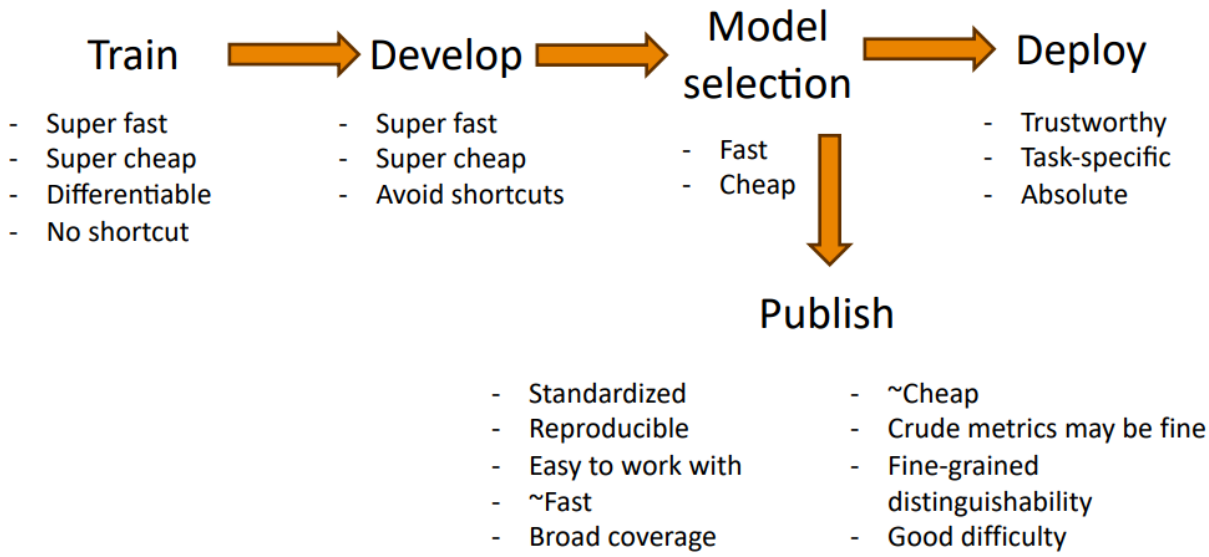
Performance measurement

Different desiderata for measuring performance



Hyperparameter

What's it?

Configuration varibles manually set (or selected through algorithm) by researchers or engineers before the model starts learning. They can't be directly learned from data, but are used to control the entire learning process itself

What kind of hyperparameter may we meet?

1. learning rate 2. batch size 3. number of training cycles
4. optimizer hyperparameter 5. network structure related hyperparameters 6. regularization strength

① Two major types of evaluations

- Close-ended evaluations: we can think about classification where we know exactly the correct label for the task
- Open-ended evaluations: think about ChatGPT, there's no correct answer

① Close-ended evaluations

[1] characteristic
- Limited number of potential answers
- often one or just a few correct answers
- Enable automatic evaluation as in ML

[2] tasks

Close-ended tasks

- Sentiment analysis: SST / IMDB / Yelp ...

  Example
  Text: Read the book, forget the movie!
  Label: Negative

- Entailment: SNLI

  Example
  Text: A soccer game with multiple males playing.
  Hypothesis: Some men are playing sport.
  Label: Entailment

- Name entity recognition: CoNLL-2003
- Part-of-Speech: PTB

Close-ended tasks

- Coreference resolution: WSC

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

- Question Answering: Squad 2

  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
  Question 2: "What was the name ofthe 1937 treaty?"
  Plausible Answer: Bald Eagle Protection Act

