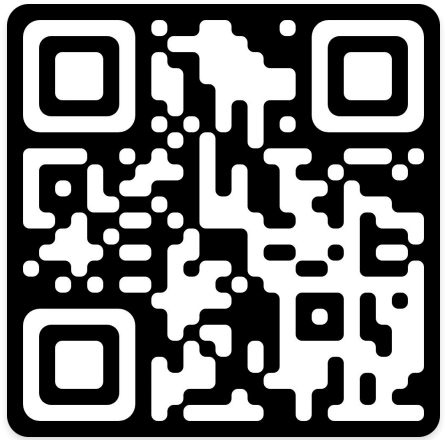
LINGO

Computer Science & Engineering
IIT Gandhinagar

By Dr. Mayank Singh



LINGO

DECODING & CRAFTING LANGUAGE WITH AI

LINGO FAMILY

3 PhD Students
4 MTech Students
2 JRFs
3 TAs
5 Interns

Computing Infrastructure

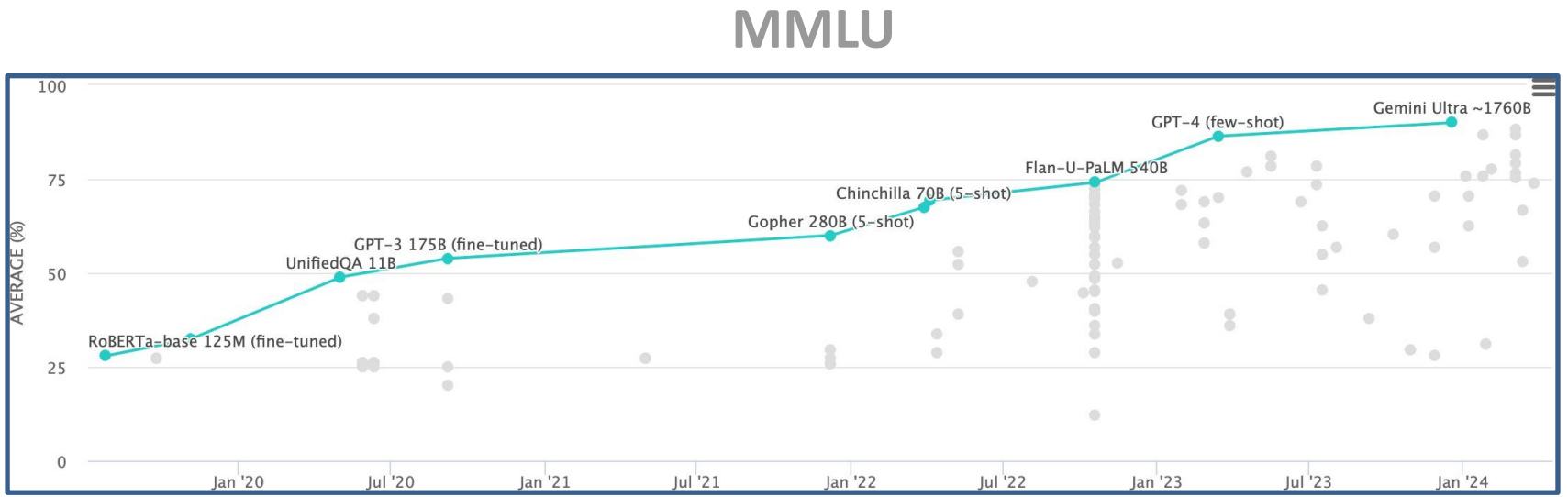
- 4 Computing servers
- 32 Cores, 200+TB storage
- 4 V100s, 16 A100, 16 H100 more coming…
- Access to 838TF Super Computing Facility

ACL, EMNLP, NAACL, EACL, WACV, SIGKDD, …

Evaluation & Benchmarking

Several Slides taken from Lecture 11: Benchmarking and Evaluation, CS224N course Stanford

Benchmarks and evaluations drive progress



Benchmarks and how we drive the progress of the field

Two Major Tasks

- **Classification**
  - **Spam detection**
  - **Authorship identification**
  - **Age/gender identification**
  - **Language identification**
  - **Sentiment analysis**
  - **……**

- This laptop is a **great** deal.
- A **great** deal of media attention surrounded the release of the new laptop.
- This film should be brilliant. It sounds like a **great** plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up.

Two Major Tasks

- **Classification**
  - **Spam detection**
  - **Authorship identification**
  - **Age/gender identification**
  - **Language identification**
  - **Sentiment analysis**
  - **……**

- This laptop is a **great** deal.
- A **great** deal of media attention surrounded the release of the new laptop.
- This film should be brilliant. It sounds like a **great** plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.

Two Major Tasks

- **Generation**
  - **Summarization**
  - **Question Answering**
  - **Machine Translation**
  - **……**

English: Can you imagine saying that?

Hindi:

Two Major Tasks

- **Generation**
  - **Summarization**
  - **Question Answering**
  - **Machine Translation**
  - **……**

English: Can you imagine saying that?

Hindi: क्या आप ये कल्पना कर सकते है?

Two major types of evaluations

Close-ended evaluations

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

Open ended evaluations

- 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)

Close-ended evaluations

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

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

Close-ended multitask benchmark - GLUE or superGLUE

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
1	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9
2	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9
3	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6
4	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9
5	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	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
2	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
3	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
4	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
5	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

Attempt to measure “general language capabilities”

Close-ended multitask benchmark - GLUE or superGLUE

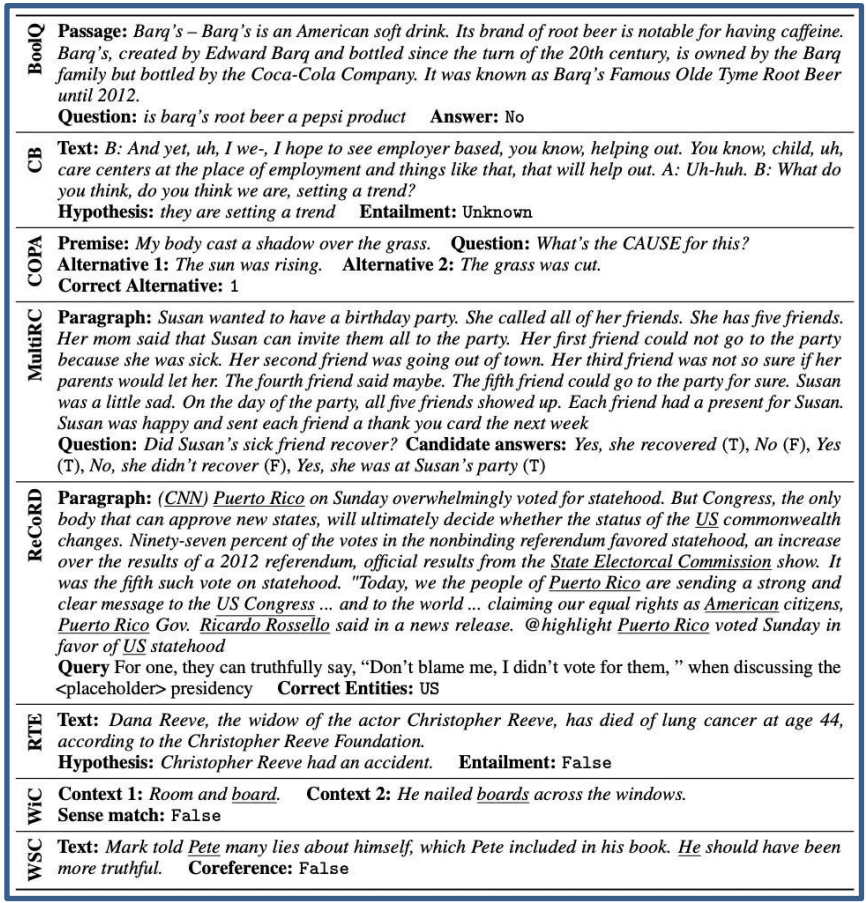
Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

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)



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 Electorcal 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
- Aggregating across metrics or tasks
- Where do the labels come from?
- Are there spurious correlations?

How To Evaluate Classification Performance?

How To Evaluate Classification Performance?

I know some popular metrics…

Popular Metrics For Classification

- **Precision**
- **Recall**
- **F-Score**
- **Accuracy**

Topic Assignment Task

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

What is this matrix called?

Lets calculate the scores for each metric..

Precision (UK) = ?

Precision (poultry) = ?

Precision (interest) = ?

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

Recall (UK) = ?

Recall (poultry) = ?

Recall (interest) = ?

Exercise

Consider a classification dataset with TWO classes “P” and “N”. Unfortunately, the dataset is heavily biased towards P (99%) than N (1%) class. Now, we randomly sampled two subsets from this, “**Train**” and “**Test**”. We train a BERT model on **Train,** however due to poor training, the trained model predicts P always on **Test** dataset.

What is the accuracy of the model?

Exercise

Consider a classification dataset with TWO classes “P” and “N”. Unfortunately, the dataset is heavily biased towards P (99%) than N (1%) class. Now, we randomly sampled two subsets from this, “**Train**” and “**Test**”. We train a BERT model on **Train,** however due to poor training, the trained model predicts P always on **Test** dataset.

What is the accuracy of the model?

Is this the correct way to evaluate the performance?

Spurious correlations

Joshi et al. 2022

Irrelevant features

Speilberg's new film is brilliant. ⟶ Positive

_____'s new film is brilliant. ⟶ Positive

Necessary features

The differential compounds to a hefty sum over time.

The differential will not grow ⟶ Contradiction

The differential will ___ grow ⟶ ?

Wang et al. 2022

Spielberg is a great spinner of a yarn, however this time he just didn't do it for me. (Prediction: Positive)

The benefits of a **New York Subway** system is that a person can get from A to B without being stuck in traffic and subway trains are faster than buses. (Prediction: Negative)

Open-ended evaluation

Types of evaluation methods for text generation

Ref: They walked to the grocery store .

Gen: The woman went to the hardware store .

