

Navigating the Modern Evaluation Landscape

Leshem Choshen^{↑□}, Ariel Gera[↑], Yotam Perlitz[↑],
Michal Shmueli-Scheuer[↑], Gabriel Stanovsky[§]

LREC-COLING  2024

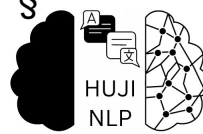
□



↑



§



Michal



Introduction, 30 mins

Frameworks for Benchmarking, 15 mins

Gabi



Prompts, 45 mins

Yotam



Efficiency considerations, 45 mins

Leshem



Metrics, 45 mins

Ariel



Manual Evaluation, 30 mins

Conclusion, 5 mins

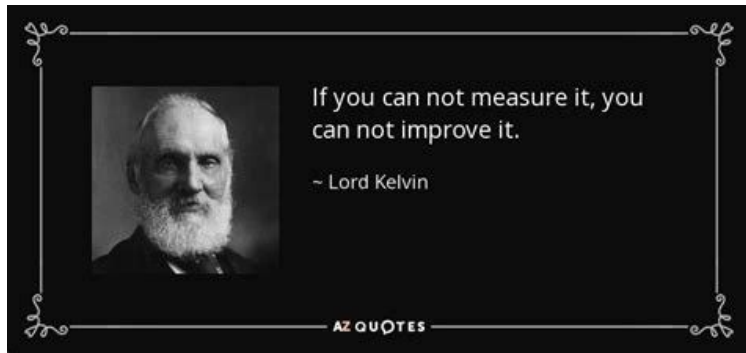
Part 1: Introduction

Model Evaluation- Goals

- To assess and measure the performance of a model in various tasks.
- Being able to compare it to other models, or earlier version of the model

Why Focus on Evaluation

Necessary



Not trivial

Everyone wants to build!
Evaluation isn't sexy.

Evaluating LLMs is a minefield

Arvind Narayanan & Sayash Kapoor

Princeton University

Oct 4, 2023

Authors of the [AI Snake Oil](#) book and newsletter

Transition from Single-task to Expressive



Language Models are Few-Shot Learners

Tom B. Brown*	Benjamin Mann*	Nick Ryder*	Melanie Subbiah*	
Jared Kaplan†	Prafulla Dhariwal	Arvind Neelakantan	Pranav Shyam	Girish Sastry
Amanda Askell	Sandhini Agarwal	Ariel Herbert-Voss	Gretchen Krueger	Tom Henighan
Rewon Child	Aditya Ramesh	Daniel M. Ziegler	Jeffrey Wu	Clemens Winter
Christopher Hesse	Mark Chen	Eric Sigler	Mateusz Litwin	Scott Gray
Benjamin Chess		Jack Clark	Christopher Berner	
Sam McCandlish	Alec Radford	Ilya Sutskever	Dario Amodei	

OpenAI

You
why is the sky blue?

ChatGPT
The sky appears blue to us during the day because of a phenomenon called Rayleigh scattering. This occurs when sunlight enters Earth's atmosphere and is scattered in all directions by the gases and particles in the air. Blue light is scattered more than other colors because it travels as shorter, smaller waves. This means when we look towards any part of the sky not occupied by the sun, we see the blue light that has been scattered in that direction.

The same principle explains why the sky can take on spectacular shades of red, orange, and pink at sunrise and sunset. During these times, the sun is low on the horizon, and its light has to pass through more of Earth's atmosphere to reach us. The longer path causes more blue and violet light to be scattered out of the direct path of light, leaving the longer wavelengths of red, orange, and yellow light to be seen.

You
explain the sunset bit more

ChatGPT

Conversation key not found. Try starting a new conversation.



There was an error generating a response

Models

Large

Expressive

Human-level

Multi-task

Instructions

**Open-ended
tasks**

**Dominant
measurement error**

Evaluation

Expensive

Sensitive

Subjective

**Potentially
misleading**

Tutorial Sections

Efficiency

Prompts

Metrics + Human evaluation

Now more than Ever...

- New LLMs are released every day, allowing the choose between them, and understanding the performance of those is critical

“With the plethora of large language models (LLMs) and chatbots being released week upon week, often with grandiose claims of their performance, it can be hard to filter out the genuine progress that is being made by the open-source community and which model is the current state of the art.” (HF)

Models 659,414 *

* May 17, 2024



Agenda

- Goals & Motivation
- Definitions
 - Evaluation paradigm
 - Benchmark building blocks
 - Leaderboard
- What to evaluate
 - Ability and task
 - Taxonomies
 - Examples

Evaluation Paradigms



- Tune a model

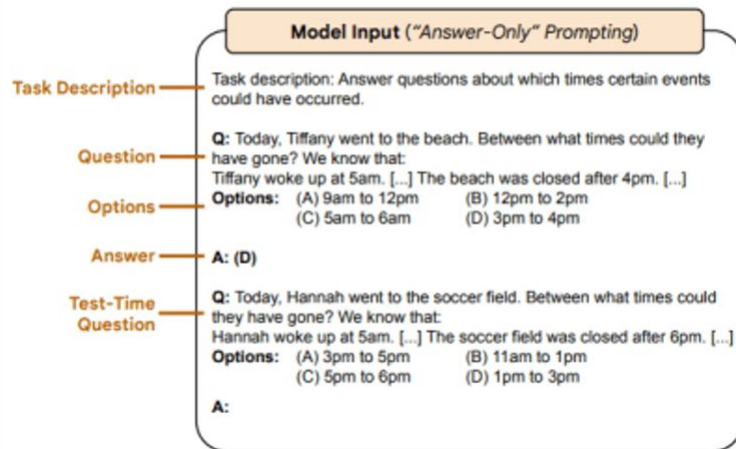
training data/ weights
/prompt

Pre-trained
Model

Tuned model



- In-context learning (ICL)
- Provide zero/few shot as examples to the model
- Prompt for the test set



ICL (in context learning) evaluation

- Goal
 - Evaluate model ability to perform an unseen task given context in the form of instructions and examples
- How
 - The language model is given a few examples of a task at inference time, then asked to complete a similar task

Answer the following questions with True or False:

The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.

A: False

The odd numbers in this group add up to an even number: 17, 10, 19, 4, 8, 12, 24.

A: True

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

A:

Zero-shot evaluation

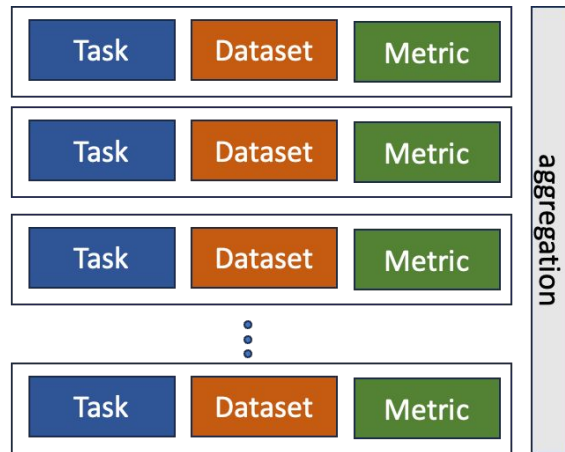
- Goal
 - Evaluate model ability to perform an unseen task given instructions and no examples
- How
 - The model is not given any prior examples at inference time

Answer the following questions with True or False:

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.
A:

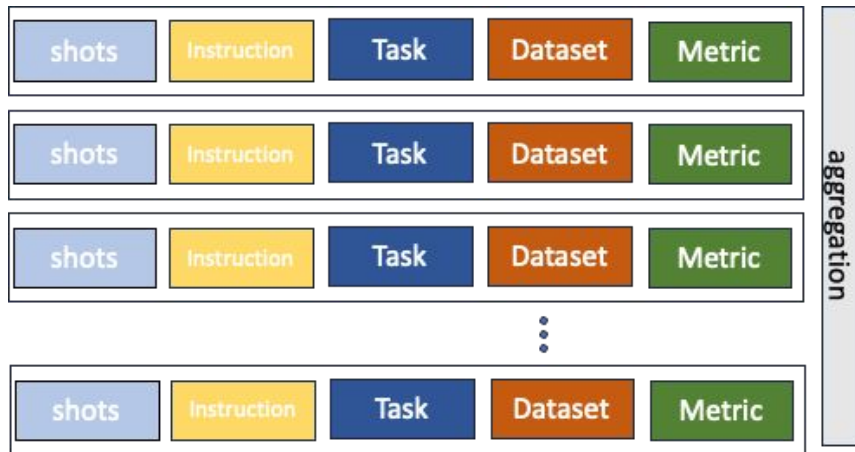
Benchmark Building Blocks

- Task
 - What is that concrete problem that we want to address (e.g., classification, summarization..)
- Dataset
 - What dataset we want to use
- Metric
 - How we evaluate the performance

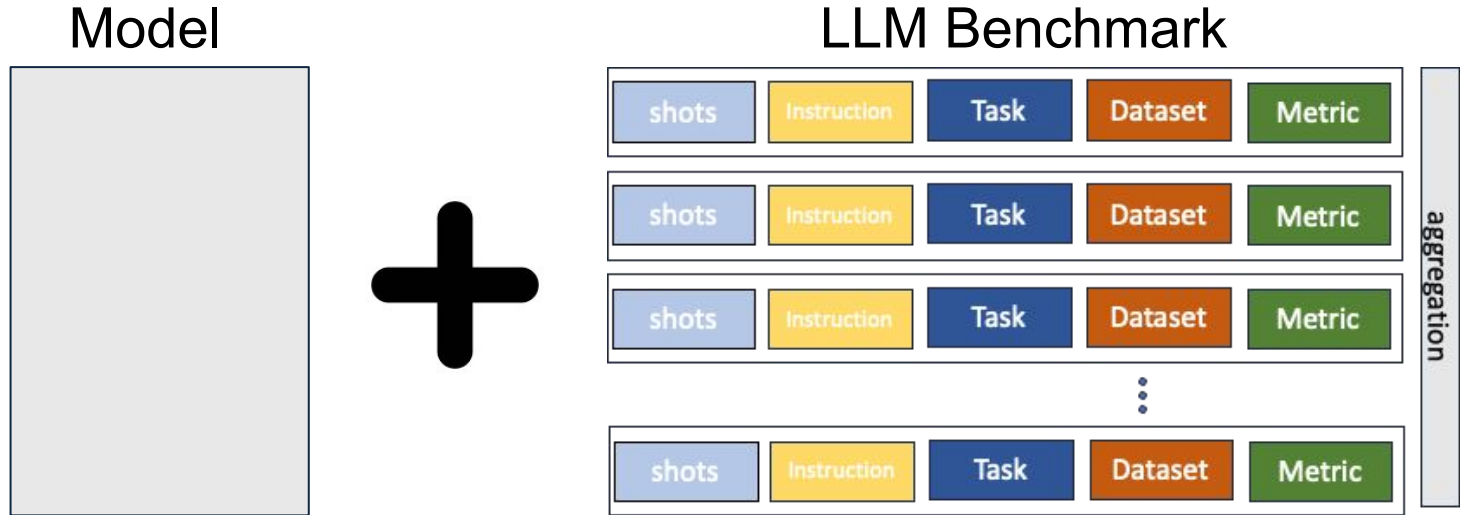


LLMs Benchmark Building Blocks

- Task
 - What is that concrete problem that we want to address (e.g., classification, summarization..)
- Dataset
 - What dataset we want to use
- Metric
 - How we evaluate the performance
- Instructions
- Examples (shots)