[3] superGLUE

Close-ended multi-task benchmark - superGLUE

SuperGLUE GLUE Leaderboard Version: **2.0**

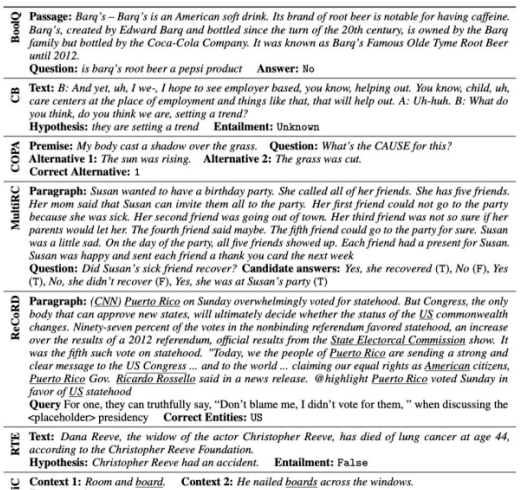
	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
+	2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	3	Microsoft Alexander 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 capabilities"

Examples from superGLUE

Cover a number of different tasks

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



WSC	Sense match: False	
	Text: *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.* Coreference: False	

The BoolQ, MultiRC, CB ... here are some datasets which are used to evaluate models

[4] Challenge

1. Choosing metrics : accuracy / precision / recall / f1-score / ROC

   Because if we choose a inappropriate metric may lead to misjudgement of model perform

2. Aggregating across metrics or tasks

   Numerical range and significance of F1-score, accuracy are different

3. where do the labels come from

4. are there spurious correlations

[5] Spurious correlation

Spurious correlation

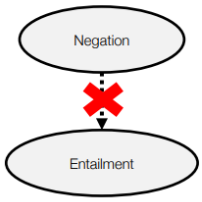
Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.

Premise:

The economy could be still better.

Hypothesis:

The economy has never been better



[Gururangan+ 2019]

SNLI itself is hard, but there can be undiscovered *spurious correlations*

the goal of the model is to minimize the loss function, rather than truly 'understanding' the language. Therefore, it will look for any shortcuts in the data that can make predictions, which are often spurious correlations

eg. here

people often use negative expressions, then model will predict no entailment when it saw the 'never' or other negative expressions

2 Open-ended tasks

1 characteristic / example

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

↳ is kind of the mother of all tasks, which is also very hard to evaluation

2 Type of evaluation methods of text generation

- Context Overlap metrics
- model-based metrics
- Human evaluation

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.)

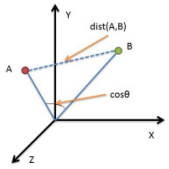
  precision recall
- Not ideal but often still reported for translation and summarization

n-gram overlap metrics have no concept of semantic relatedness

2. model-based metrics to capture more semantic

- 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**

Model-based metrics: Word distance functions



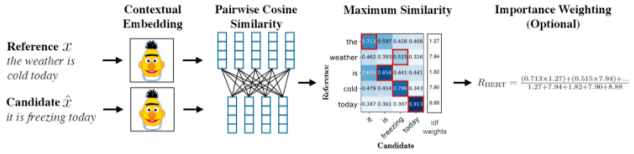
Vector Similarity

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)



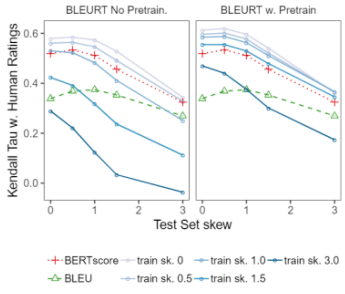
It utilizes the powful understanding ability of large-scale LM such as BERT for contextual semantics, and can identify semantically related phrases

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 grammatical and conveys the meaning of the reference text.

(Sellam et.al. 2020)



It take a pretrained BERT and do some continual pretraining by trying to predict the BLEU score and some other metrics, then finetune it.

important part: finetuned pretraing model actually do the evaluation.

An important failure case:

When the quality of the reference text is poor, automatic evaluation indicaters based on the reference (such as ROUGE) will become invalid

Reference-based measures *are only as good as their references.*

lead to

Reference free evals

- **Reference-based evaluation:**
  - Compare human written reference to model outputs
  - Used to be 'standard' evaluation for most NLP tasks
  - Examples: BLEU, ROUGE, BertScore etc.
- **Reference free evaluation**
  - Have a model give a score
  - No human reference
  - Was nonstandard – now becoming popular with GPT4
  - Examples: AlpacaEval, MT-Bench

3.