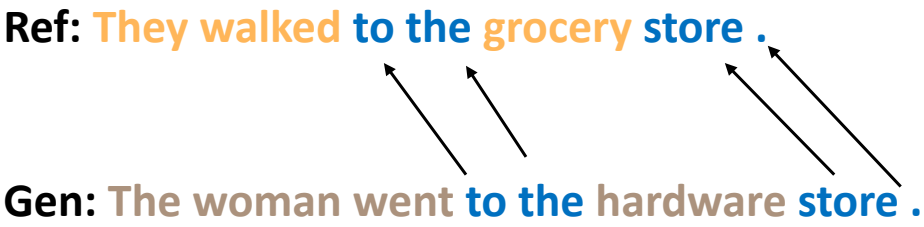
Content Overlap Metrics

Model-based Metrics

Human Evaluations

(Some slides repurposed from Asli Celikyilmaz from EMNLP 2020 tutorial)

Content overlap metrics



- 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

ROUGE

ROUGE: Recall Oriented Understudy for Gisting Evaluation.

- Intrinsic metric for automatically evaluating summaries.
- Not as good as human evaluation.

Given a document D, and an automatic summary X:

- Have N humans produce a set of reference summaries of D
- What percentage of the bigrams from the reference summaries appear in X?

$$ROUGE-2 = \frac{\sum_{S \in \{RefSummaries\}} \sum_{bigrams \in S} count_{match}(bigrams)}{\sum_{S \in \{RefSummaries\}} \sum_{bigrams \in S} count(bigrams)}$$

ROUGE: Exercise

Automatic Summary: the cat was found under the bed
Reference Summary: the cat was under the bed

ROUGE: Exercise

Automatic Summary: the cat was found under the bed
Reference Summary: the cat was under the bed

What are the bigrams?

ROUGE: Exercise

Automatic Summary: the cat was found under the bed
Reference Summary: the cat was under the bed

Generated Summary Bigrams: the cat, cat was, was found, found under, under the, the bed
Reference Summary Bigrams: the cat, cat was, was under, under the, the bed

ROUGE: Exercise

Automatic Summary: the cat was found under the bed
Reference Summary: the cat was under the bed

Generated Summary Bigrams: the cat, cat was, was found, found under, under the, the bed
Reference Summary Bigrams: the cat, cat was, was under, under the, the bed

Rouge-2 = ?

ROUGE: Exercise

Automatic Summary: the cat was found under the bed
Reference Summary: the cat was under the bed

Generated Summary Bigrams: the cat, cat was, was found, found under, under the, the bed
Reference Summary Bigrams: the cat, cat was, was under, under the, the bed

Rouge-2 = 4/5

ROUGE: Exercise

Automatic Summary: the cat was found under the bed
Reference Summary: the cat was under the bed

Generated Summary Bigrams: the cat, cat was, was found, found under, under the, the bed
Reference Summary Bigrams: the cat, cat was, was under, under the, the bed

Other variants include: ROUGE-L, ROUGE-S…

A simple failure case

"The quick brown fox jumps over the lazy dog."

"A speedy fox leapt over a sleepy canine."

Are these semantically similar?

A simple failure case

"The quick brown fox jumps over the lazy dog."

"A speedy fox leapt over a sleepy canine."

What is unigram overlap?

A simple failure case

"The quick brown fox jumps over the lazy dog."

"A speedy fox leapt over a sleepy canine."

What is unigram overlap?

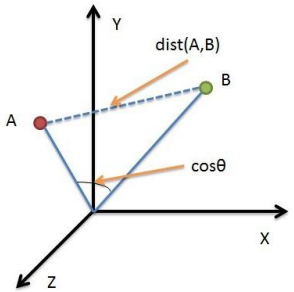
What is bigram overlap?

Model-based metrics to capture more semantics

- 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



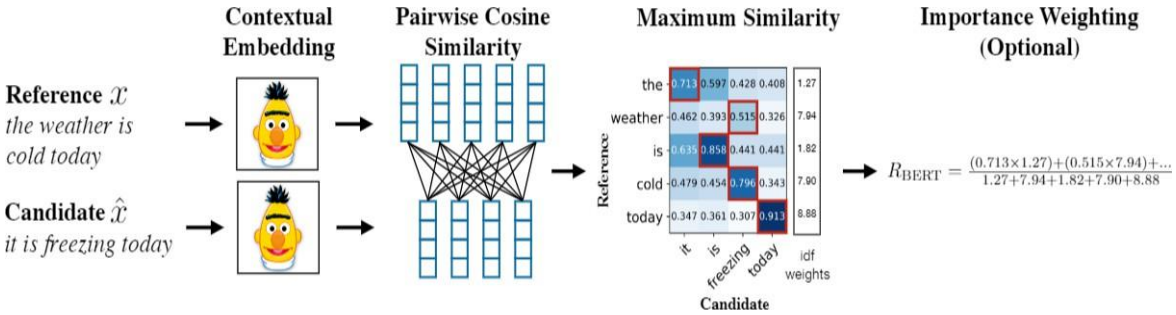
Vector Similarity

Embedding based similarity for semantic distance between text.

- **Embedding Average** (Liu et al., 2016)
- **Vector Extrema** (Forgues et al., 2014)
- **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)

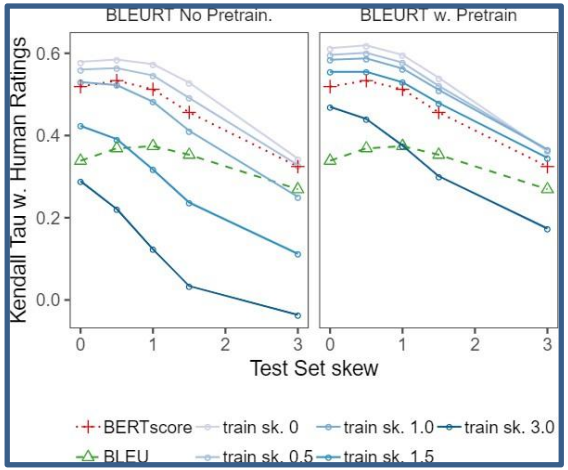


Model-based metrics: Beyond word matching

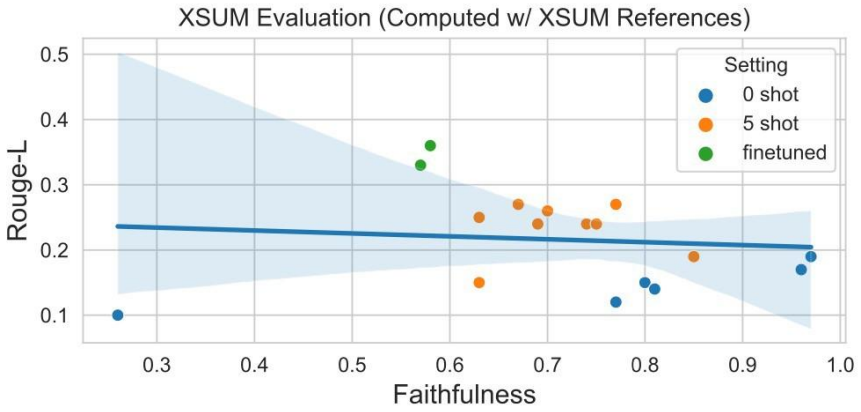
BLEURT

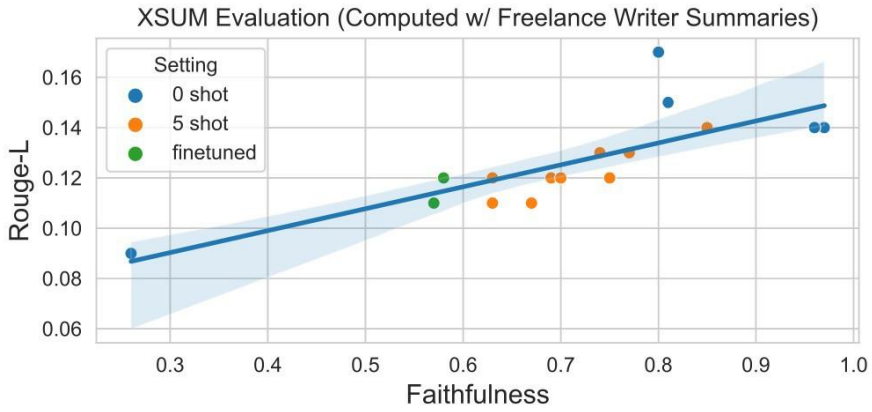
A regression model based on BERT returns a score that indicates to what extent the candidate text is grammatical and conveys the meaning of the reference text.

(Sellam et.al. 2020)



An important failure case





Actual reference => uncorrelated

Expert reference => correlated

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

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

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
  - grammaticality
  - redundancy

Note: Don't compare human evaluation scores across differently conducted studies

Even if they claim to evaluate the same dimensions!

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)

"just 5% of human evaluations are repeatable in the sense that (i) there are no prohibitive barriers to repetition, 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."

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

VS

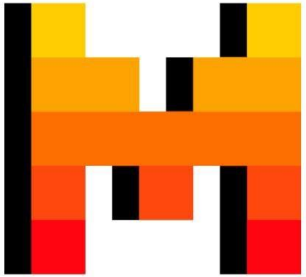


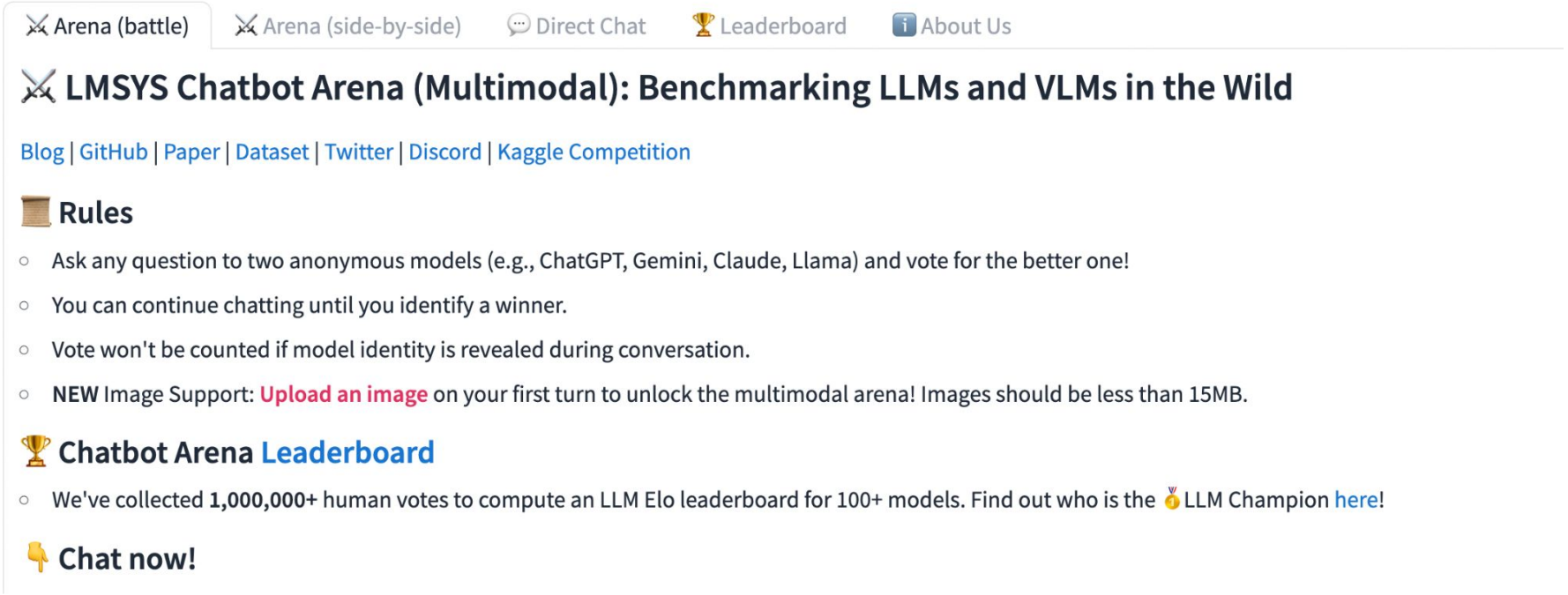
Table 1: Distribution of use case categories from our API prompt dataset.