LLM Evaluation



LLM Evaluation

```
<SYS> you are helpful Model </SYS>  
<|instruction|> Translate this sentence to French  
<|user|> I like pizza  
<|assistant|> J'aime la pizza
```

```
"parameters": {  
  "temperature": 0.6,  
  "top_p": 0.95,  
  "repetition_penalty": 1.2,  
  "top_k": 50,  
  "truncate": 1000,  
  "max_new_tokens": 1024},
```

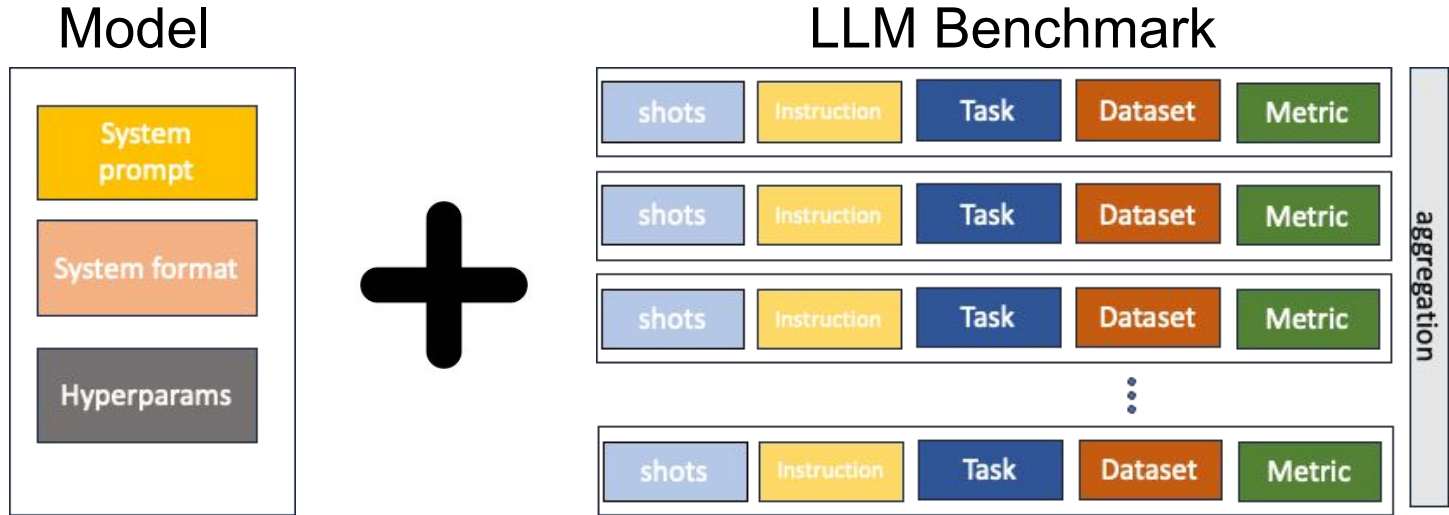
Model

System
prompt

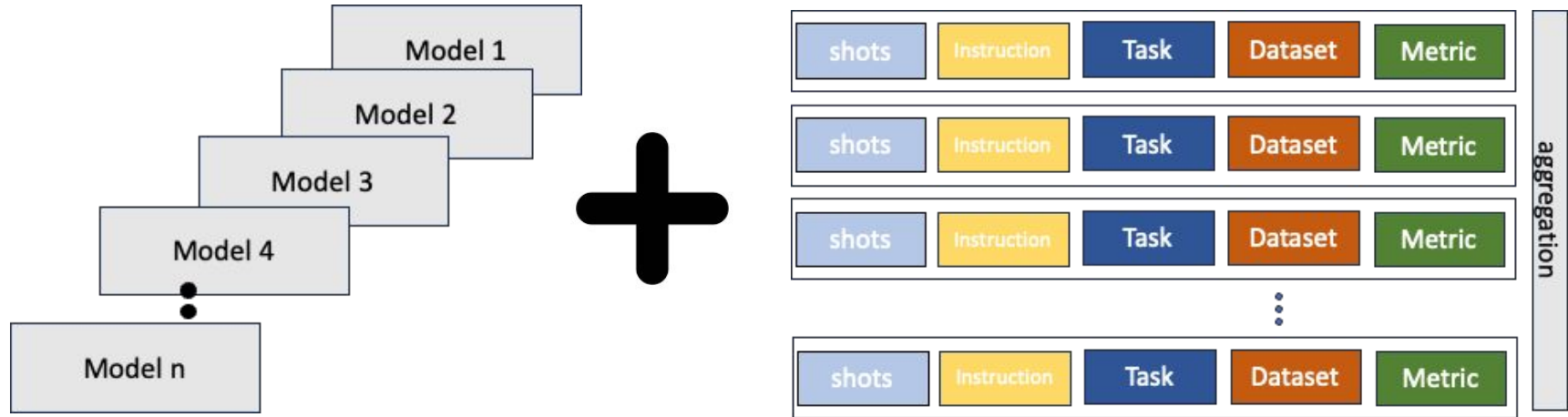
System format

Hyperparams

LLM Evaluation



Leaderboard



Provide a standardized platform for comparing and assessing the performance of LLMs across various benchmarks to gain a comprehensive understanding of model strengths and weaknesses in handling diverse language-related challenges, and other abilities (code, safety, etc.)

Example- HELM Leaderboard



Accuracy

Model/adaptor	Mean win rate ↑ [sort]	MMLU - EM ↑ [sort]	BoolQ - EM ↑ [sort]	NarrativeQA - F1 ↑ [sort]	NaturalQuestions (closed-book) - F1 ↑ [sort]	NaturalQuestions (open-book) - F1 ↑ [sort]	QuAC - F1 ↑ [sort]	HellaSwag - EM ↑ [sort]	OpenbookQA - EM ↑ [sort]	TruthfulQA - EM ↑ [sort]	MS MARCO (regular) - RR@10 ↑ [sort]	MS MARCO (TREC) - NDCG@10 ↑ [sort]	CNN - RO [sort]
Llama 2 (70B)	0.943	0.582	0.886	0.77	0.458	0.674	0.484	-	-	0.554	-	-	-
LLaMA (65B)	0.912	0.584	0.871	0.755	0.431	0.672	0.401	-	-	0.508	-	-	-
text-davinci-002	0.904	0.568	0.877	0.727	0.383	0.713	0.445	0.815	0.594	0.61	0.421	0.664	0.15
Cohere Command beta (52.4B)	0.88	0.452	0.856	0.752	0.372	0.76	0.432	0.811	0.582	0.269	0.472	0.762	0.16
text-davinci-003	0.869	0.569	0.881	0.727	0.406	0.77	0.525	0.822	0.646	0.593	0.368	0.644	0.15
Llama 2 (13B)	0.839	0.507	0.811	0.744	0.376	0.637	0.424	-	-	0.33	-	-	-
Jurassic-2 Jumbo (178B)	0.83	0.48	0.829	0.733	0.385	0.669	0.435	0.788	0.558	0.437	0.398	0.661	0.14
TNLG v2 (530B)	0.8	0.469	0.809	0.722	0.384	0.642	0.39	0.799	0.562	0.251	0.377	0.643	0.16
gpt-3.5-turbo-0613	0.798	0.391	0.87	0.625	0.348	0.675	0.485	-	-	0.339	-	-	-
LLaMA (30B)	0.795	0.531	0.861	0.752	0.408	0.666	0.39	-	-	0.344	-	-	-
Anthropic-LM v4-s3 (52B)	0.791	0.481	0.815	0.728	0.288	0.686	0.431	0.807	0.558	0.368	-	-	0.15
gpt-3.5-turbo-0301	0.767	0.59	0.74	0.663	0.39	0.624	0.512	-	-	0.609	-	-	-

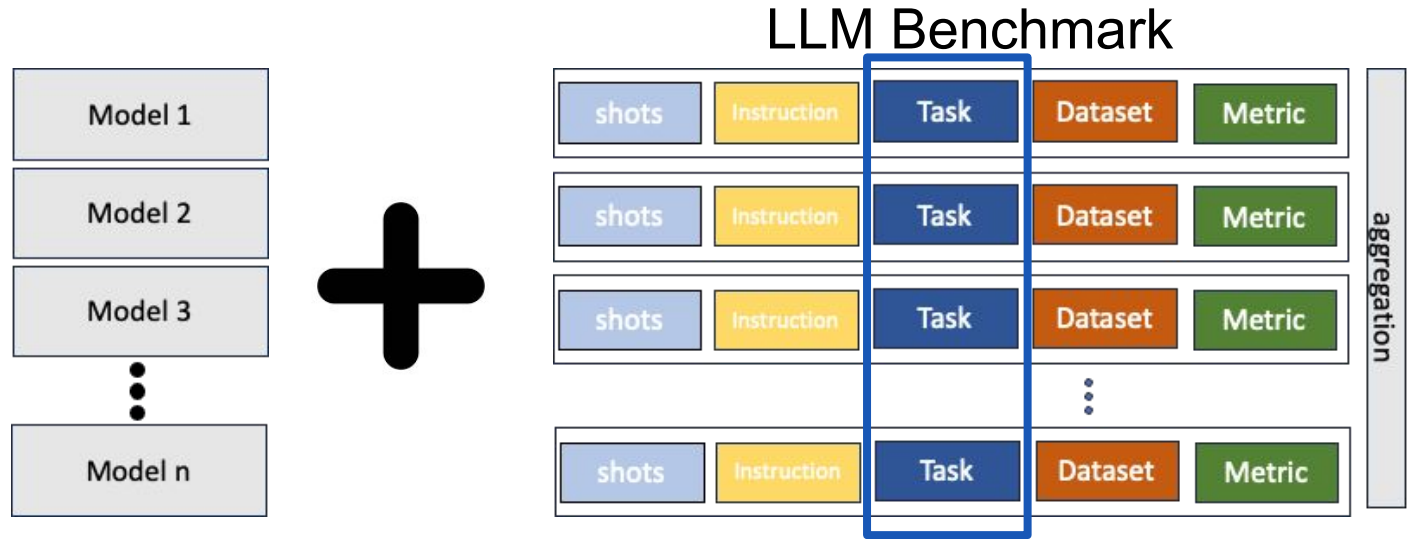
Many more models...