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

Note: Don't compare human evaluation scores across

- commonsense
- style / formality
- grammaticality
- redundancy

differently conducted studies

Even if they claim to evaluate the same dimensions!

People usually do also when they have human's annotations is that they have metrics for computing the inter-annotator, basically, they try to achieve a certain inter-annotator agreement.

Human evaluation: 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)

→ workers may tend to find shortcut to maximize ther salary.

Human evaluation: 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, …

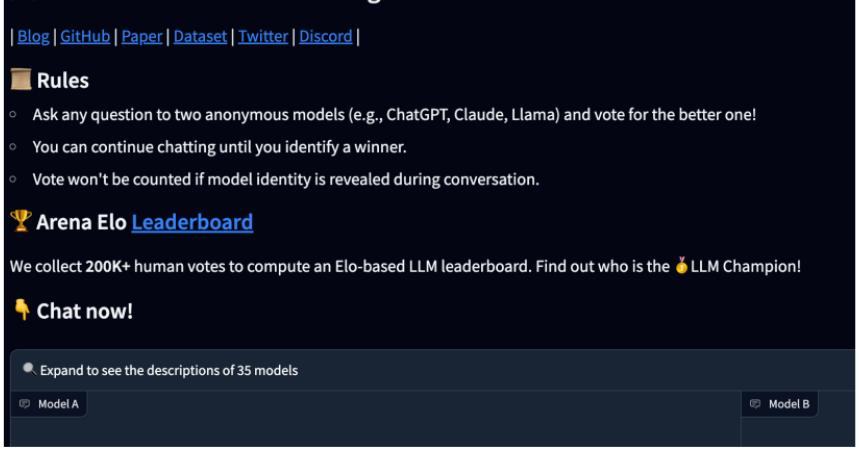
Reference-free eval: chatbots

- 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

Chatbot Arena: Benchmarking LLMs in the Wild



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

Chatbot arena is used as gold standard

What's missing with side-by-side human eval?

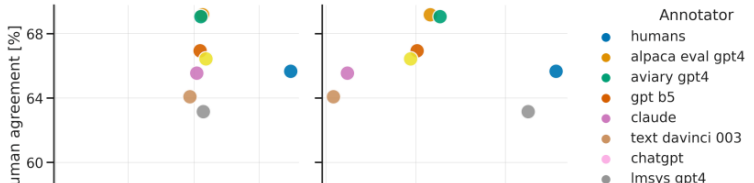
- Current gold standard for evaluation of chat LLM
- **External validity**
  - Typing random questions into a head-to-head website may not be representative
- **Cost**
  - Human annotation takes large, community effort
  - New models take a long time to benchmark
  - Only notable models get benchmarked

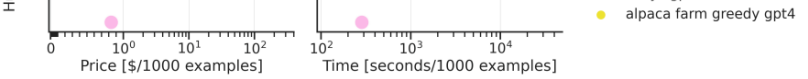
Lowering the cost - use LM evaluator

- Use a LM as a reference free evaluator
- Surprisingly high correlations with human
- Common versions: AlpacaEval, MT-bench

AlpacaFarm

AlpacaFarm : Human agreement

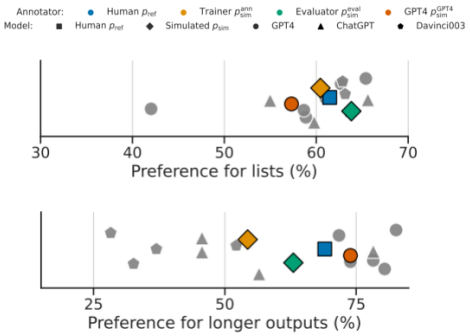




- 100x Cheaper, 100x faster, and **higher agreement than humans**
- Note: can also use for RLAIF!