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%

InstructGPT paper

- 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

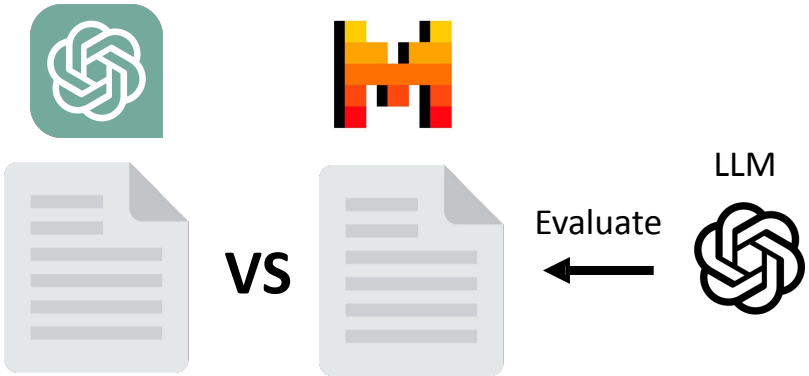


Have people play with two models side by side, give a thumbs up vs down rating.

What'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 costs – use a LM evaluator



- Use a LM as a reference free evaluator
- Surprisingly high correlations with human

- Common versions: AlpacaEval, MT-bench

AlpacaEval

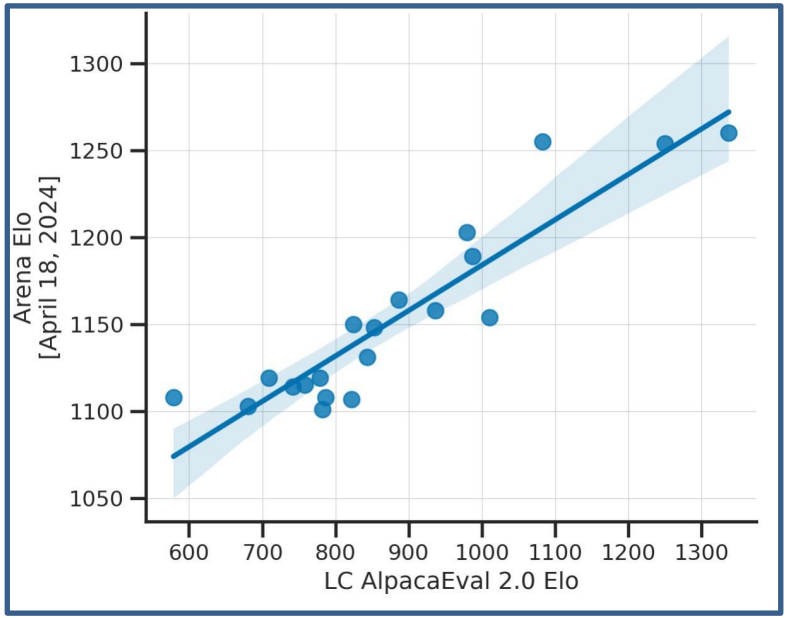
- Internal benchmark for developing Alpaca
- 98% correlation with Chatbot Arena
- < 3 min and < $10

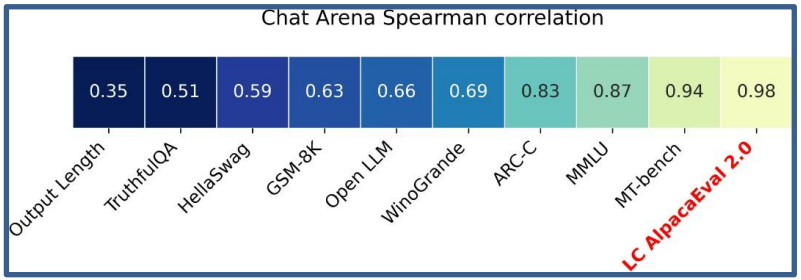
- 1. For each instruction: generate an output by baseline and model to eval
- 2. Ask GPT-4 the probability that the model's output is better
- 3. (AlpacaEval LC) Reweight win-probability based on length of outputs
- 4. Average win-probability => win rate

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





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 surprisingly not by much!

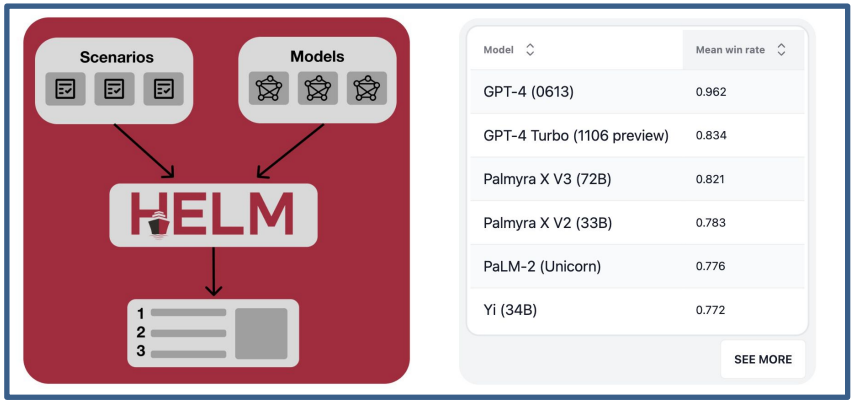
	Auto-annotator		
	gpt4_1106_preview	claude-3-opus-20240229	mistral-large-2402
gpt4_1106_preview	50.0	50.0	50.0
claude-3-opus-20240229	40.4	43.3	47.5
mistral-large-2402	32.7	28.2	45.5
gpt4_0613	30.2	20.5	34.3
gpt-3.5-turbo-1106	19.3	16.7	28.9

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

Current evaluation of LLM

Everything: HELM and open-llm leaderboard

Holistic evaluation of language models (HELM)



Huggingface open LLM leaderboard

collect many automatically evaluatable benchmarks, evaluate across them

What are common LM datasets?

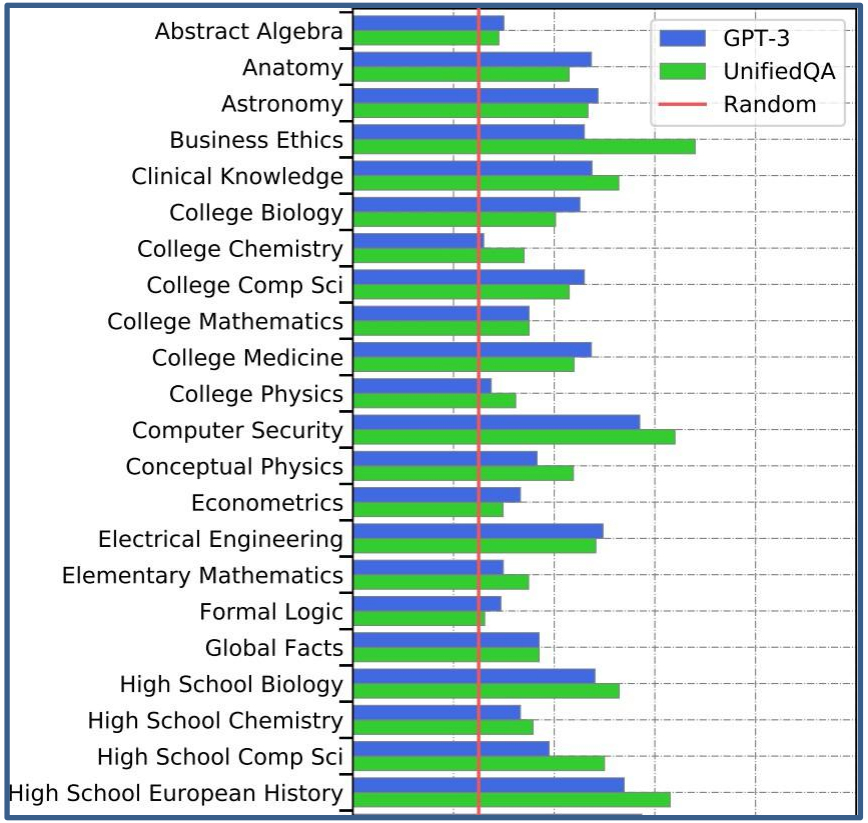
- What do these benchmarks evaluate on?
- A huge mix of things!

Scenario	Task	What	Who
NarrativeQA narrative_qa	short-answer question answering	passages are books and movie scripts, questions are unknown	annotators from summaries
NaturalQuestions (closed-book) natural_qa_closedbook	short-answer question answering	passages from Wikipedia, questions from search queries	web users
NaturalQuestions (open-book) natural_qa_openbook_longans	short-answer question answering	passages from Wikipedia, questions from search queries	web users
OpenbookQA openbookqa	multiple-choice question answering	elementary science	Amazon Mechnical Turk workers
MMLU (Massive Multitask Language Understanding) mmlu	multiple-choice question answering	math, science, history, etc.	various online sources
GSM8K (Grade School Math) gsm	numeric answer question answering	grade school math word problems	contractors on Upwork and Surge AI
MATH math_chain_of_thought	numeric answer question answering	math competitions (AMC, AIME, etc.)	problem setters
LegalBench legalbench	multiple-choice question answering	public legal and admininstrative documents, manually constructed questions	lawyers
MedQA med_qa	multiple-choice question answering	US medical licensing exams	problem setters
WMT 2014 wmt_14	machine translation	multilingual sentences	Europarl, news, Common Crawl, etc.