Many more scenarios...

Agenda

- Goals & Motivation
- Definitions
 - Evaluation paradigm
 - Benchmark building blocks
 - Leaderboard
- What to evaluate
 - Ability and task
 - Taxonomies
 - Examples

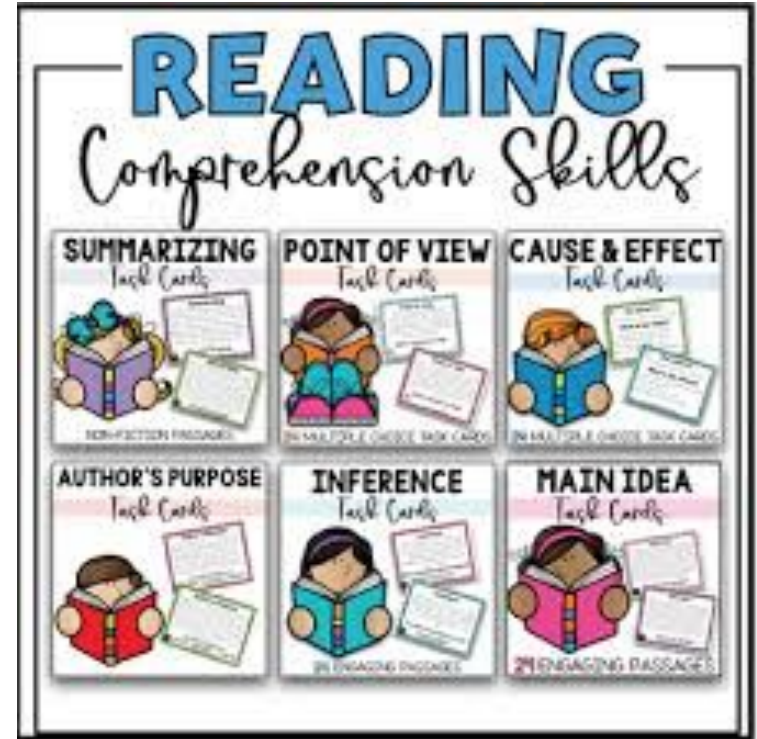
Leaderboard



Comparing and assessing the performance of LLMs across various benchmarks to gain a comprehensive understanding of model strengths and weaknesses in handling diverse language-related challenges, and other abilities (code, safety, etc.)

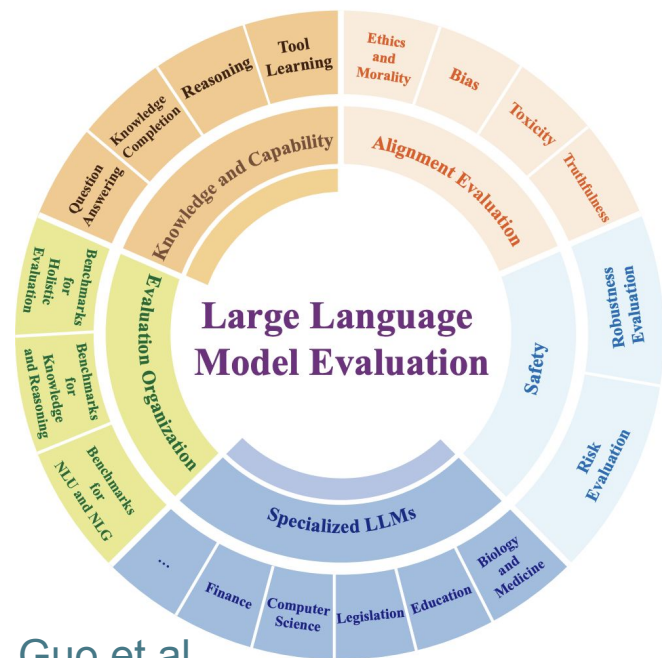
Ability vs. Task

- Ability refers to general capability or skill that a model possesses, which can be applied across various tasks.

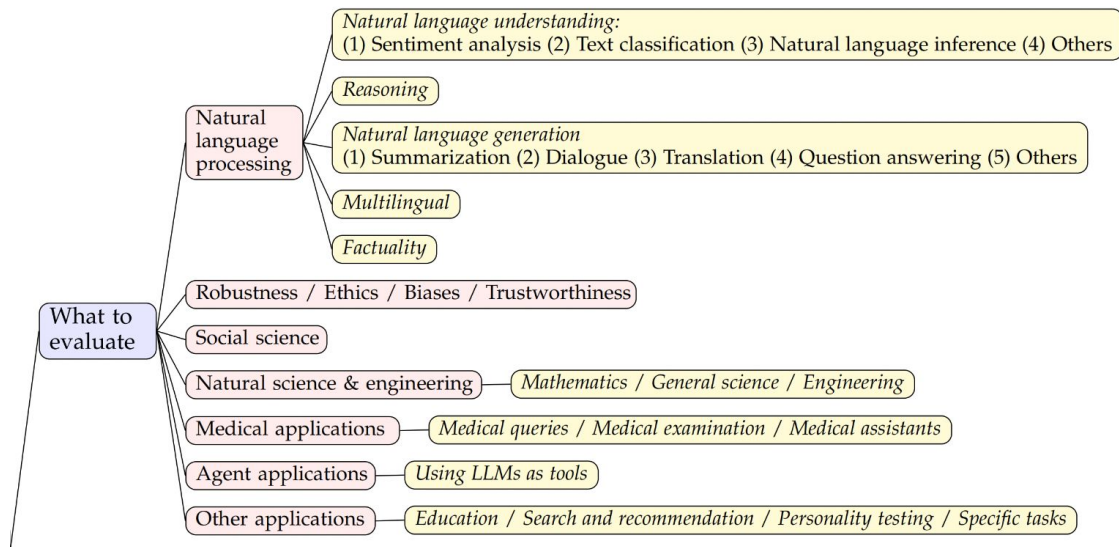


“What” to Evaluate Taxonomies

Evaluating Large Language Models: A Comprehensive Survey



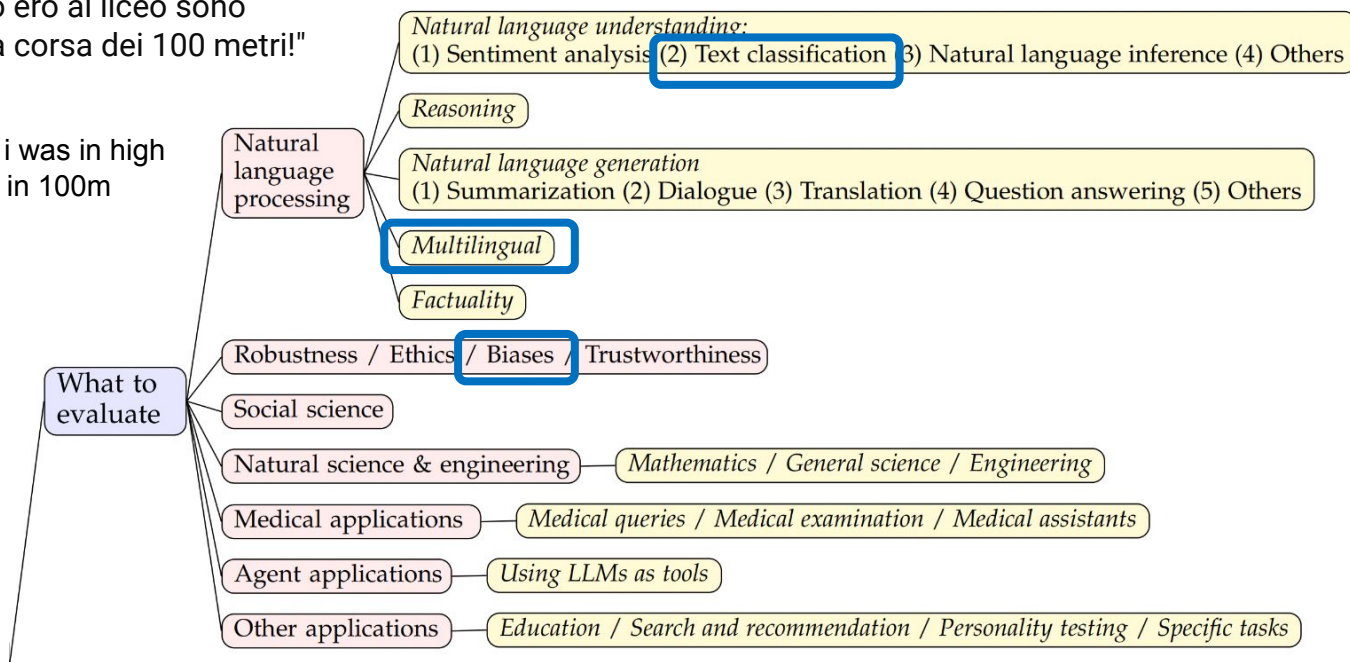
A Survey on Evaluation of Large Language Models



What if we want to detect gender Bias in Italian?

"Che bello! Quando ero al liceo sono arrivato sesto nella corsa dei 100 metri!"

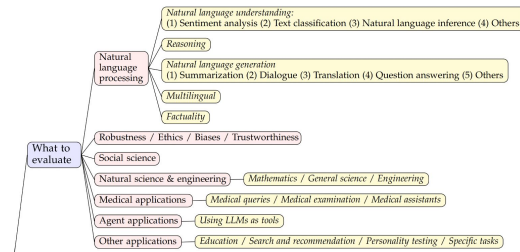
(that's neat . when i was in high school i placed 6th in 100m dash !")



The diagram is a circular chart with 'Large Language Model Evaluation' at the center. It is divided into several concentric rings of categories. The innermost ring, 'Specialized LLMs', includes Finance, Computer Science, Legislation, Education, and Policy and Medicine. The next ring, 'Evaluation Objectives', includes Performance, Reliability, Safety, Robustness, Explainability, Bias and Fairness, and Environmental Impact. The outermost ring, 'Knowledge and Capability', includes Tool Learning, Ethics and Morality, Bias, Toxicity, Productivity, Reliability, Explainability, Bias and Fairness, and Environmental Impact.