Humans have low agreement because of variance

Things to be careful with



- Same issues as before: Spurious correlations!
  - Length
  - Position (but everyone randomizes this away)
  - GPT-4 self bias

Alpaca Eval

AlpacaEval

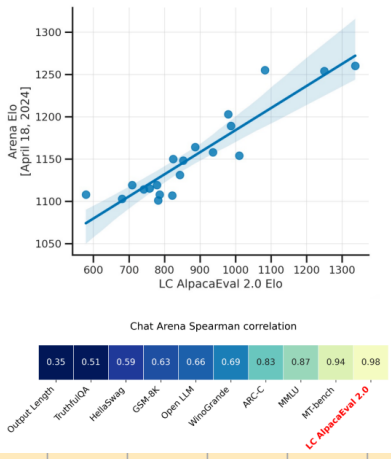
- Internal benchmark for developing Alpaca
- 98% correlation with Chatbot Arena
- < 3 min and < $10

- 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%

AlpacaEval : System level correlation



Length bias / self-bias

AlpacaEval Length Controlled

- Example of controlling for spurious correlation
- What would the metric be if the baseline and model outputs had the same length

	AlpacaEval			Length-controlled AlpacaEval		
	concise	standard	verbose	concise	standard	verbose
gpt4_1106_preview	22.9	50.0	64.3	41.9	50.0	51.6
Mixtral-8x7B-Instruct-v0.1	13.7	18.3	24.6	23.0	23.7	23.2
gpt4_0613	9.4	15.8	23.2	21.6	30.2	33.8
claude-2.1	9.2	15.7	24.4	18.2	25.3	30.3
gpt-3.5-turbo-1106	7.4	9.2	12.8	15.8	19.3	22.0
alpaca-7b	2.0	2.6	2.9	4.5	5.9	6.8

Self-bias

- The annotator is biased to its outputs, but suprisingly not by much!

	Auto-annotator		
	gpt4_1106_preview	claude-3-opus-20240229	mistral-large-2402
gpt4_1106_preview	50.0	50.0	50.0
claude-3-opus-20240229	40.4	43.3	47.5
mistral-large-2402	32.7	28.2	45.5
gpt4_0613	30.2	20.5	34.3
gpt-3.5-turbo-1106	19.3	16.7	28.9

Figure 7: Length-controlled win rate has the best Arena Correlation and gameability from considered methods, while still being relatively robust to adversarial attacks.

② Current evaluation of LLM

- Perplexity → Pretraining
- Everything → finetuned
- Arena-like → finetuned

① everything

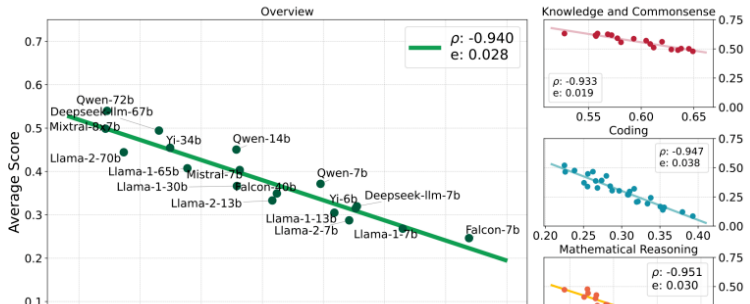
two evaluation ranking of LLMs:
- Holistic evaluation of LM (HELM)
- Huggingface open LLM leaderboard (open-llm leaderboard)

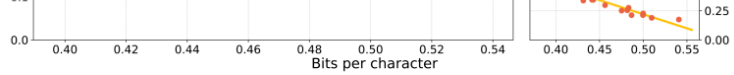
Common LM datasets: NarrativeQA, NaturalQuestions (closed/open-book), OpenbookQA, MMLU (noted at lecture 10), GSM8K, MATH

Other capabilities:
- code
- agent: LMs often get used for more than text — sometimes for things like actuating agents
  - ↓
  - Challenge: evaluation need to be done in sandbox environments
- ...

② perplexity

Perplexity





Perplexity is highly correlated with downstream performance

But depends on data & tokenizer

don't compare the perplexity with different datasets!

↳ why?