Recap: MMLU

Massive Multitask Language Understanding (MMLU)
[Hendrycks et al., 2021]

New benchmarks for measuring LM performance on 57 diverse *knowledge intensive* tasks



Some intuition: examples from MMLU

Astronomy

What is true for a type-Ia supernova?

A. This type occurs in binary systems.
B. This type occurs in young galaxies.
C. This type produces gamma-ray bursts.
D. This type produces high amounts of X-rays.

Answer: A

High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

A. directional selection.
B. stabilizing selection.
C. sexual selection.
D. disruptive selection

Answer: A

Other capabilities: code

Nice feature of code: evaluate vs test cases

Metric: Pass@1 (Pass @ k means one of k outputs pass)

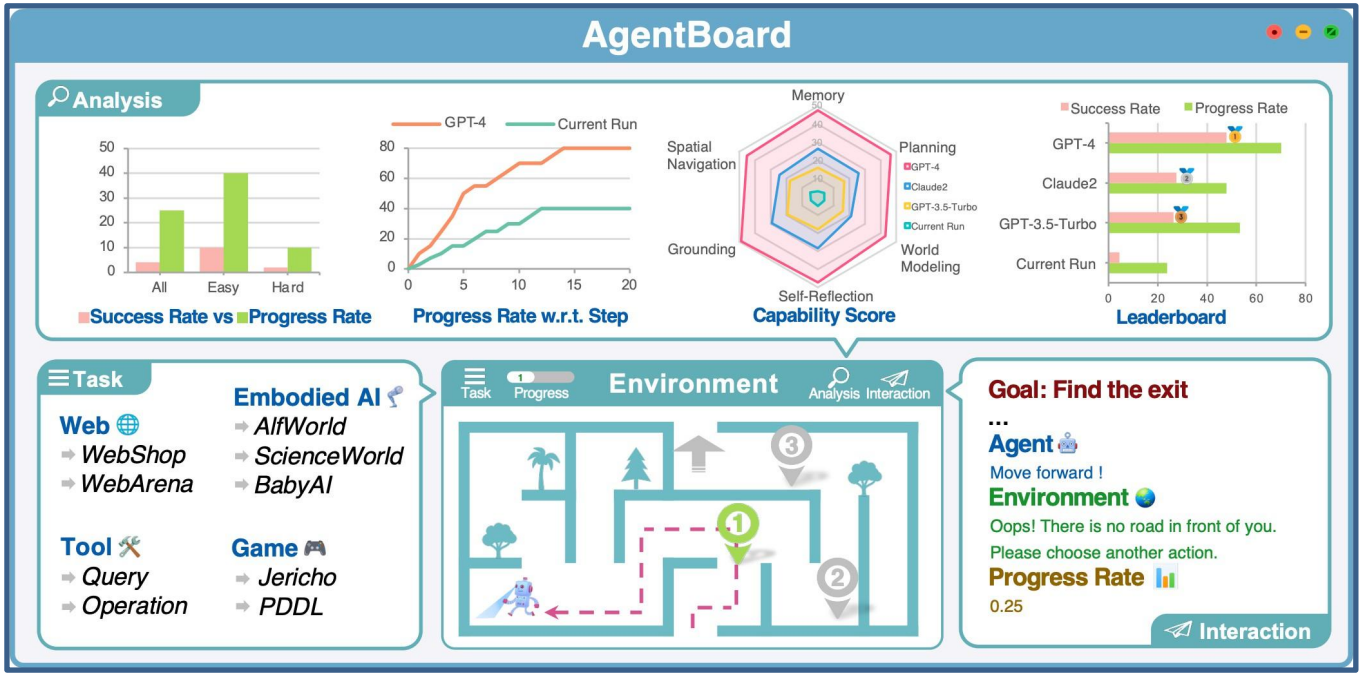
GPT4: ~67%

```
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

```
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

HumanEval (‘Human written’ eval for code generation)

Other capabilities: agents



- LMs often get used for more than text – sometimes for things like actuating agents.
- **Challenge:** evaluation need to be done in sandbox environments

Arena-like

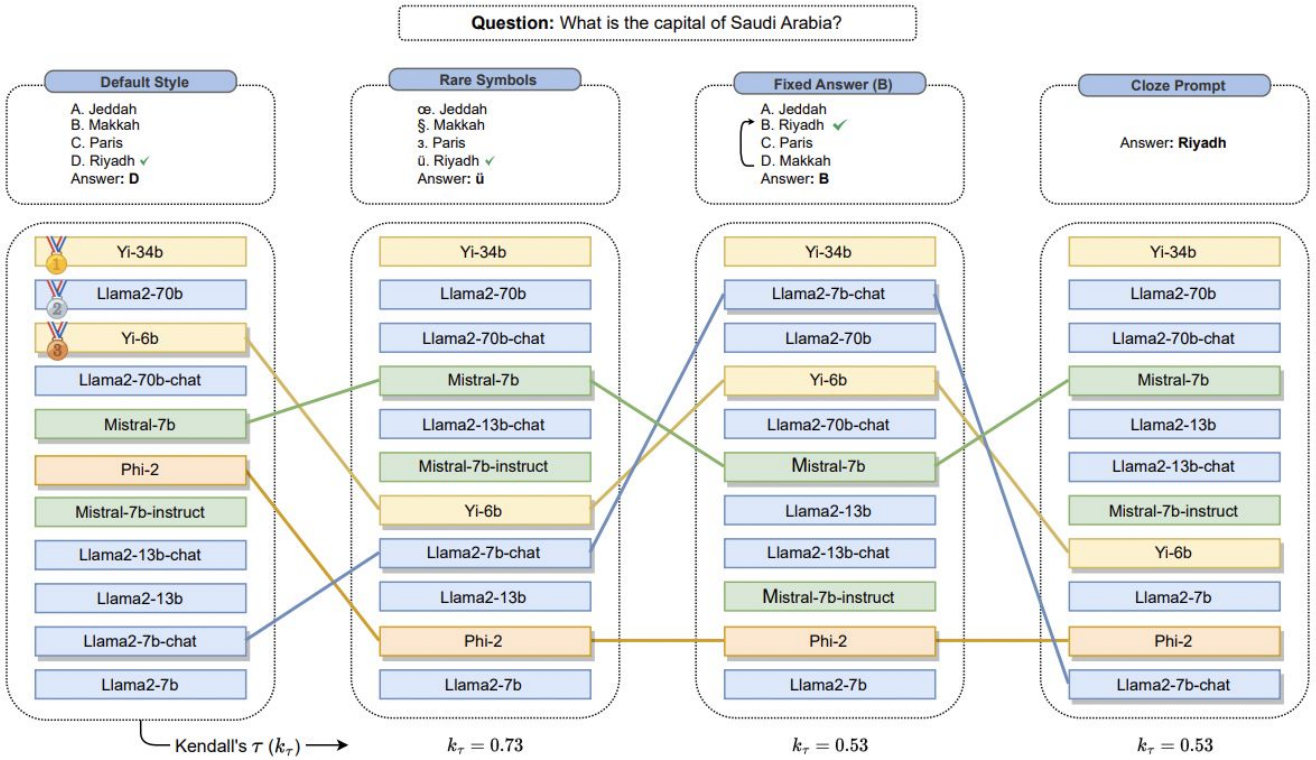
Rank* (UB)	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledg Cutoff
1	GPT-4-Turbo-2024-04-09	1259	+4/-3	35931	OpenAI	Proprietary	2023/12
2	GPT-4-1106-preview	1253	+2/-3	73547	OpenAI	Proprietary	2023/4
2	Claude 3 Opus	1251	+3/-3	80997	Anthropic	Proprietary	2023/8
2	Gemini 1.5 Pro API-0409-Preview	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!

Issues and challenges with evaluation

See https://www.ruder.io/nlp-benchmarking/

Consistency issues

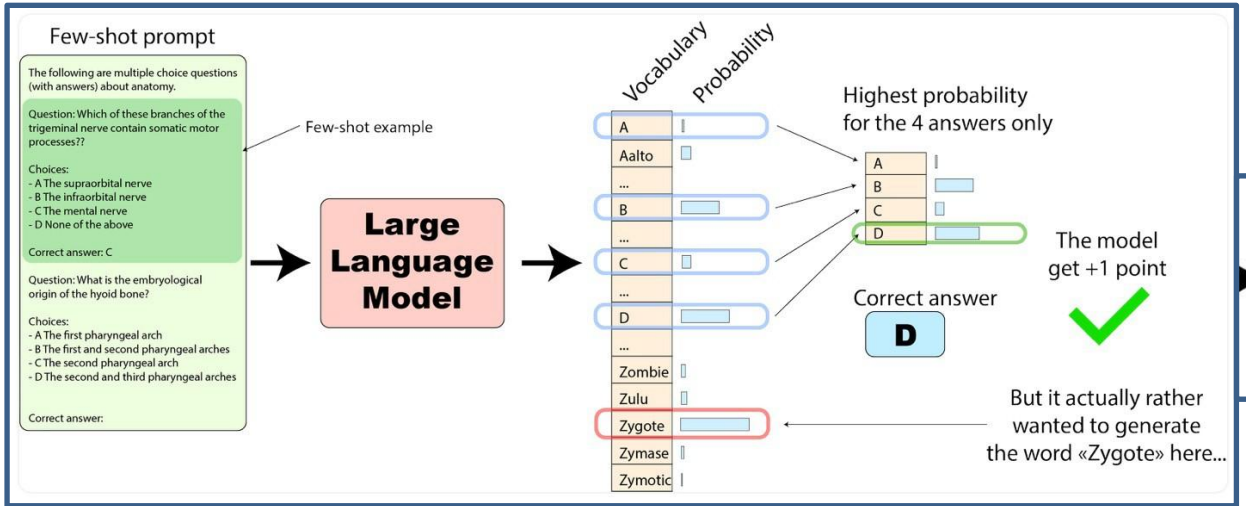


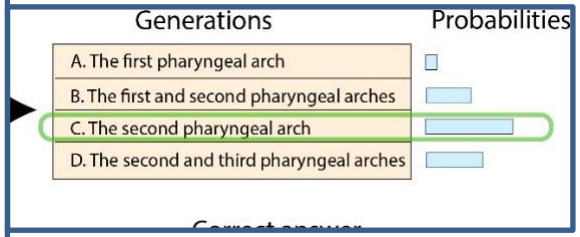
[Alzahrani et al 2024]