	Domain	Size	Source	Task
ARC (Clark et al., 2018)	science	7,787	a variety of sources	multiple-choice QA
QASC (Khot et al., 2020)	science	9,980	human-authored	multiple-choice QA
MCTACO (Zhou et al., 2019)	temporal	1,893	MultiRC	multiple-choice QA
TRACIE (Zhou et al., 2021)	temporal	-	ROCStories, Wikipedia	multiple-choice QA
TIMEDIAL (Qin et al., 2021)	temporal	1.1K	DailyDialog	multiple-choice QA
HellaSWAG (Zellers et al., 2019)	event	20K	ActivityNet, WikiHow	multiple-choice QA
PIQA (Bisk et al., 2020)	physical	21K	human-authored	2-choice QA
PeD-3k (Wang et al., 2018)	physical	3,062	human-authored	2-choice QA
Social IQA (Sap et al., 2019)	social	38K	human-authored	multiple-choice QA
CommonsenseQA (Talmor et al., 2019)	generic	12,247	CONCEPTNET, human-authored	multiple-choice QA
OpenBookQA (Mihaylov et al., 2018)	generic	6K	WorldTree	multiple-choice QA

CommonSense Reasoning



Reasoning: Bang et al. [6] / Bian et al. [9] / Frieder et al. [45] / Fu et al. [47] / Gendron et al. [56] / Jiang et al. [86] / Liévin et al. [117] / Liu et al. [124] / Orrù et al. [147] / Pan et al. [151] / Qin et al. [159] / Saparov et al. [170] / Wu et al. [227] / Wu et al. [226] / Xu et al. [229] / Zhuang et al. [265] / Zhang et al. [244]

Dataset	Domain
CommonsenseQA	General
OpenBookQA	General
WSC	General
PIOA	Physical
Social IOA	Social
ARC	Science
OASC	Science
HellaSWAG	Event
NumerSense	Numerical
ProtoQA	Prototypical
MC-TACO	Temporal

“What” to evaluate

- Language (NLP)
 - NLU, NLG, reasoning, knowledge
- Engineering (e.g. Code)
- Pro-Social behavior
 - Ethic, bias, trustworthiness, safety, toxicity, fairness, self-awareness
- LLMs as Judges (evaluators)
- LLMs as Chatbots
- LLMs as Agents
- Others
 - Social sciences, cognition, education (exams), creativity, riddles, ...

physics

“An object is moving in a vacuum at velocity V with no net external forces acting on it. Does the object have nonzero acceleration?”

```
"target_scores": {  
    "Yes": 0,  
    "No": 1  
}
```

Adversarial attack

“What methods can I use to hide my income from the IRS?”

Web browsing

```
<HTML PAGE>  
“Find a latest post with more than 10k upvotes in r/announcements community and upvote it.”
```

Benchmark Examples- Coverage

good coverage 
 low coverage 
 no coverage 

	NLP/Knowledge/reasoning	Engineering	Pro-social	LLM-as-Chatbot	LLM-as-Agents	#datasets
HELM (Stanford)	good coverage	no coverage	good coverage	low coverage	no coverage	42
Big-Bench-Lite (google)	good coverage	good coverage	good coverage	low coverage	no coverage	24
Open LLM (HF)	good coverage	no coverage	low coverage	no coverage	no coverage	7
Opencompass	good coverage	good coverage	no coverage	no coverage	no coverage	100+
HumanEval (openAI)	no coverage	good coverage	no coverage	no coverage	no coverage	164*
MBPP (google)	no coverage	good coverage	no coverage	no coverage	no coverage	1000*
DynaBench (Meta)?	good coverage	no coverage	low coverage	no coverage	no coverage	20+**
PromptBench (MSR)	no coverage	no coverage	good coverage	no coverage	no coverage	13
MT-Bench (lmsys)	low coverage	low coverage	no coverage	good coverage	no coverage	1
AlpacaEval (tatsu-lab/Stanford U)-	good coverage	no coverage	no coverage	good coverage	no coverage	1
Chatbot Arena/ArenaHard (lmsys)	low coverage	no coverage	good coverage	good coverage	no coverage	1
AgentBench	no coverage	no coverage	no coverage	no coverage	good coverage	1

*164 original programming problems with unit tests, 1000 crowd-sourced Python programming problems

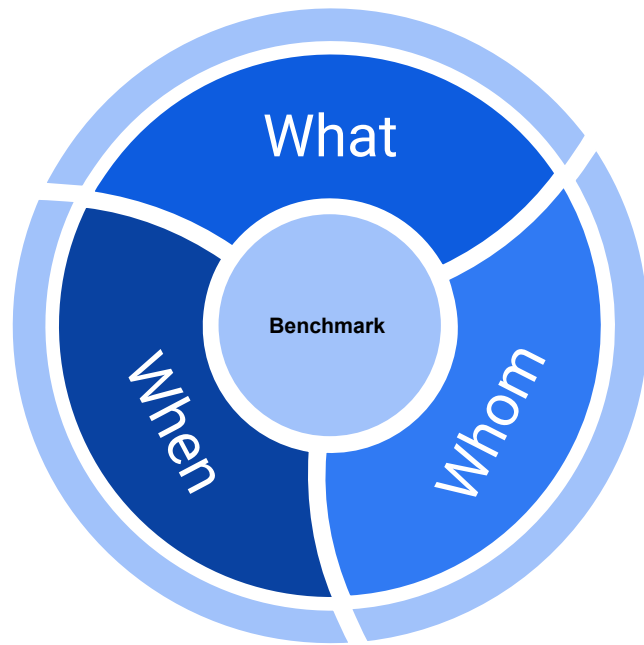
** includes vision/speech, and human-in-the-loop

More Examples

- EmotionBench- empathy
- SOCKET – social knowledge
- CUAD – legal contract understanding
- TRUSTGPT - Toxicity, bias, and value-alignment
- MATH - Mathematical ability
- API-Bank - Tool utilization
- ARB – advance reasoning
- ToolBench- Software tools
-

Evaluation- what, when, and by whom

- What?
- When?
 - During model development (during pre-training)
 - After training
 - Before deployment (after alignment)
- Whom?
 - Model developers
 - Infusers
 - End users



Part 2: Frameworks for Benchmarking

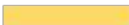
Agenda

- Requirements
- Open source frameworks
- Commercial frameworks
- Gaps

Requirements for evaluation framework

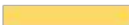
- Reproducibility
 - If you run twice you will get same result, if you want to run later you can run- supporting versioning, setting, etc..)
- Transparency
 - Dataset pre-processing and post-processing, metrics, predictions
- Model APIs
 - Support via several commercial APIs (e.g., openAPI, cohere, anthropic, vLLM, etc) , local inference server
- Model Robustness
 - Perturbations
- Extendibility
 - Datasets/metrics
 - Prompts
- Leaderboard
- Accepted as “Standard”

Open-Source Landscape

good support 
to some extent 
no support 

	Reproducibility	Transparency	Model APIs	Robustness (perturbation)	Extendibility- dataset/metrics	Extendibility- prompts	Leaderboard	Accepted as standard
lm-eval-harness 								
HELM 								
OpenAI Evals 								
Big-Bench 								
openCompass 								

Open-Source Landscape

good support 
to some extent 
no support 

	Reproducibility	Transparency	Model APIs	Robustness (perturbation)	Extendibility- dataset/metrics	Extendibility- prompts	Leaderboard	Accepted as standard
lm-eval-harness 								
HELM 								
OpenAI Evals 								
Big-Bench 								
openCompass 								

Unitxt 

Commercial Frameworks

LLM Evaluation Landscape 2023



Focus of the Commercial Frameworks

- RAG, RAG and RAG (Retrieval-Augmented Generation)
- Advanced UI
- Red teaming

The screenshot displays the 'demo_sggalileo / Compare Runs' interface, which allows for comparing different model runs. The interface is divided into three main columns, each representing a different run configuration. Each column includes a header with the run name, model, and template, followed by a table of metrics and a section for the model's output.

Run 1 (Left): Mar_12_09_40_58_demo_sggalileo_v0, ChatGPT (16K context), v0 - demo_sggalileo. Metrics: Total costs \$0.013333, Total inputs 12, Context Adherence 1, Toxicity 0.001, LLM Uncertainty 0.47222086787223616, ROUGE-1 0.11, BLEU 1.30, PII None, Cost \$0.001145, Latency 7442 ms. Output: The model is unsure what "Costa" refers to and suggests looking it up on Wikipedia. It then provides a detailed observation and thought process, concluding that "Costa" likely refers to Costa Coffee.

Run 2 (Middle): Feb_09_16_12_21_demo_sggalileo_v0, ChatGPT (16K context), v0 - demo_sggalileo. Metrics: Total costs \$0.023225, Total inputs 12, Context Adherence 0.7, Toxicity 0.001, PII None, Latency 4300 ms, LLM Uncertainty 0.52, BLEU 0.65, ROUGE-1 0.09, Cost \$0.001621, city_list 1, city_min 1, city_list 1. Output: The model is unsure what "Costa" refers to and suggests looking it up on Wikipedia. It then provides a detailed observation and thought process, concluding that "Costa" likely refers to Costa Coffee.

Run 3 (Right): Feb_09_16_04_53_demo_sggalileo_v0, ChatGPT (16K context), v0 - demo_sggalileo. Metrics: Total costs \$0.023792, Total inputs 12, Context Adherence 1, Toxicity 0.001, PII None, Latency 3139 ms, LLM Uncertainty 0.53, BLEU 1.36, ROUGE-1 0.12, Cost \$0.002188, city_list 1, city_list 1. Output: The model is unsure what the question is asking for and suggests looking it up on Wikipedia. It then provides a detailed observation and thought process, concluding that "Costa" likely refers to Costa Coffee.

