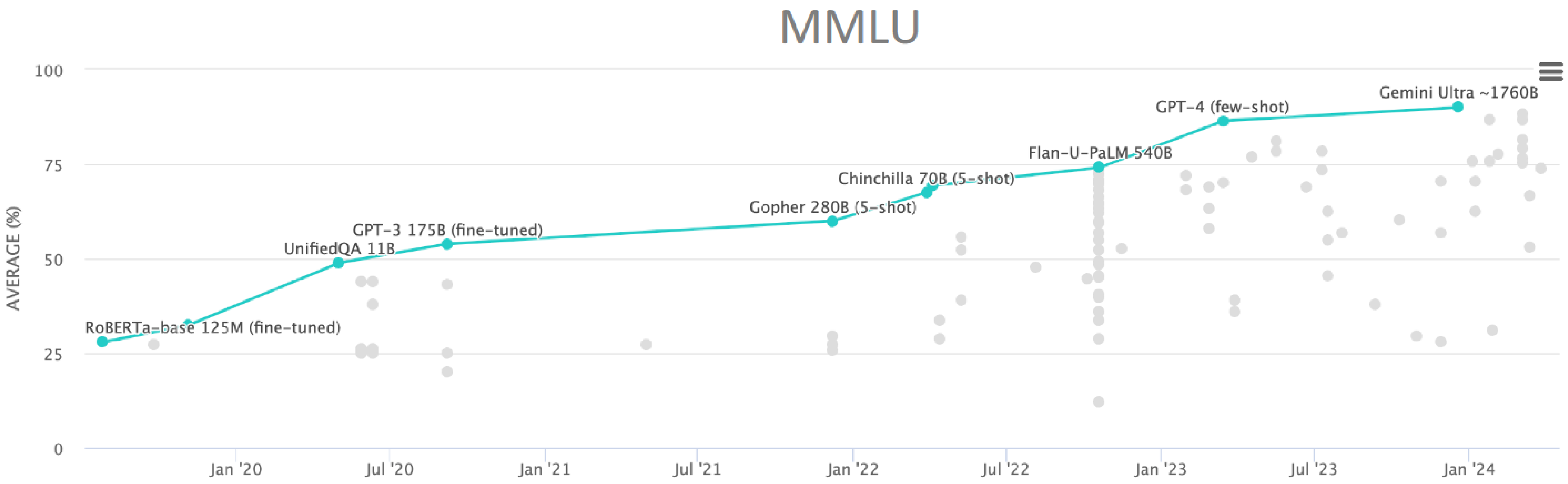


Benchmarks and evalua:ons drive progress

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

MMLU (Massive Multitask Language Understanding)



Two major types of evaluations

Close-ended evaluations

Example

Text: Read the book, forget the movie!

Label: Negative

Open ended evaluations

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.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Close-ended mul:-task benchmark - superGLUE

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

https://super.gluebenchmark.com/

SuperGLUE GLUE

Leaderboard Version: **2.0**

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	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
+	2	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
	3	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
	4	ERNIE Team - Baidu	ERNIE 3.0		90.6	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
	5	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
+	6	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
+	7	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
	8	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
+	9	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

Attempt to measure “general language capabilites”

Close-ended mul:-task benchmark - superGLUE

Cover a number of different tasks

- BoolQ, Mul;RC (reading texts)
- CB, RTE (Entailment)
- COPA (cause and effect)
- ReCoRD (QA+reasoning)
- WiC (meaning of words)
- WSC (coreference)

BoolQ	**Passage:** *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.* **Question:** *is barq's root beer a pepsi product* **Answer:** No
CB	**Text:** *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?* **Hypothesis:** *they are setting a trend* **Entailment:** Unknown
COPA	**Premise:** *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?* **Alternative 1:** *The sun was rising.* **Alternative 2:** *The grass was cut.* **Correct Alternative:** 1
MultiRC	**Paragraph:** *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week* **Question:** *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (T), *No* (F), *Yes* (T), *No, she didn't recover* (F), *Yes, she was at Susan's party* (T)
ReCoRD	**Paragraph:** *(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electoral Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood* **Query** For one, they can truthfully say, "Don't blame me, I didn't vote for them, " when discussing the <placeholder> presidency **Correct Entities:** US
RTE	**Text:** *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.* **Hypothesis:** *Christopher Reeve had an accident.* **Entailment:** False
WiC	**Context 1:** *Room and board.* **Context 2:** *He nailed boards across the windows.* **Sense match:** False
WSC	**Text:** *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.* **Coreference:** False

Close-ended: challenges

Choosing your metrics: accuracy / precision / recall / f1-score / ROC

https://github.com/cgpotts/cs224u/blob/main/evaluation_metrics.ipynb

Aggregating across metrics or tasks

- Where do the labels come from?
- Are there spurious correlations?