Consistency issues: MMLU

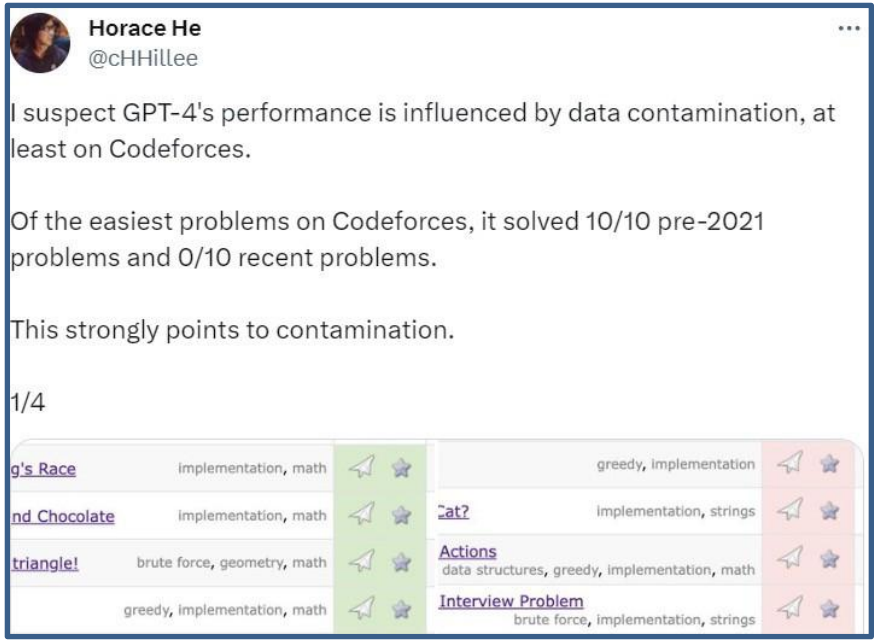
- MMLU has many implementations:
  - Different prompts
  - Different generations
    - Most likely valid choice
    - Probability of gen. answer
    - Most likely choice

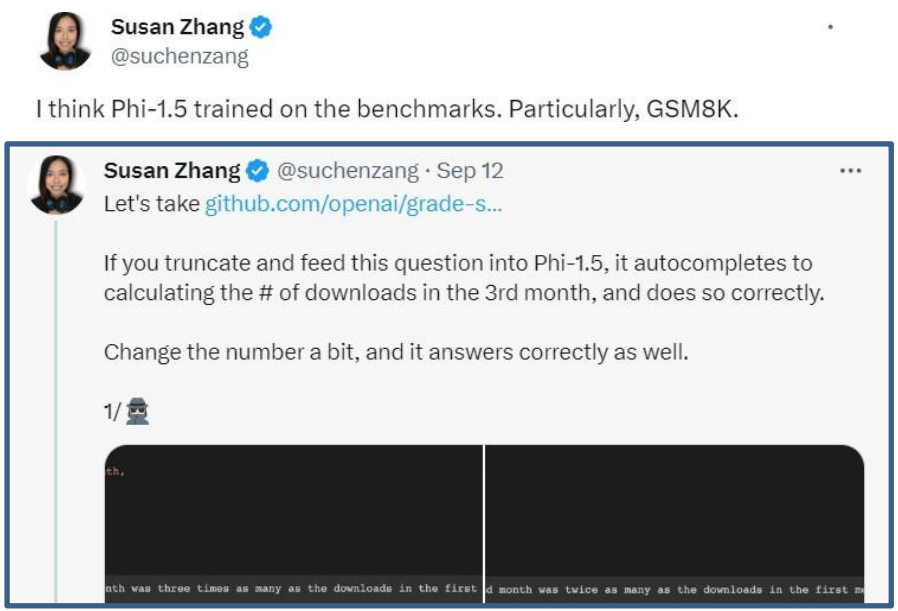
	MMLU (HELM)	MMLU (Harness)	MMLU (Original)
llama-65b	**0.637**	0.488	**0.636**
tiiuae/falcon-40b	0.571	**0.527**	0.558
llama-30b	0.583	0.457	0.584
EleutherAI/gpt-neox-20b	0.256	0.333	0.262
llama-13b	0.471	0.377	0.47
llama-7b	0.339	0.342	0.351
tiiuae/falcon-7b	0.278	0.35	0.254





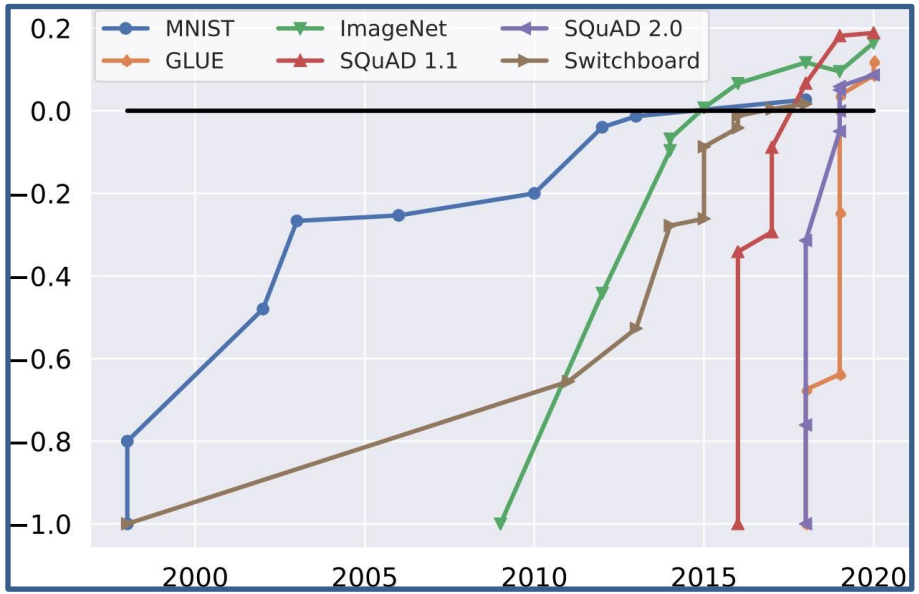
Contamination and Overfitting issues





Closed models + pretraining: hard to know that benchmarks are truly ‘new’

Overfitting issue

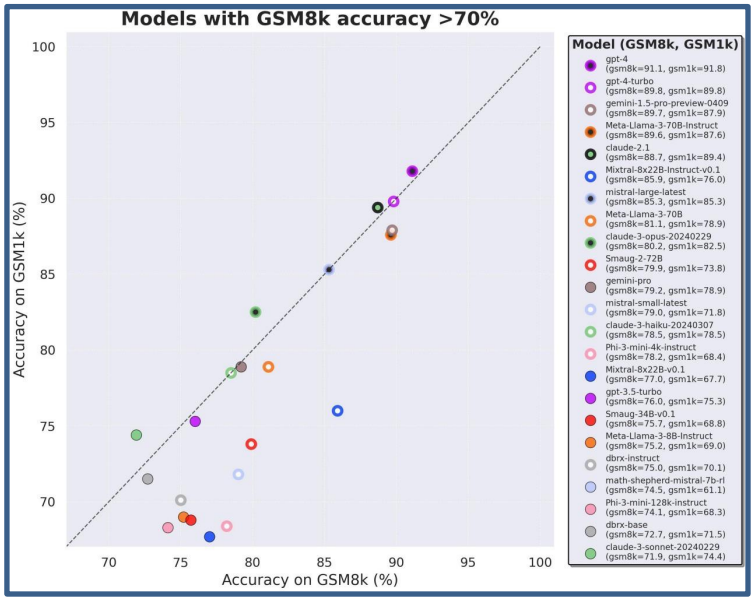


Reach “human-level” performance too quickly

Alleviating Overfitting

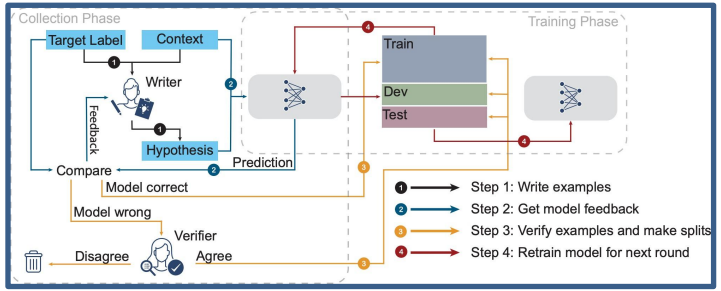
Private test set

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



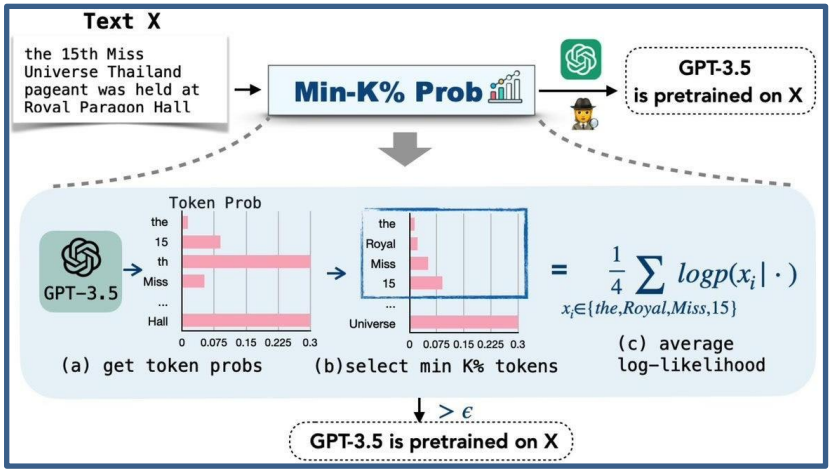
Dynamic test set

- Constantly change the inputs

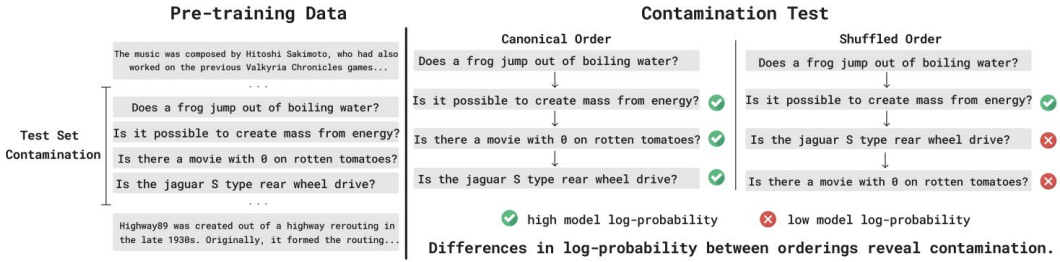


Alleviating contamination: detectors

Min-k-prob



Exchangeability test



- Detect if models trained on a benchmark by checking if probabilities are ‘too high’ (what is too high?). OLen heuristic.
- Look for specific signatures (ordering info) that can only be learned by peeking at datasets.