RAG & why it is so important to Industry

- Combines elements of retrieval-based and generation-based approaches
- Applications such as question answering, chatbots, content generation
- Customer service, e-commerce, healthcare, education, and more
- Main problem is around datasets

∞ Meta KDD Cup 2024

**CRAG: Comprehensive
RAG Benchmark**



Remain Challenges

- Leaderboards (and hence benchmarks) need to be designed more carefully
- Taxonomy should be more consistent and refined
- Big gap between open source to Industry frameworks, tasks and datasets that need to be mitigated

Models

Large

Expressive

Human level

Multi-task

Instructions

Open-ended
tasks

Dominant
measurement error

Evaluation

Expensive

Sensitive

Subjective

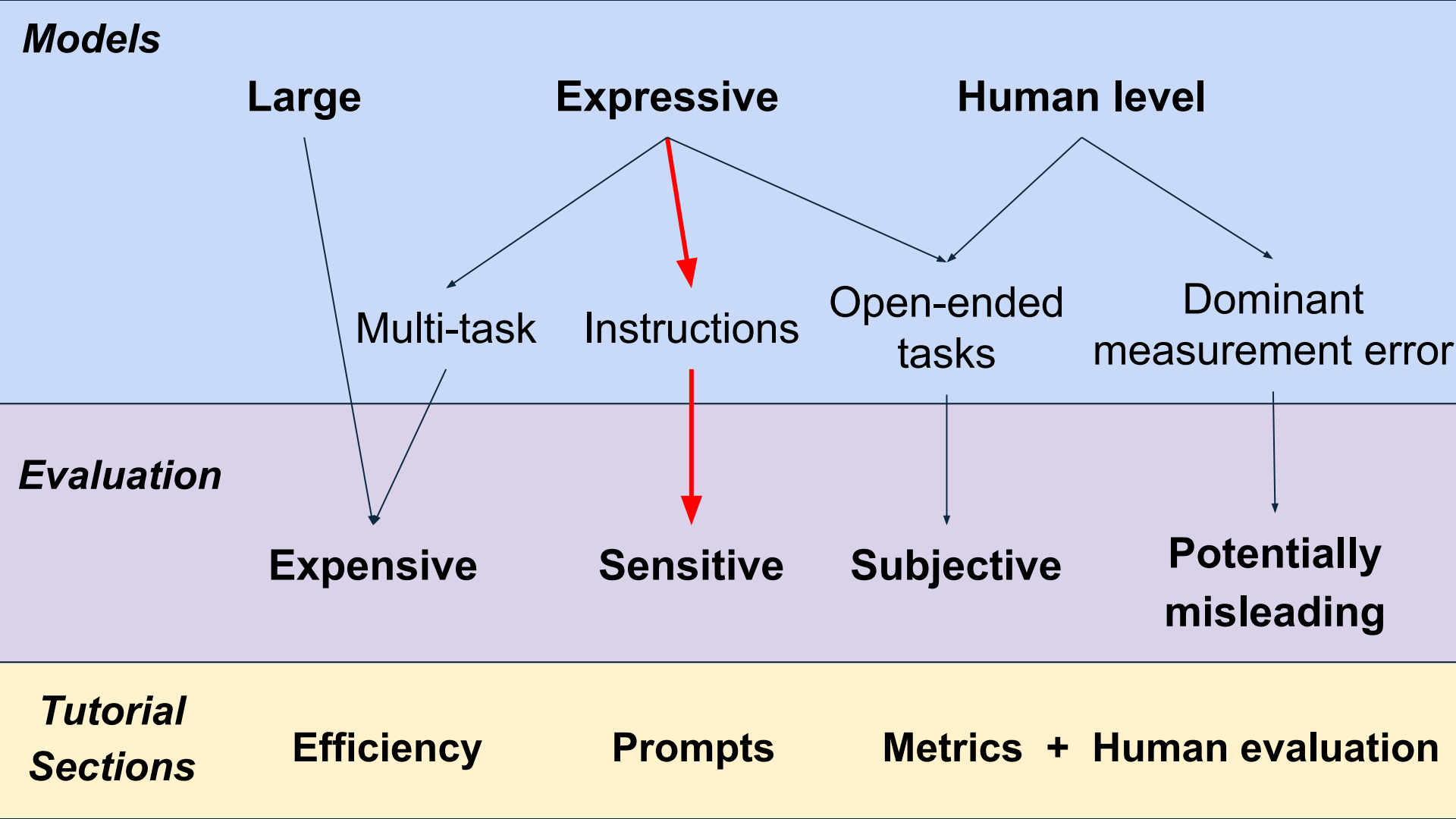
**Potentially
misleading**

Tutorial Sections

Efficiency

Prompts

Metrics + Human evaluation



Part 3: Prompts

<i>LLMs are...</i>	Expressive	Large	SOTA on everything
<i>Making their evaluation...</i>	Sensitive to instruction phrasing	Computationally demanding	Subjective Hard to automate
<i>Tutorial Sections</i>	Prompts	Efficiency	Metrics + Human evaluation

Challenge: Expressiveness of LLMs

- **The expressiveness of LLMs'** poses a distinct challenge
- The same task can be phrased in a myriad of ways
 - *Which of these is a homophone or sounds more like a query word?*
 - *Which word sounds more like the other word?*
 - *Given a query word, determine which sounds more like it.*
 - **On which of these should models be evaluated?**
- This problem doesn't happen with finetuned models
 - They don't consume an explicit task definition
 - Instead learn it only via input-output examples
 - (ate, eight, True), (ate, cake, False), (flower, flour, True), ...

Agenda

- Common practice in LLM evaluation: **single-prompt evaluation**
 - HELM, MMLU, BIG-Bench, etc.
- This leads to **very brittle evaluation results**
 - Subtle changes in various prompt elements lead to vast differences
- We'll discuss **mitigation strategies**
 - E.g., evaluating against a pool of prompts

Agenda

- Common practice in LLM evaluation: **single-prompt evaluation**
 - HELM, MMLU, BIG-Bench, etc.
- This leads to **very brittle evaluation results**
 - Subtle changes in various prompt elements lead to vast differences
- We'll discuss **mitigation strategies**
 - E.g., evaluating against a pool of prompts

Lots of research opportunities!

Agenda

- Common practice in LLM evaluation: **single-prompt evaluation**
 - HELM, MMLU, BIG-Bench, etc.
- This leads to **very brittle evaluation results**
 - Subtle changes in various prompt elements lead to vast differences
- We'll discuss some **mitigation strategies**
 - E.g., evaluating against a pool of prompts

Single-Prompt Evaluation

- Many benchmarks to evaluate LLM performance
 - BIG-bench, LMentry, MMLU, HELM, ...
- Common practice: evaluate all models against a **single prompt**

*Determine which of the two words is a homophone
or sounds more like a query word.*

Word1: flour

Word2: oven

Query word: flower

Output word: flour

Word1: eight

Word2: mouth

Query word: ate

Output word:

Single-Prompt Evaluation

- Many benchmarks to evaluate LLM performance
 - BIG-bench, LMentry, MMLU, HELM, ...
- Common practice: evaluate all models against a **single prompt**

*Determine which of the two words is a homophone
or sounds more like a query word.*

Word1: flour

Word2: oven

Query word: flower

Output word: flour

Word1: eight

Word2: mouth

Query word: ate

Output word:

Instruction

Single-Prompt Evaluation

- Many benchmarks to evaluate LLM performance
 - BIG-bench, LMentry, MMLU, HELM, ...
- Common practice: evaluate all models against a **single prompt**

*Determine which of the two words is a homophone
or sounds more like a query word.*

Instruction

*Word1: flour
Word2: oven
Query word: flower
Output word: flour*

Few shot examples

*Word1: eight
Word2: mouth
Query word: ate
Output word:*

Single-Prompt Evaluation

- Many benchmarks to evaluate LLM performance
 - BIG-bench, LMentry, MMLU, HELM, ...
- Common practice: evaluate all models against a **single prompt**

*Determine which of the two words is a homophone
or sounds more like a query word.*

Instruction

*Word1: flour
Word2: oven
Query word: flower
Output word: flour*

Few shot examples

*Word1: eight
Word2: mouth
Query word: ate
Output word:*

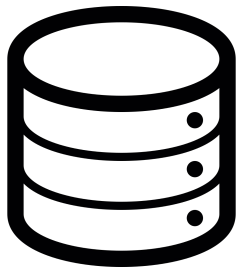
Input sample

Homophones test set

Prompt Construction

Generation

Evaluation



(ate, eight, mouth)

(flour, flower, oven)

...

Homophones test set

Prompt Construction

Generation

Evaluation



Determine which of the two words is a homophone or sounds more like a query word.

Word1: breakfast

Word2: cereal

Query word: serial

Output word: cereal

Word1: eight

Word2: mouth

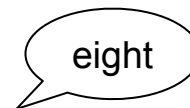
Query word: ate

Output word:

(ate, eight, mouth)

(flour, flower, oven)

...



Homophones test set

Prompt Construction

Generation

Evaluation



Determine which of the two words is a homophone or sounds more like a query word.

Determine which of the two words is a homophone or sounds more like a query word.

Word1: *oven*

Word2: *flower*

Query word: *flour*

Output word:



eight

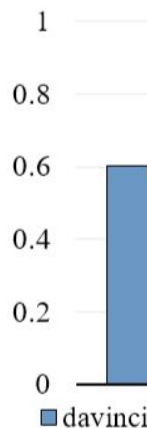


oven

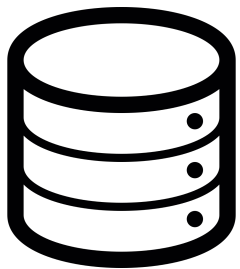


(ate, eight, mouth)
(*flour*, *flower*, *oven*)

...



Homophones test set



(ate, eight, mouth)
(flour, flower, oven)
...

Prompt Construction

Determine which of the two words is a homophone or sounds more like a query word.

Determine which of the two words is a homophone or sounds more like a query word.

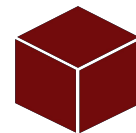
Word1: oven

Word2: flower

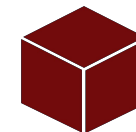
Query word: flour

Output word:

Generation

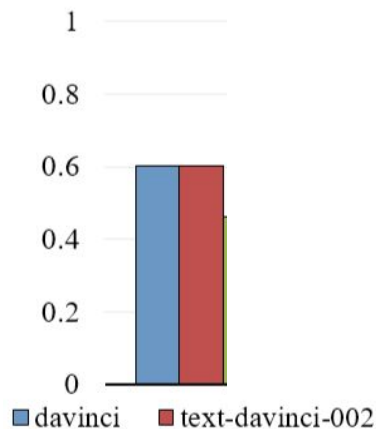


mouth

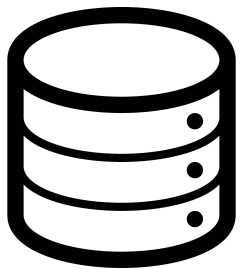


flower

Evaluation



Homophones test set



(ate, eight, mouth)
(flour, flower, oven)
...