SuperGLUE Tasks

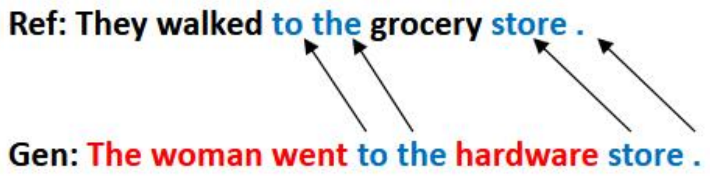
Matthew's Corr	F1a / EM	F1 / Accuracy
Avg. F1 / Accuracy	Accuracy	
Accuracy	Accuracy	Gender Parity / Accuracy

Open-ended tasks

Long generations with too many possible correct answers to enumerate

- => can’t use standard ML metrics
- There are now better and worse answers (not just right and wrong)
- Example:
- Summarization: CNN-DM / Gigaword
- Translation: WMT
- Instruction-following: Chatbot Arena / AlpacaEval / MT-Bench

Types of evalua:on methods for text generation



Content Overlap Metrics

Model-based Metrics

Human Evaluations

Content Overlap Metrics

Ref: They walked to the grocery store .

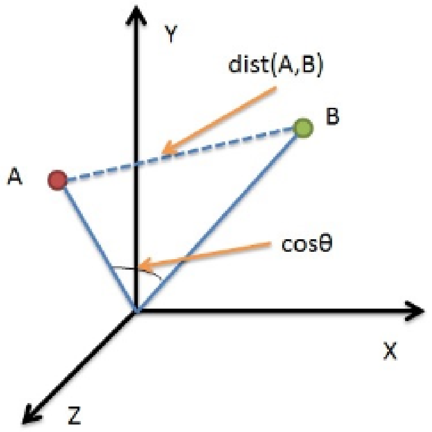
Gen: The woman went to the hardware store .

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

Model-based metrics to capture more semanantics

- Use learned representations of words and sentences to compute semantic similarity between generated and reference texts

- The embeddings are pretrained, distance metrics used to measure the similarity can be fixed



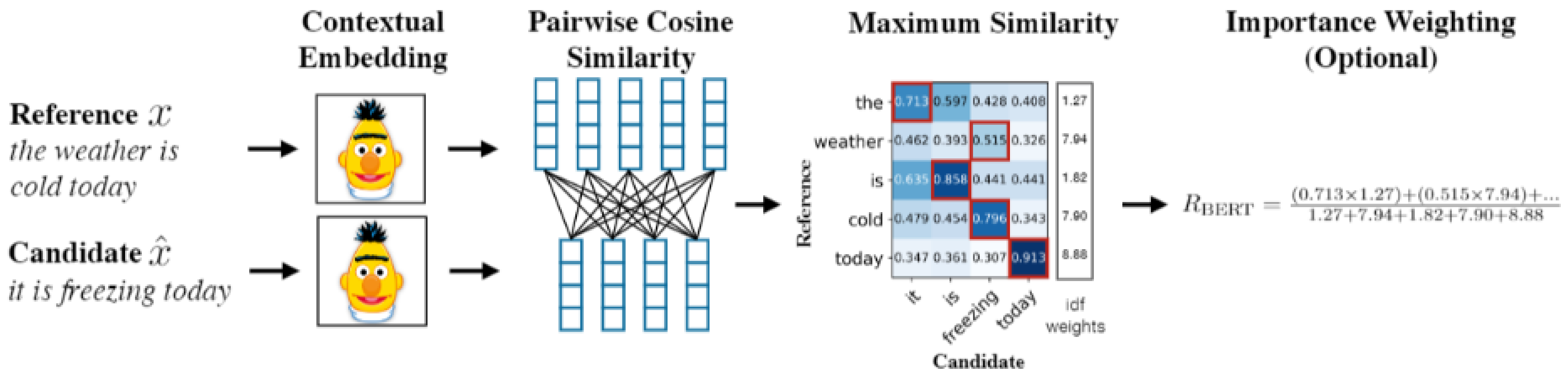
Vector Similarity

Embedding based similarity for semantic distance between text.