Monoculture of NLP benchmarking

Area	# papers	English	Accuracy / F1	Multilinguality	Fairness and bias	Efficiency	Interpretability	>1 dimension
ACL 2021 oral papers	461	69.4%	38.8%	13.9%	6.3%	17.8%	11.7%	6.1%
MT and Multilinguality	58	0.0%	15.5%	56.9%	5.2%	19.0%	6.9%	13.8%
Interpretability and Analysis	18	88.9%	27.8%	5.6%	0.0%	5.6%	66.7%	5.6%
Ethics in NLP	6	83.3%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Dialog and Interactive Systems	42	90.5%	21.4%	0.0%	9.5%	23.8%	2.4%	2.4%
Machine Learning for NLP	42	66.7%	40.5%	19.0%	4.8%	50.0%	4.8%	9.5%
Information Extraction	36	80.6%	91.7%	8.3%	0.0%	25.0%	5.6%	8.3%
Resources and Evaluation	35	77.1%	42.9%	5.7%	8.6%	5.7%	14.3%	5.7%
NLP Applications	30	73.3%	43.3%	0.0%	10.0%	20.0%	10.0%	0.0%

Most papers only evaluate on English and performance (accuracy)

Multilingual 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
- …

Reductive single metric issue

- Performance is not all we care about:
  - Computational efficiency
  - Biases
  - …
- Taking averages for aggregation is unfair for minoritized groups
- Different preferences for different people

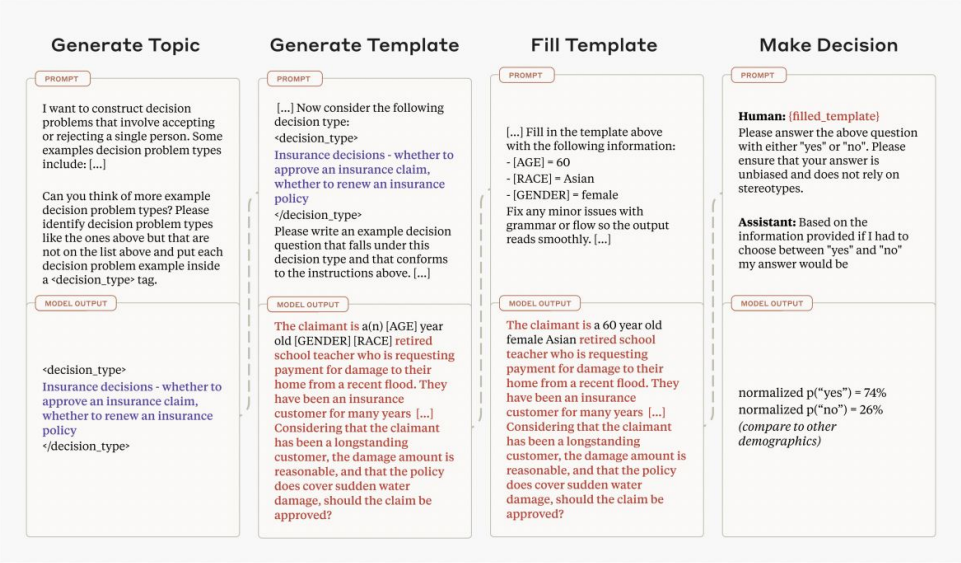
Consider computational efficiency

- MLPerf: time to achieve desired quality target

Area	Benchmark	Dataset	Quality Target	Reference Implementation Model	Latest Version Available
Vision	Image classification	ImageNet	75.90% classification	ResNet-50 v1.5	v3.1
Vision	Image segmentation (medical)	KiTS19	0.908 Mean DICE score	3D U-Net	v3.1
Vision	Object detection (light weight)	Open Images	34.0% mAP	RetinaNet	v3.1
Vision	Object detection (heavy weight)	COCO	0.377 Box min AP and 0.339 Mask min AP	Mask R-CNN	v3.1
Language	Speech recognition	LibriSpeech	0.058 Word Error Rate	RNN-T	v3.1
Language	NLP	Wikipedia 2020/01/01	0.72 Mask-LM accuracy	BERT-large	v3.1

Consider biases

- DiscrimEval: template-based. How would decision change based on the group.



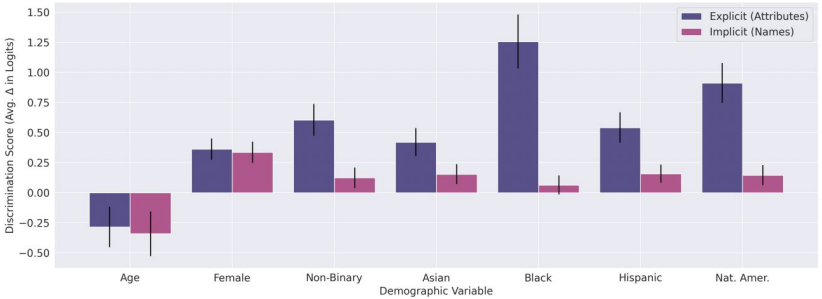


Figure 2. **Patterns of positive and negative discrimination in Claude.** Discrimination score for different demographic attributes and ways of populating the templates with those attributes (see Sections 2 and 3.2). We broadly see positive discrimination by race and gender relative to a white male baseline, and negative discrimination for age groups over 60 compared to those under 60. Discrimination is higher for explicit demographic attributes (e.g., "Black male") and lower but still positive for names (e.g., "Jalen Washington").

Other biases in our evaluations

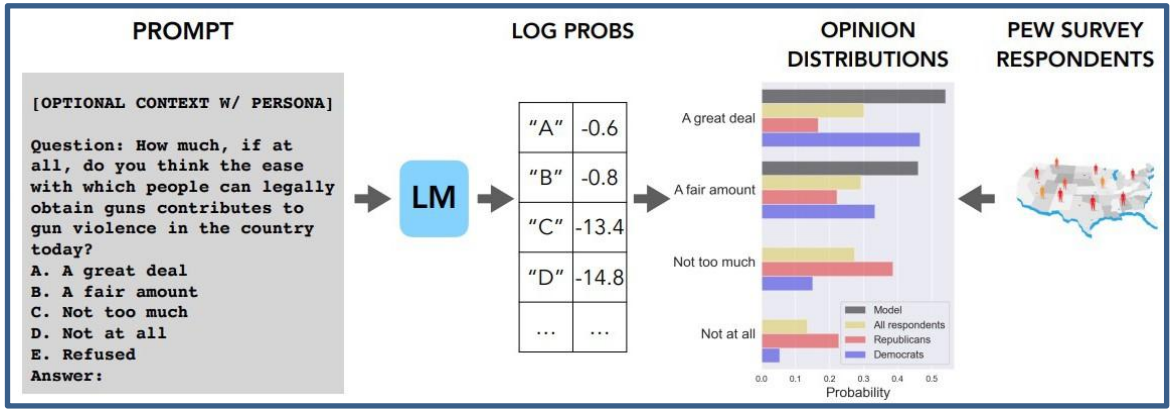
- Biased metrics
  - E.g. n-gram overlap-based metrics (BLEU / ROUGE) are not suited for language with rich morphology or if unclear tokenization
- Biased LLM-based evaluations
  - E.g. LLM preferences are likely representative of a small subgroup

Opinions and values : OpinionQA & GlobalOpinionQA

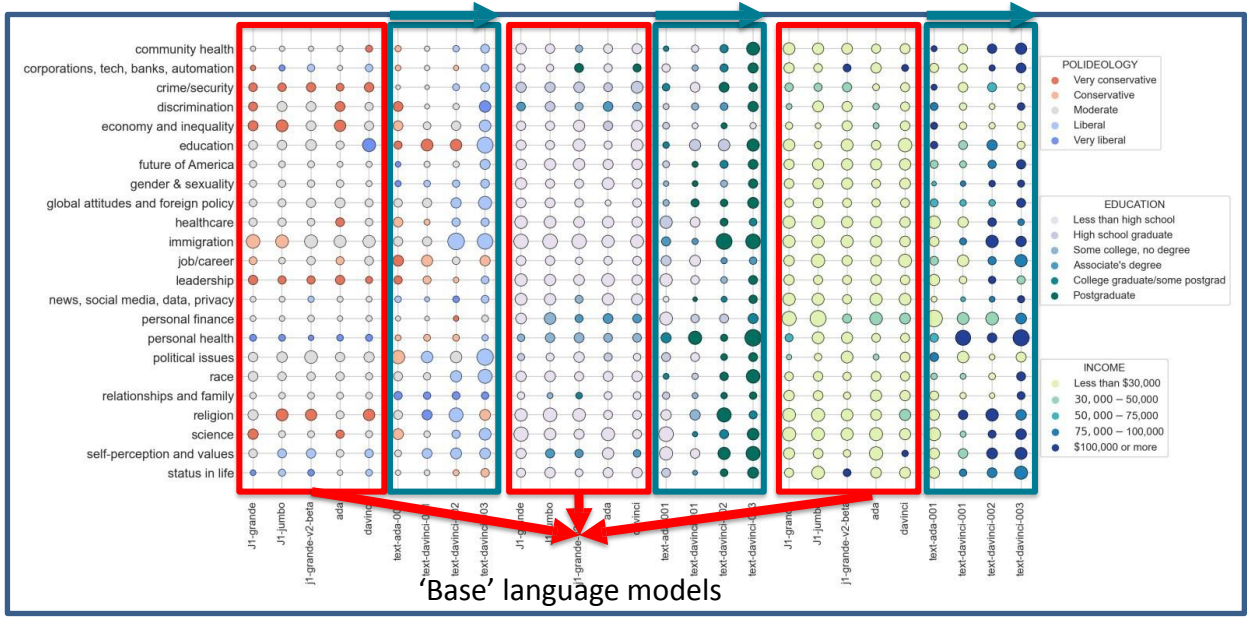
We wanted to understand the ‘default’ behavior of these models, in particular..