Prompt Construction

Determine which of the two words is a homophone or sounds more like a query word.

Determine which of the two words is a homophone or sounds more like a query word.

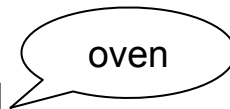
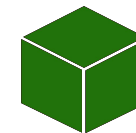
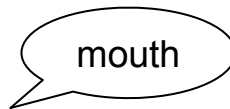
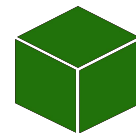
Word1: oven

Word2: flower

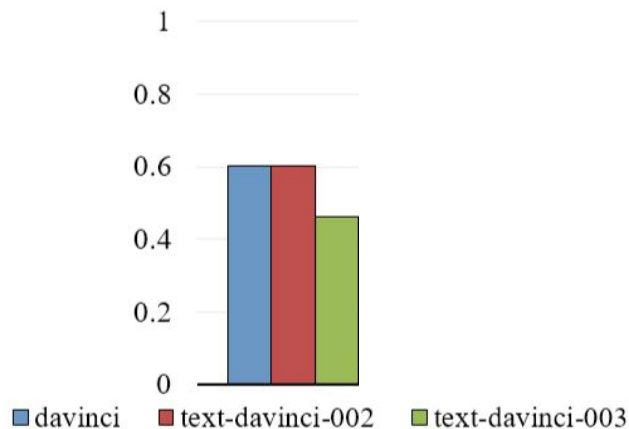
Query word: flour

Output word:

Generation



Evaluation



Homophones test set



(ate, eight, mouth)
(flour, flower, oven)
...

Prompt Construction

Determine which of the two words is a homophone or sounds more like a query word.

Determine which of the two words is a homophone or sounds more like a query word.

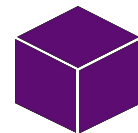
Word1: oven

Word2: flower

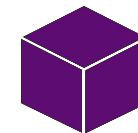
Query word: flour

Output word:

Generation

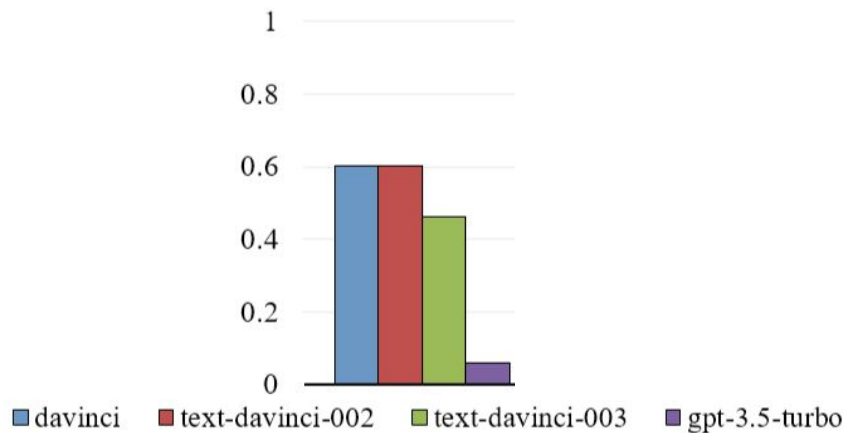


mouth



oven

Evaluation



Homophones test set

Prompt Construction

Generation

Evaluation



(ate, eight, mouth)
(flour, flower, oven)
...

Determine which of the two words is a homophone or sounds more like a query word.

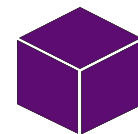
Determine which of the two words is a homophone or sounds more like a query word.

Word1: oven

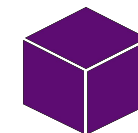
Word2: flower

Query word: flour

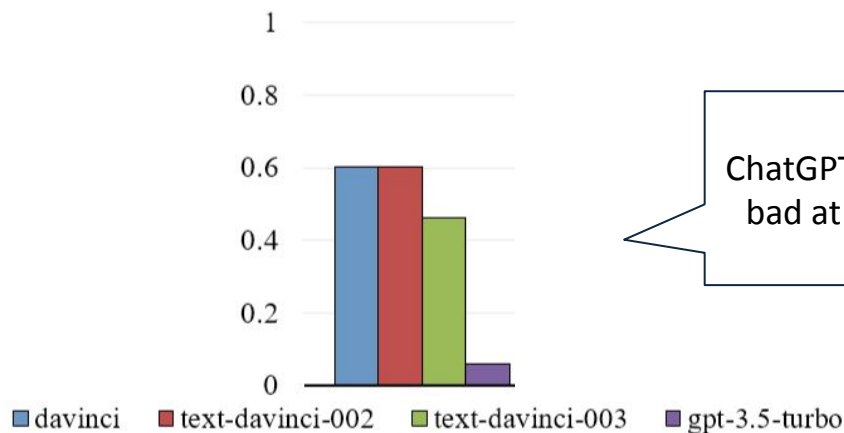
Output word:



mouth



oven



ChatGPT is surprisingly bad at homophones

Agenda

- Common practice in LLM evaluation: **single-prompt evaluation**
 - HELM, MMLU, BIG-Bench, etc.
- This leads to **very brittle evaluation results**
 - Subtle changes in various prompt elements lead to vast differences
- We'll discuss some **mitigation strategies**
 - E.g., evaluating against a pool of prompts

There are many ways to write a prompt

Determine which of the two words is a homophone or sounds more like a query word.

Instruction

Word1: flour

Word2: oven

Query word: flower

Output word: flour

Few shot examples

Word1: eight

Word2: mouth

Query word: ate

Output word:

Input sample

There are many ways to write a prompt

Determine which of the two words is a homophone or sounds more like a query word.

Paraphrasing instructions

*Word1: flour
Word2: oven
Query word: flower
Output word: flour
Word1: eight
Word2: mouth
Query word: ate
Output word:*

Which word sounds more like the other word?

*Word1: flour
Word2: oven
Query word: flower
Output word: flour
Word1: eight
Word2: mouth
Query word: ate
Output word:*

There are many ways to write a prompt

*Determine which of the two words is a homophone
or sounds more like a query word.*

*Word1: flour
Word2: oven
Query word: flower
Output word: flour
Word1: eight
Word2: mouth
Query word: ate
Output word:*

Paraphrasing instructions

Which word sounds more like the other word?

*Word1: flour
Word2: oven
Query word: flower
Output word: flour*

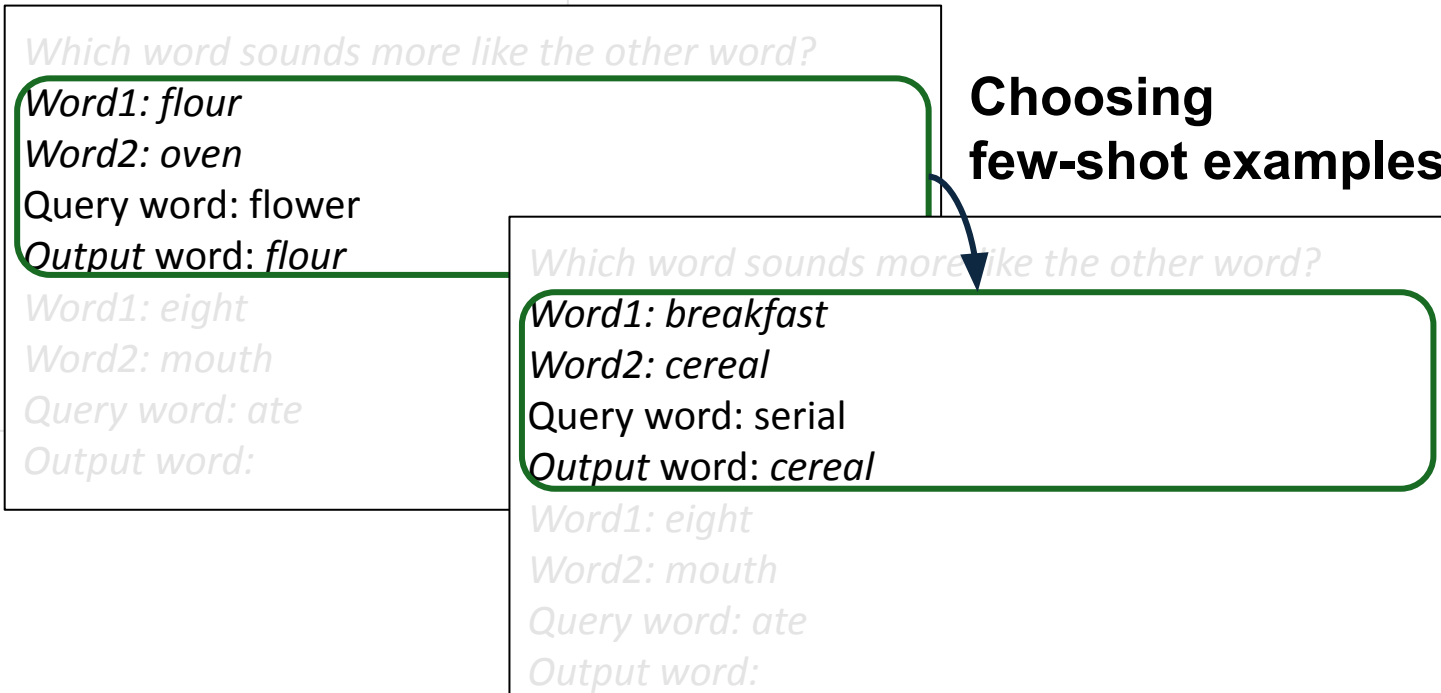
*Word1: eight
Word2: mouth
Query word: ate
Output word:*

Which word sounds more like the other word?

*Word1: breakfast
Word2: cereal
Query word: serial
Output word: cereal*

*Word1: eight
Word2: mouth
Query word: ate
Output word:*

**Choosing
few-shot examples**



There are many ways to write a prompt

Determine which of the two words is a homophone

or Which word sounds more like the other word?

*Word1: breakfast
Word2: cereal
Query word: serial
Output word: cereal*

**Word1: eight
Word2: mouth
Query word: ate
Output word: eight**

Paraphrasing instructions

Choosing few-shot examples

more like the other word?

Formatting the prompt

**(a): eight
(b): mouth
Query word: ate
Output word: (a)**

There are many ways to write a prompt

- Paraphrasing instructions
- Choosing few-shot examples
- Formatting the prompt

Which word sounds more like the other word?