↳ the easy answer is that the vocabulary changes

(3) Arena-like

Arena-like

Rank* (UB)	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledg Cutoff
1	GPT-4-Turbo-2024-04-09	1259	+4/-3	35931	OpenAI	Proprietary	2023/12
2	GPT-4-1106-preview	1253	+2/-3	73547	OpenAI	Proprietary	2023/4
2	Claude 3 Opus	1251	+3/-3	80997	Anthropic	Proprietary	2023/8
2	Gemini 1.5 Pro API-0409-Preview	1250	+3/-3	39482	Google	Proprietary	2023/11
2	GPT-4-0125-preview	1247	+3/-2	67354	OpenAI	Proprietary	2023/12
6	Llama-3-70b-Instruct	1210	+3/-4	53404	Meta	Llama 3 Community	2023/12

Let users decide!

(3) Issues and challenges with evaluation

⚠ consistency issue

eg. if just change A.B.C.D to random symbols, the generations will be different also the ranking.

=> it will be very dependant on exactly how you format those choices

eg. MMLU has many implementations { different prompts
different generations
{ most likely valid choice
Probality of gen. answer
most likely choice
...

=> this has led to the loss of fair comparsion

② Contamination issue / Overfitting issue

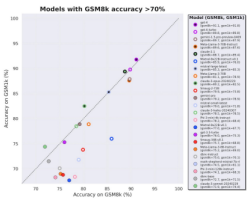
Closed source model + pretraining: hard to know that benmark are truly 'new'

⇒ we don't know if the model has already been trained on a certain dataset

Alleviating overfitting

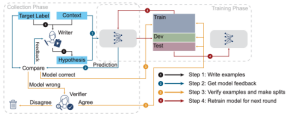
Private test set

- Control the number of times one can see the test set



Dynamic test set

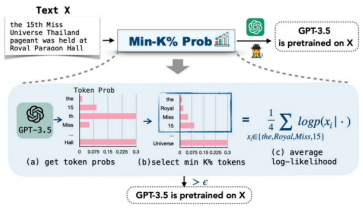
- Constantly change the inputs



Dyna Bench

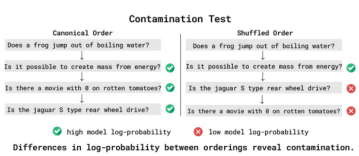
Alleviating contamination: detectors

Min-k-prob



- Detect if models trained on a benchmark by checking if probabilities are 'too high' (what is too high?). Often heuristic.

Exchangeability test



- Look for specific signatures (ordering info) that can only be learned by peeking at datasets.

③ mono culture of NLP benchmarking

Most papers only evaluate on English and performance (accuracy)

⇓ recommand using these dataset

Multi-lingual benchmarking

- Benchmarks exist, we should use them!
- MEGA: Multilingual Evaluation of Generative AI
  - 16 datasets, 70 languages
- GlobalBench:
  - 966 datasets in 190 languages.
- XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization
  - 9 tasks, 40 languages
- Multilingual Large Language Models Evaluation Benchmark
  - MMLU / ARC / HellaSwag translated in 26 languages
- ...

④ Reductive Single metric issue

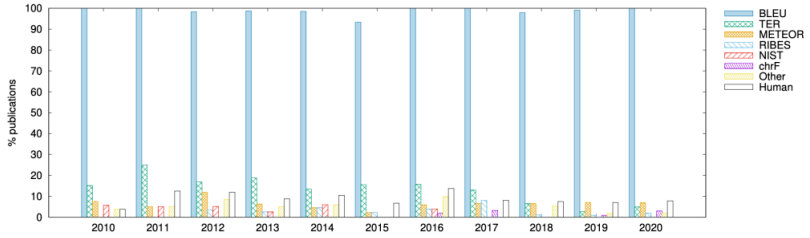
1. Performance is not all we care about
   1. computational efficiency ⇒ ML Perf
   2. bias ⇒ DiscrimEval
   ...
2. Taking averages for aggregation is unfair for minoritized groups
3. different preferences for different people

other bias { Bias metric
Bias LLM-Based evaluation

5

The challenges of challenges: statu quo issue

- Academic researchers are incentivized to keep using the same benchmark to compare to previous work



- 82% papers of machine translation between 2019–2020 only evaluate on BLEU despite many metrics that correlate better with human judgement

=> we will be incentivized not to look at something (metrics) else

Evaluation: 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 evaluating the right thing)
  - Contamination (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!**