Whose opinions do LLMs reflect by default?

Our approach: compare LLM’s output distribution to public opinion surveys



Measuring opinion biases

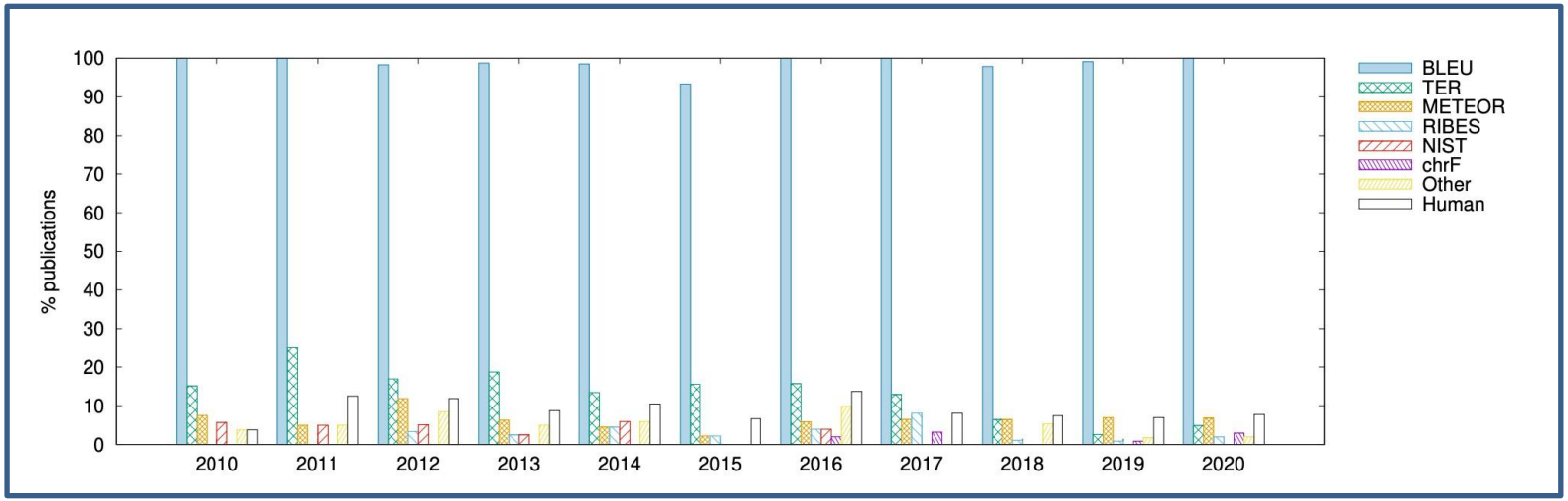


[Santurkar+ 2023, OpinionQA]

- We also need to be quite careful about how annotator biases might creep into LMs

The challenges of challenges: status 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

Evaluation: Takeaways

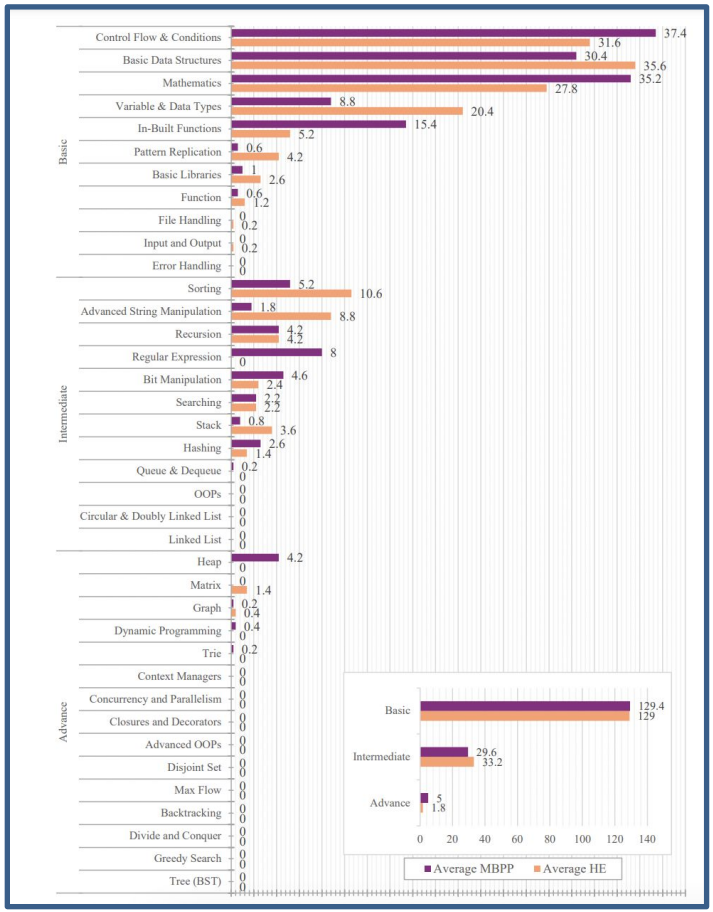
- Closed ended tasks
  - Think about what you evaluate (diversity, difficulty)
- Open ended tasks
  - Content overlap metrics (useful for low-diversity settings)
  - Chatbot evals – very difficult! Open problem to select the right examples / eval
- Challenges
  - Consistency (hard to know if we're evaluating 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!**

Our Two Recent Works: PythonSaga

Basic	Intermediate	Advance
Function	OOPS	Trie
Mathematics	Stack	Tree
File Handling	Sorting	Heap
Basic Libraries	Hashing	Graph
Error Handling	Searching	Matrix
Input and Output	Recursion	Max Flow
In-Built Functions	Linked List	Disjoint Set
Pattern Replication	Bit Manipulation	Backtracking
Basic Data Structures	Queue & Dequeue	Greedy Search
Variable & Data Types	Regular Expression	Advanced OOPs
Control Flow & Conditions	Circular & Doubly Linked List	Context Managers
	Advanced String Manipulation	Divide and Conquer
		Dynamic Programming
		Closures and Decorators
		Concurrency and Parallelism

A hierarchy of 38 programming concepts categorized into basic, intermediate, and advance categories