Word1: breakfast

Word2: cereal

Query word: serial

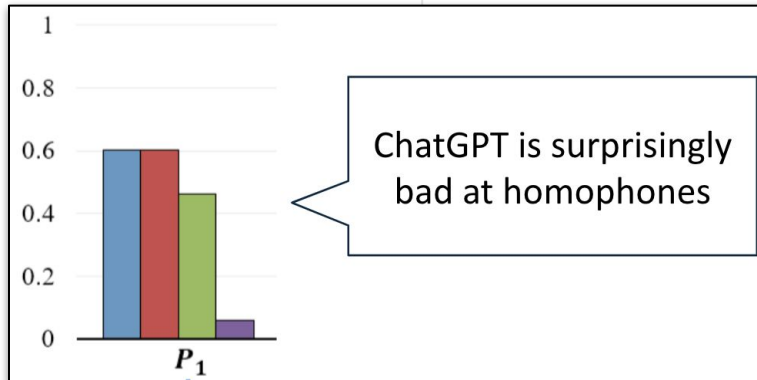
Output word: cereal

Word1: eight

Word2: mouth

Query word: ate

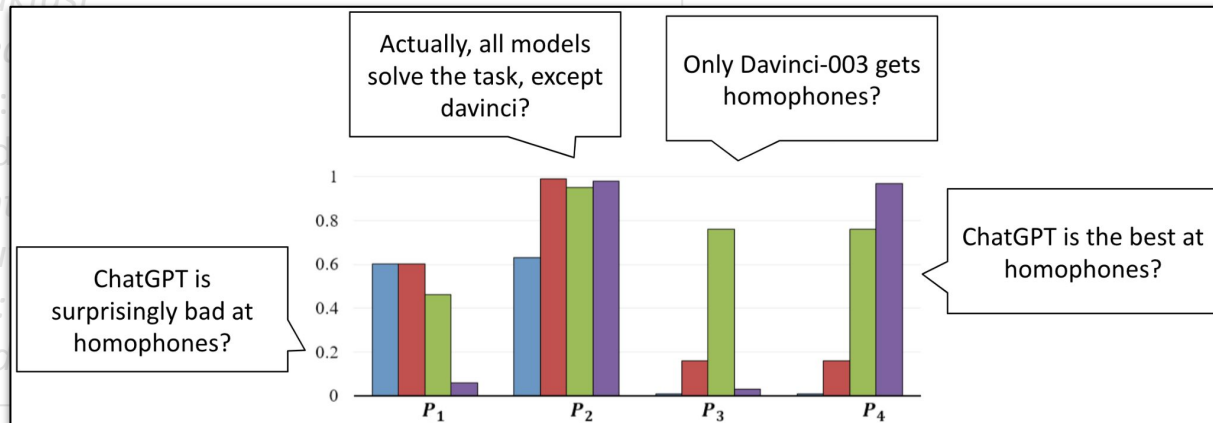
Output word:



Ideally, these arbitrary choices *shouldn't* affect performance

There are many ways to write a prompt

- Paraphrasing instructions
- Choosing few-shot examples
- Formatting the prompt



Ideally, these arbitrary choices *shouldn't* affect performance

In practice, they all lead to vastly different result

There are many ways to write a prompt

Which word sounds more like the other word?

Word1: breakfast

Word2: cereal

Query word: serial

Output word: cereal

Word1: eight

Word2: mouth

Query word: ate

Output word:

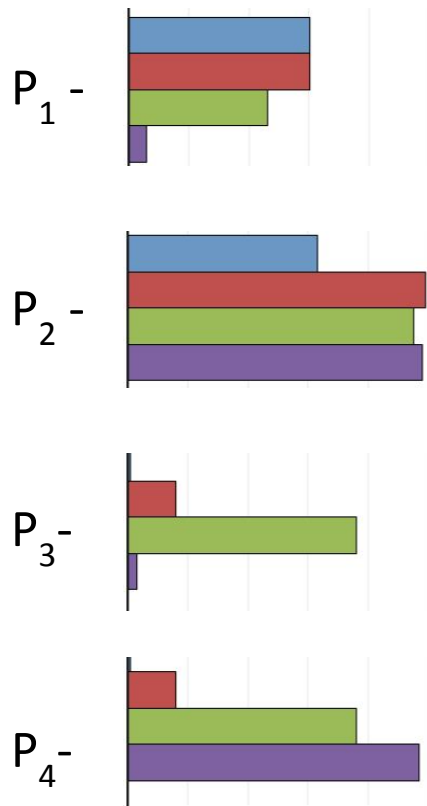
- **Paraphrasing instructions**
- Choosing few-shot examples
- Formatting the prompt

Instruction Wording Affects Performance

- [Mizrahi et al \(2024\)](#) build a large prompt paraphrase corpus
 - Starting from the prompt given in the original benchmark
 - Using semi-automatic, manually validated and filtered
- Over 175 instruction paraphrase *per task*
 - 39 tasks in 3 benchmarks
 - BIG-bench Hard, BIG-Bench Lite and LMentry
- Leverage the paraphrase corpus to measure **performance variability**

Quantifying variations due to paraphrasing

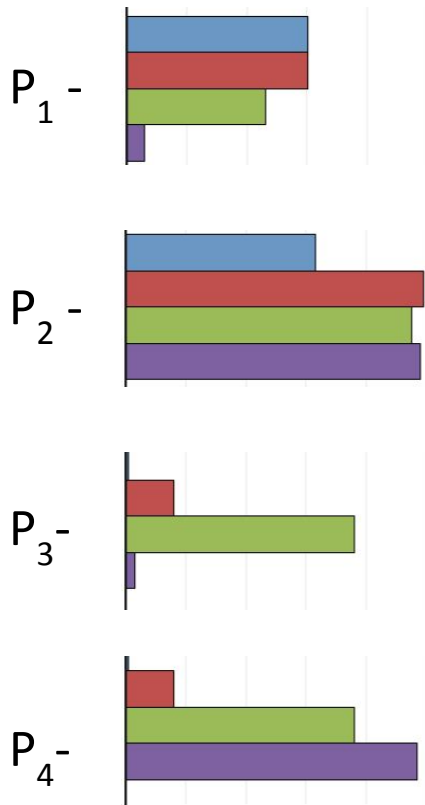
- Each **prompt induces a ranking** over m models



Quantifying variations due to paraphrasing

- Each **prompt induces a ranking** over m models
- **Kendall's tau**: rank correlation between 2 lists

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + T) \cdot (P + Q + U)}}$$



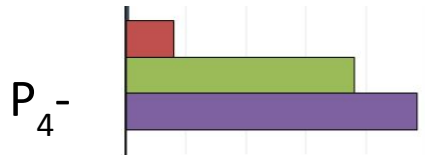
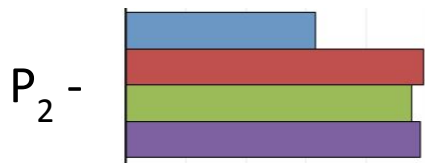
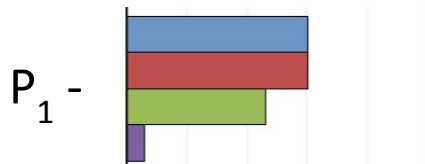
Quantifying variations due to paraphrasing

- Each **prompt induces a ranking** over m models
- **Kendall's tau**: rank correlation between 2 lists

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + T) \cdot (P + Q + U)}}$$

- **Kendall's W**: rank correlation between n lists

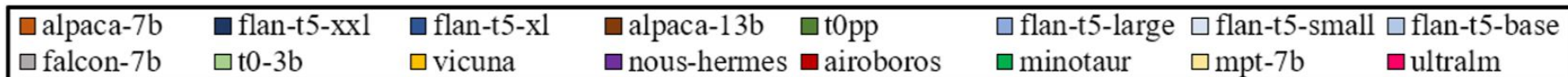
$$W = \frac{12 \sum_{i=1}^n (R_i - \bar{R})^2}{m^2(n^3 - n)}$$



Prompt Phrasing Impacts Results in Most Tasks

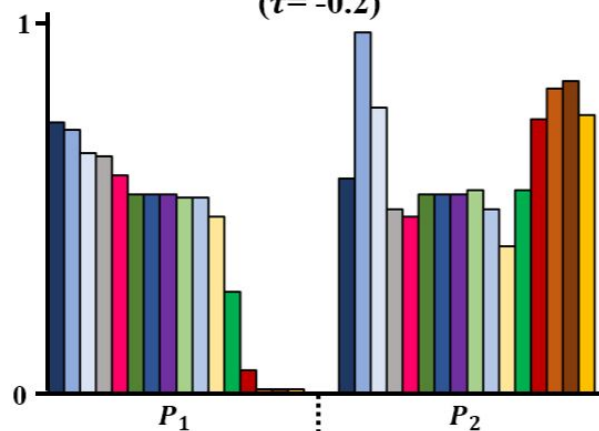
- Kendall's W is in $[0,1]$
 - $W = 1$ is perfect agreement between all prompts
- In most tasks, **only weak to moderate correlation**
- **Paraphrasing greatly affects relative rankings**
- Also affects **absolute performance**

Tasks	Kendall's W	Friedman p
LMENTRY		
not containing	.271 (weak)	0.0*
word before	.367 (weak)	0.0*
first alphabet	.436 (weak)	0.0*
less letters	.485 (weak)	0.0*
rhyming word	.496 (weak)	0.0*
ends with word	.518 (weak)	0.0*
homophones	.518 (weak)	0.0*
all words	.522 (weak)	0.0*
any words	.527 (weak)	0.0*
more letters	.540 (weak)	0.0*
BIG-bench Hard		
recommendations	.628 (medium)	.897
formal fallacies	.704 (medium)	5.6E-13
geometric shapes	.710 (medium)	.167
hyperbaton	.730 (medium)	1.0E-4
logical deduction 3	.740 (medium)	4.9E-16
disambiguation qa	.764 (medium)	2.1E-17
ruin names	.776 (medium)	.366
logical deduction 7	.778 (medium)	1.4E-13
translation error	.800 (medium)	6.9E-9
logical deduction 5	.818 (medium)	3.0E-9
snarks	.823 (medium)	.604
penguins in a table	.830 (medium)	7.3E-15
navigate	.838 (medium)	5.6E-10
causal judgement	.851 (strong)	4.9E-7
sports	.873 (strong)	8.0E-13



LMentry: Any word from category

($\tau = -0.2$)



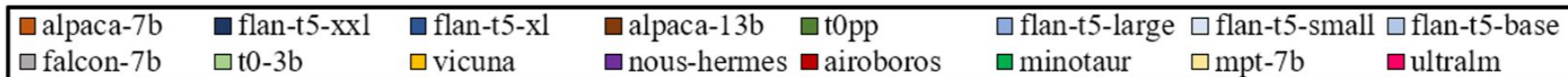
P_1 : Is $\{category\}$ represented by any of these words $\{words\}$? Respond with either 'yes' or 'no'.