- **Embedding Average** (Liu et al., 2016)
- **Vector Extrema** (Liu et al., 2016)
- **MEANT** (Lo, 2017)
- **YISI** (Lo, 2019)

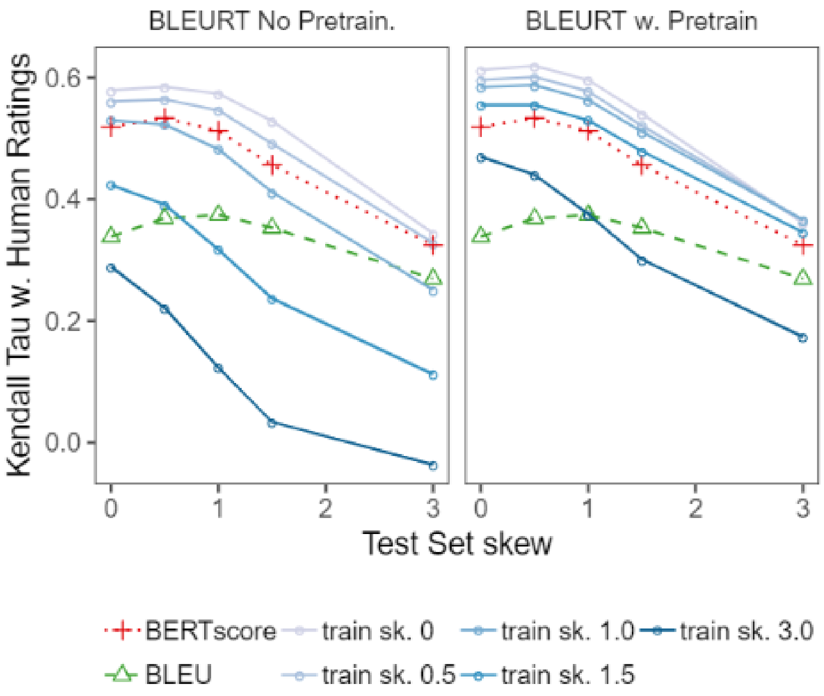
BERTSCORE

Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. (Zhang et.al. 2020)



Model-based metrics: Beyond word matching

BLEURT:
A regression model based on BERT returns a score that indicates to what extent the candidate text is gramatical and conveys the meaning of the reference text.
(Sellam et.al. 2020)



Reference free evals

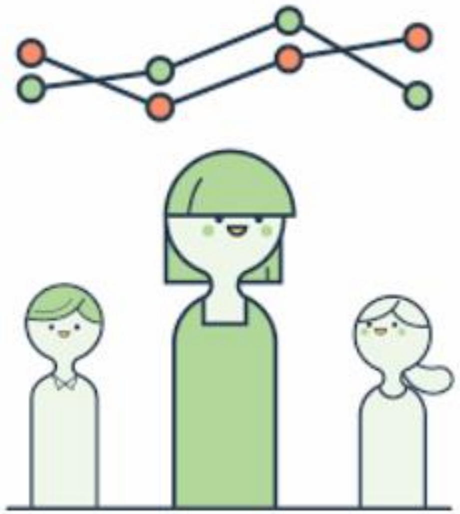
Reference-based evaluation:

- Compare human written reference to model outputs
- Used to be ‘standard’ evalua;on for most NLP tasks
- Examples: BLEU, ROUGE, BertScore etc.

Reference free evaluaEon

- Have a model give a score
- No human reference
- Was nonstandard – now becoming popular with GPT4
- Examples: 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 and correctness
- commonsense
- style / formality
- gramma;cality
- redundancy

Note: Don't compare human evaluation scores across differently conducted studies Even if they claim to evaluate the same dimensions!

Human Evaluations: Issues

Human judgments are regarded as the gold standard, but it also has issues:

- Slow
- Expensive
- Inter-annotator disagreement (esp. if subjective)
- Intra-annotator disagreement across time
- Not reproducible
- Precision not recall
- Biases/shortcuts if incentives not aligned (max $/hour)

“just 5% of human evaluations are repeatable in the sense that (i) there are no prohibitive barriers to repeat on, and (ii) sufficient information about experimental design is publicly available for rerunning them. Our estimate goes up to about 20% when author help is sought.”

Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP

Anya Belz[a,b] **Craig Thomson**[b] **Ehud Reiter**[b] **Simon Mille**[a]

[a]ADAPT, Dublin City University
Dublin, Ireland
{anya.belz,simon.mille}@adaptcentre.ie

[b]University of Aberdeen
Aberdeen, UK
{c.thomson,e.reiter}@abdn.ac.uk

Human Evaluations: Issues

Challenges with human evaluation

- How to describe the task?
- How to show the task to the humans?
- What metric do you use?
- Selecting the annotators
- Monitoring the annotators: time, accuracy, …

Reference-free eval: Chatbots

VS

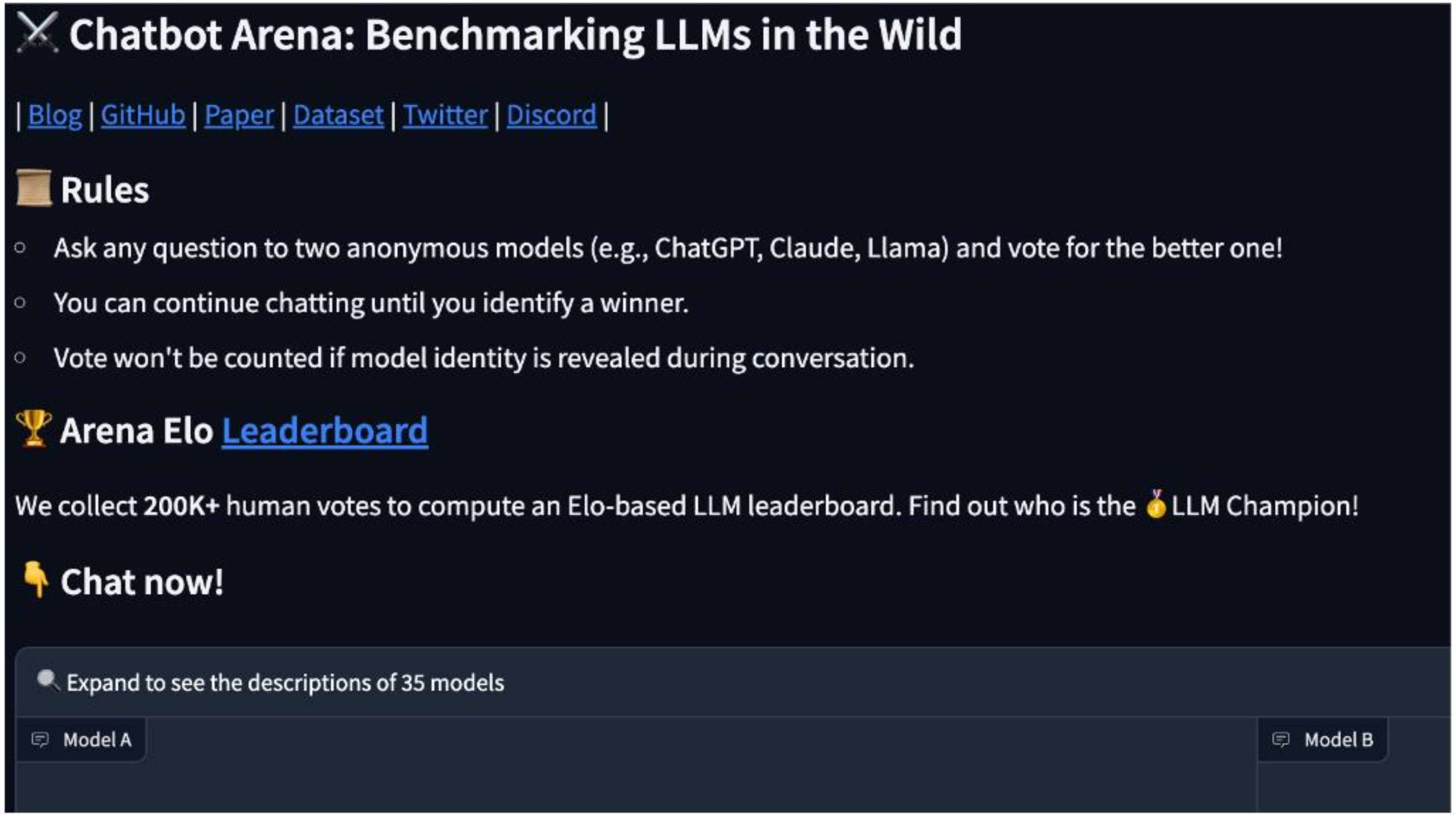
Table 1: Distribution of use case categories from our API prompt dataset.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

How do we evaluate something like ChatGPT?

- So many different use cases it’s hard to evaluate
- The responses are also long-form text, which is even harder to evaluate.

Side-by-side ratings



Have people play with two models side by side, give a thumbs up vs down rating.