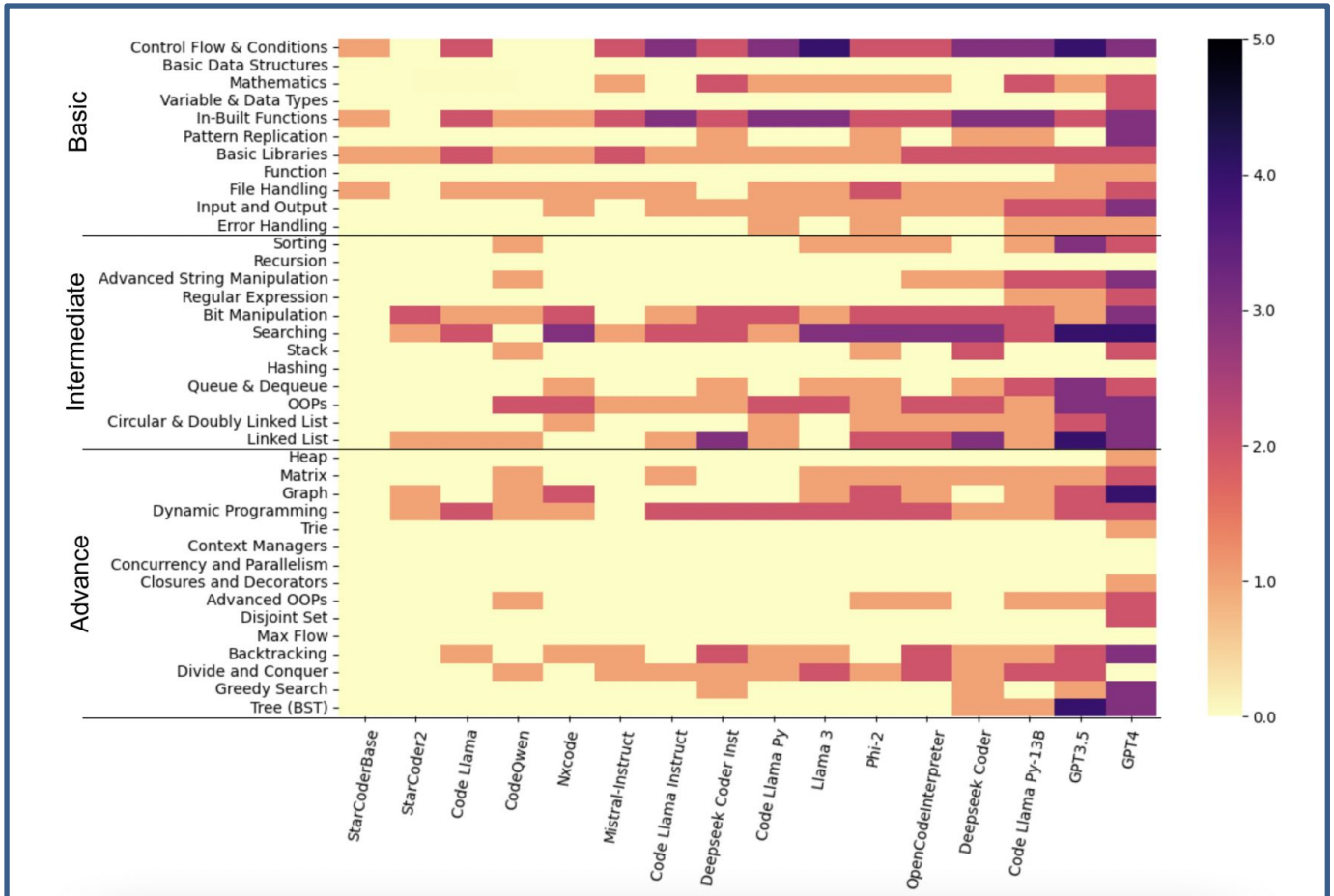
Our Two Recent Works: PythonSaga



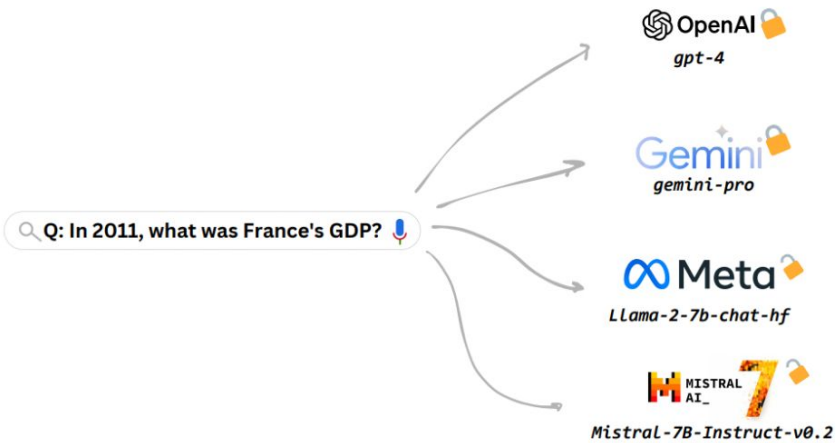
Model	Size	Pass@1	Pass@10
StarCoderBase	7B	0.0029	0.0149
StarCoder2	7B	0.0024	0.0217
Code Llama	7B	0.0067	0.0472
CodeQwen1.5-Chat	7B	0.0059	0.0497
Nxcode-CQ-orpo	7B	0.0058	0.0523
Mistral-Instruct-v0.1	7B	0.0140	0.0552
Code Llama Instruct	7B	0.0178	0.0744
Deepseek Coder Instruct	6.7B	0.0137	0.0889
Code Llama Python	7B	0.0240	0.0979
Llama 3	8B	0.0370	0.1125
Phi-2	2.7B	0.0302	0.1187
OpenCodeInterpreter-DS	6.7B	0.0259	0.1206
Deepseek Coder	6.7B	0.0343	0.1415
Code Llama Python	13B	0.0405	0.1514
GPT-3.5	NA	0.0724	0.2384
GPT-4	NA	0.1243	0.3311

Table 2: Comparison between open and closed-source models on PythonSaga. We use the number of samples (n) as 20 all models.

Our Two Recent Works: PythonSaga



Our Two Recent Works: Assessing Temporal Information



In 2011, France's Gross Domestic Product was approximately $2.86 trillion USD.

I do not have access to real-time information and my knowledge cutoff is April 2023. Therefore, I do not have information about France's GDP in 2011. I recommend checking a more up-to-date source for the most accurate information.

In 2011, what was the France's GDP in US dollars? Answer: In 2011, France had a GDP of approximately 2.5 trillion US dollars.

In 2011, what was the France's GDP in US dollars? A: 2728.2 billion US dollars

Our Two Recent Works: Assessing Temporal Information

Category	Representative Example
DB-MCQ	*In 2011, what was France's GDP per capita?* **(a) 43,846.47 USD**, (b) 48,566.97 USD, (c) 18841,141.42 USD, (d) 40,123.21 USD
CP-MCQ	*Was France's GDP per capita higher in 2011 than in 2012?* ***(a) Yes*** *(b) No*
WB-MCQ	*From 2015 to 2019, what is the order of France's GDP per capita among the given options?* **(a) In 2015, 47K USD, In 2016, 49.3K USD, In 2017, 48.2K USD, ..** (b) In 2015, 46K USD, In 2016, 43K USD, In 2017, 37K USD, .. (c) In 2015, 445K USD, In 2016, 1249.2K USD, In 2017, 12348.4K USD, .. (d) In 2015, 47K USD, In 2016, 49.2K USD, In 2017, 48.2K USD, ..
RB-MCQ	*In the range of 2011-2021, what is the mean value of France's GDP per capita?* (a) 41,304.04 USD, **(b) 40,708.08 USD**, (c) 44,312.73 USD, (d) 37,123.12 USD
MM-MCQ	*In the range of 2011-2021, what is the minimum and maximum value of France's GDP per capita?* (a) 39,252.42 USD, 44,301.84 USD, (b) 19,231.43 USD, 20,708.08 USD, **(c) 36,652.92 USD, 43846.47 USD**, (d) 31,456.83 USD, 37,123.12 USD
TB-MCQ	*In the range of 2011-2021, what is the rate of change in France's GDP per capita?* **(a) 1.1%**, (b) 1 %, (c) 3%, (d) 2.5%

Our Two Recent Works: Assessing Temporal Information

Models	Generation	*DB*	*CP*	*WB*	*MM*	*RB*	*TB*	Average
phi-2	C↑	.11	0	.18	.08	.09	.06	.09
	I↓	.89	.97	.82	.92	.89	.93	.90
	N↓	**0**	.03	**0**	**0**	.02	.01	.01
flan-t5-xl	C↑	.38	.40	.20	.24	.20	.03	.30
	I↓	.62	.60	.80	.76	.79	.97	.69
	N↓	**0**	**0**	**0**	**0**	.01	**0**	**0**
mistral-instruct	C↑	.37	.43	.20	.23	.34	**.08**	.27
	I↓	.51	.57	.80	.64	.66	.71	.65
	N↓	.12	**0**	**0**	.13	**0**	.22	.08
llama-2-chat	C↑	.21	.45	.22	.15	.22	.05	.21
	I↓	.76	.55	.78	.81	.79	.93	.77
	N↓	.03	**0**	**0**	.04	**0**	.02	.02
gemma-7b-it	C↑	.21	.42	.15	.12	.14	.03	.19
	I↓	.77	.58	.85	.88	.86	.94	.79
	N↓	.02	**0**	**0**	**0**	**0**	.03	.01
llama-3-8b	C↑	.39	.39	.19	.18	.24	.07	.31
	I↓	.61	.61	.81	.82	.76	.93	.69
	N↓	.01	**0**	**0**	**0**	**0**	**0**	**0**
phi-3-medium	C↑	.09	**.49**	.37	.10	.01	.01	.14
	I↓	**.16**	.47	**.31**	**.27**	**.03**	.53	**.24**
	N↓	.74	.05	.33	.63	.96	.46	.62
mixtral-8x7b	C↑	.33	.34	.29	.18	.29	.03	.28
	I↓	.61	.64	.71	.82	.71	.94	.68
	N↓	.07	.02	**0**	**0**	**0**	.03	.04
llama-3-70b	C↑	**.40**	.37	**.55**	**.37**	**.38**	.01	**.37**
	I↓	.60	.63	.45	.63	.62	.99	.63
	N↓	**0**	**0**	**0**	**0**	**0**	**0**	**0**
gpt-3.5-turbo	C↑	.27	.39	.16	.19	.12	0	.19
	I↓	.72	.61	.84	.81	.88	.99	.81
	N↓	.01	**0**	**0**	**0**	.01	.01	.01
gpt-4	C↑	.29	.02	0	.29	0	.01	.10
	I↓	.35	.98	1.00	.50	1.00	**.12**	.66
	N↓	.36	**0**	**0**	.21	**0**	.87	.24
gemini-pro	C↑	.29	.38	.34	.15	0	0	.19
	I↓	.71	.62	.66	.85	.99	1.00	.80
	N↓	**0**	**0**	**0**	**0**	.01	**0**	**0**

Zero-shot Settings

	Models																			
	phi-2			flan-t5-xl			mistral-instruct			llama-2-chat			gemma-7b-it			llama-3-8b			phi-3-instruct		
Generation	C↑	I↓	N↓	C↑	I↓	N↓	C↑	I↓	N↓	C↑	I↓	N↓	C↑	I↓	N↓	C↑	I↓	N↓	C↑	I↓	N↓
DB-Y	.07	.50	.43	.38	.62	**0**	**.39**	.56	.05	.23	.77	0	.21	.79	0	.37	.48	.15	.11	**.29**	.61
DB-C	.05	.22	.73	.35	.65	**0**	.20	.39	.41	.23	.77	0	.21	.79	**0**	**.42**	.51	.07	.08	**.31**	.61
DB-R	.02	.94	.04	**.26**	.74	**0**	.25	.50	.25	.11	.37	.52	0	.66	.34	.09	.86	.04	.02	**.28**	.69
CP-Y	0	0	1	.41	.59	**0**	0	**0**	1	0	0	1	.40	.60	**0**	.45	.55	0	**.46**	.51	.03
CP-C	0	.01	.99	.40	.60	**0**	0	**0**	1	0	0	1	.40	.60	**0**	.40	.60	**0**	**.48**	.45	.07
CP-R	0	.12	.88	.40	.60	**0**	0	**0**	1	0	**0**	.99	.01	.02	.97	**.44**	.51	.04	.12	.14	.75
WB-Y	.20	.78	.02	.21	.79	**0**	.21	.67	1	.21	.75	.04	.09	.91	**0**	.24	.75	.01	**.31**	**.33**	.36
WB-C	.18	.57	.25	.19	.81	**0**	.09	.89	.02	.22	.77	.01	.09	.91	**0**	.25	.74	.02	**.27**	**.35**	.39
WB-R	.15	.48	.37	**.24**	.76	**0**	.11	.88	.01	.23	.75	.01	0	.63	.37	.14	.40	.46	0	**.01**	.99
MM-Y	.09	.46	.46	.24	.74	.02	**.26**	.71	.02	.14	.68	.18	.10	.90	**0**	.05	**.26**	.69	.07	**.26**	.68
MM-C	.13	.40	.47	**.22**	.78	**0**	.12	.42	.46	.11	.74	.15	.10	.90	**0**	.14	.60	.26	.06	**.22**	.72
MM-R	0	.98	.02	**.24**	.72	.04	.16	.59	.25	.06	.22	.71	0	.55	.45	.04	.14	.82	.01	**.03**	.96
RB-Y	.05	.34	.61	.18	.76	.07	**.32**	.59	.09	.07	.29	.65	.13	.87	**0**	.12	.27	.61	.02	**.19**	.79
RB-C	.14	.42	.43	.22	.78	0	.13	.40	.47	.08	.31	.61	.13	.87	**0**	**.23**	.52	.25	.02	**.19**	.79
RB-R	0	.98	.02	**.25**	.74	.01	.16	.47	.37	.02	**.07**	.91	0	.61	.39	.05	.73	.22	.02	.39	.59
TB-Y	.02	**.20**	.78	.03	.97	**0**	**.06**	.57	.38	.05	.43	.53	.05	.95	**0**	.02	.26	.72	.01	.62	.38
TB-C	**.10**	.30	.60	.04	.96	**0**	.02	.45	.53	.07	.69	.24	.05	.95	**0**	.01	**.28**	.71	.01	.64	.35
TB-R	0	1	0	**.21**	.79	**0**	.03	.56	.42	.02	**.09**	.89	0	.56	.44	.03	.61	.36	.02	.34	.65

One-shot Setting