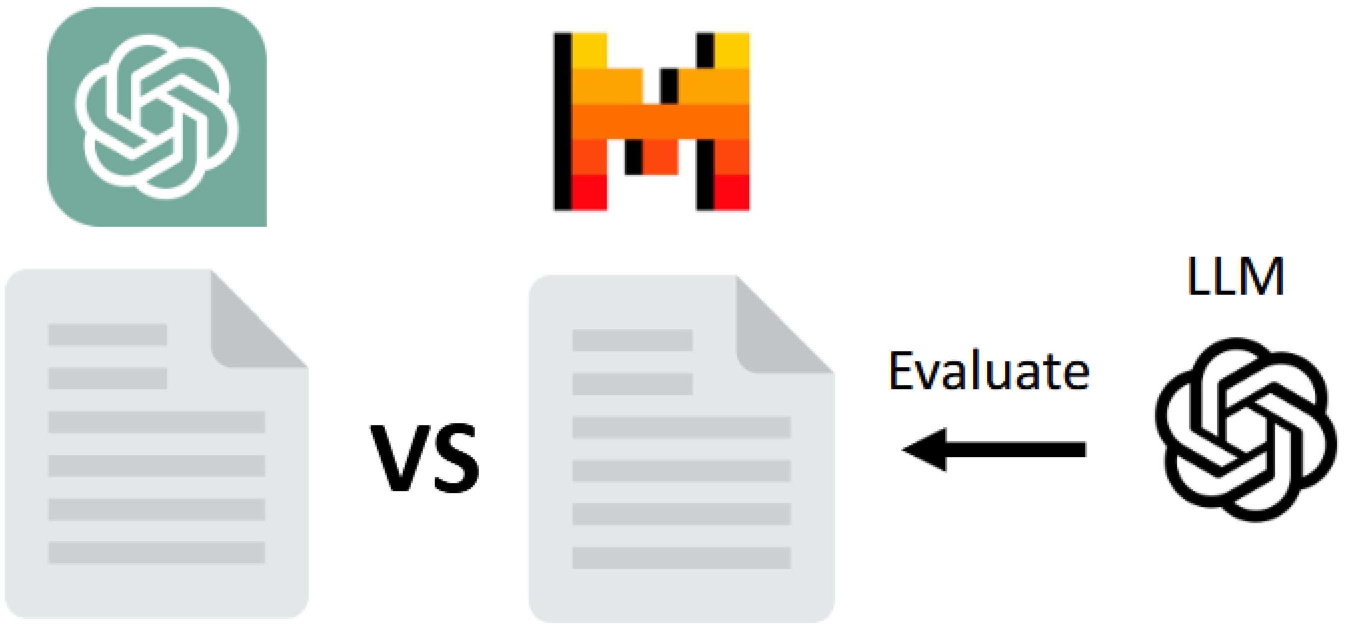
What is missing with Side-by-side ratings

Current gold standard for evaluation of chat LLM

- External validity
- Typing random questions into a head-to-head website may not be representative
- Cost
- Human annotation takes large, community effort
- New models take a long time to benchmark
- Only notable models get benchmarked

Lowering the cost – use a LM evaluator

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería



Use a LM as a reference free evaluator

- Surprisingly high correla;ons with human
- Common versions: AlpacaEval, MT-bench

AlpacaEval

AlpacaEval

- Internal benchmark for developing Alpaca
- 98% correlation with Chatbot Arena
- < 3 min and < $10

- 1. For each instruction: generate an output by baseline and model to eval
- 2. Ask GPT-4 the probability that the model's output is better
- 3. (AlpacaEval LC) Reweight win-probability based on length of outputs
- 4. Average win-probability => win rate

AlpacaEval Leaderboard

Model Name	LC Win Rate	Win Rate
GPT-4 Turbo (04/09)	55.0%	46.1%
GPT-4 Preview (11/06)	50.0%	50.0%
Claude 3 Opus (02/29)	40.5%	29.1%
GPT-4	38.1%	23.6%

TakeAways

Closed ended tasks
- Think about what you evaluate (diversity, difficulty)

Open ended tasks
- Content overlap metrics (useful for low-diversity settings)
- Chatbot evals – very difficult! Open problem to select the right examples / eval

Challenges
- Consistency (hard to know if we're evalua&ng the right thing)
- Contamina&on (can we trust the numbers?)
- Biases

In many cases, the best judge of output quality is YOU!
- Look at your model generations. Don't just rely on numbers!

Readings

https://arxiv.org/html/2412.05579v1

https://www.datacamp.com/blog/llm-evaluation

https://arxiv.org/abs/2404.18796

https://huggingface.co/papers/2404.18796