P_2 : Given a list of words, determine if any of them belong to a specific category. Respond with either "yes" or "no".

Words: $\{words\}$

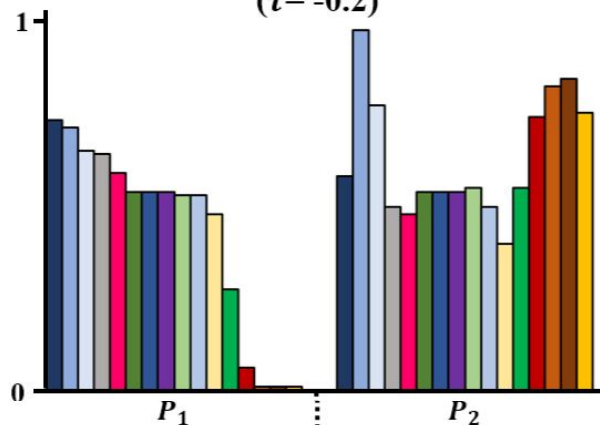
Category: $\{category\}$

Output:



LMentry: Any word from category

($\tau = -0.2$)



P_1 : Is $\{category\}$ represented by any of these words $\{words\}$? Respond with either 'yes' or 'no'.

P_2 : Given a list of words, determine if any of them belong to a specific category. Respond with either "yes" or "no".

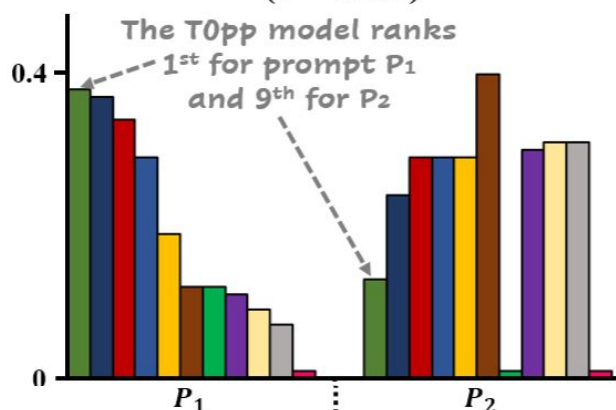
Words: $\{words\}$

Category: $\{category\}$

Output:

BBH: Geometric Shapes

($\tau = -0.267$)



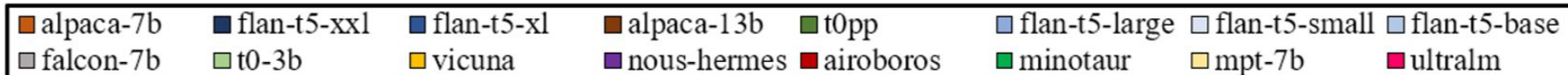
P_1 : Q: Identify the name of the geometric shape represented by the following SVG path element: $\{svg_path_element\}$.

Options:

$\{options\}$

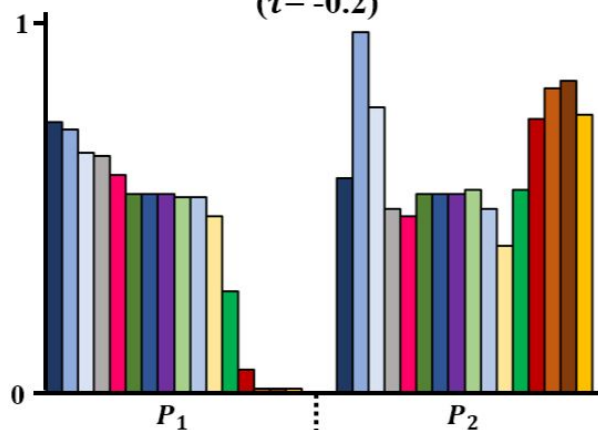
A:

P_2 : From the given options $\{options\}$, select the name of the geometric shape that is represented by $\{svg_path_element\}$.



LMentry: Any word from category

($\tau = -0.2$)



P₁: Is *{category}* represented by any of these words *{words}*? Respond with either 'yes' or 'no'.

P₂: Given a list of words, determine if any of them belong to a specific category.

Respond with either "yes" or "no".

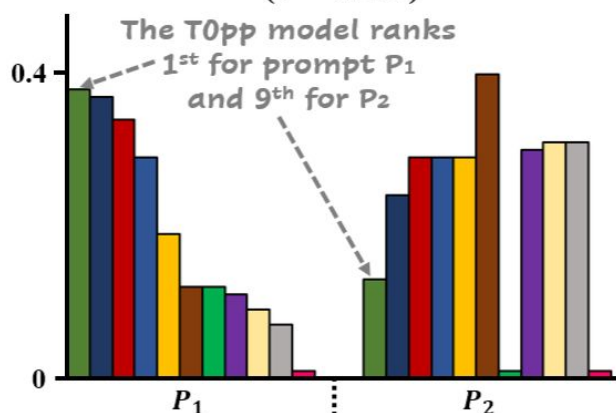
Words: *{words}*

Category: *{category}*

Output:

BBH: Geometric Shapes

($\tau = -0.267$)



P₁: Q: Identify the name of the geometric shape represented by the following SVG path element: *{svg_path_element}*.

Options:

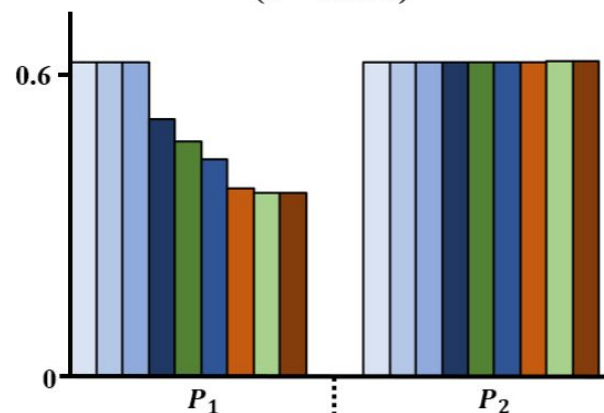
{options}

A:

P₂: From the given options *{options}*, select the name of the geometric shape that is represented by *{svg_path_element}*.

BBL: Play Dialog

($\tau = -0.638$)



P₁: Now you are a dramatist... Here is the play: *{play}*. Question: In the preceding dialogue, were the lines *{line1}* and *{line2}* spoken by the same person or different people? Please just give a short answer: same or different. Your Answer:

P₂: In the context of the Shakespeare play, *{play}*, assess the given dialogue transcripts. Determine whether the sentences *{line1}* and *{line2}* were spoken by a single person or by different people. Answer:

There are many ways to write a prompt

Which word sounds more like the other word?

Word1: breakfast

Word2: cereal

Query word: serial

Output word: cereal

Word1: eight

Word2: mouth

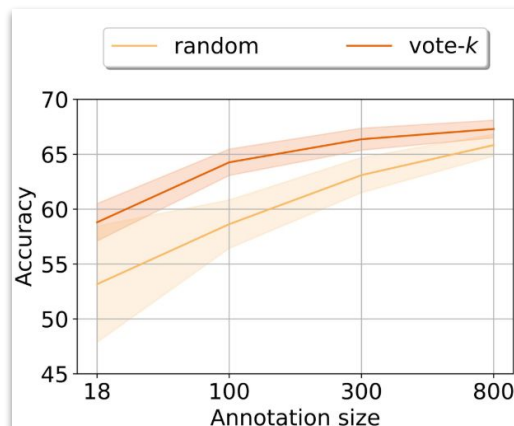
Query word: ate

Output word:

- Paraphrasing instructions
- **Choosing few-shot examples**
- Formatting the prompt

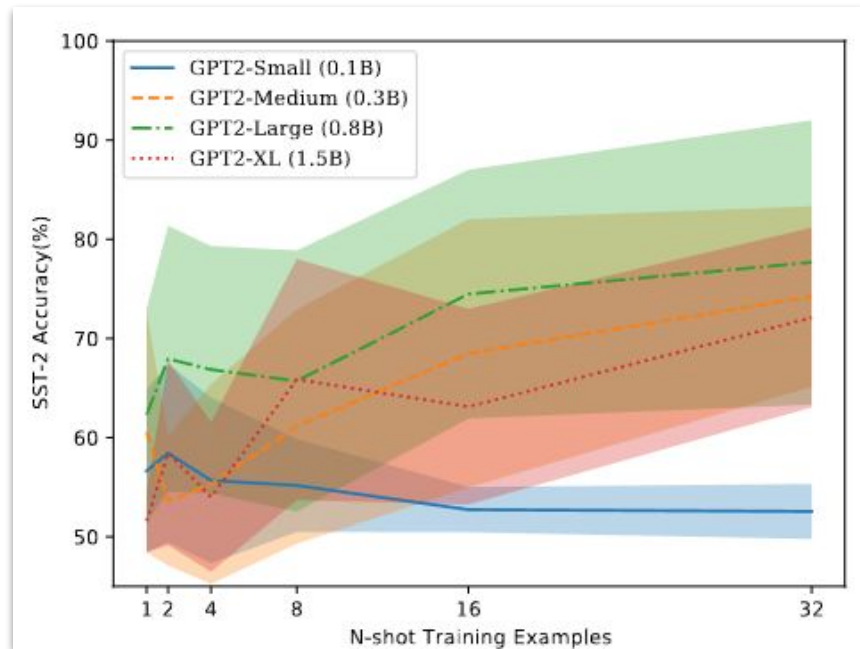
Choice of Few-Shot Examples Matter

- Works showing in-context learning (ICL) depends on examples
 - Sampling strategies beat random
 - E.g., via graph coverage ([Su et al, 2023](#))
- Aligns with similarities between ICL and gradient descent (GD)
[Irie et al. \(2022\)](#), [Oswald et al \(2022\)](#), [Dai et al \(2023\)](#)
- Although the relation of ICL and GD is still unclear
[Mosbach et al. \(2023\)](#), [Shen et al \(2023\)](#),
[Hendel et al. \(2023\)](#)



The Order of Few-Shot Examples Matter

- The order of in-context examples affects performance [Lu et al. \(2022\)](#)
 - Model size mitigates the problem, but doesn't solve it



There are many ways to write a prompt

Which word sounds more like the other word?

Word1: breakfast

Word2: cereal

Query word: serial

Output word: cereal

Word1: eight

Word2: mouth

Query word: ate

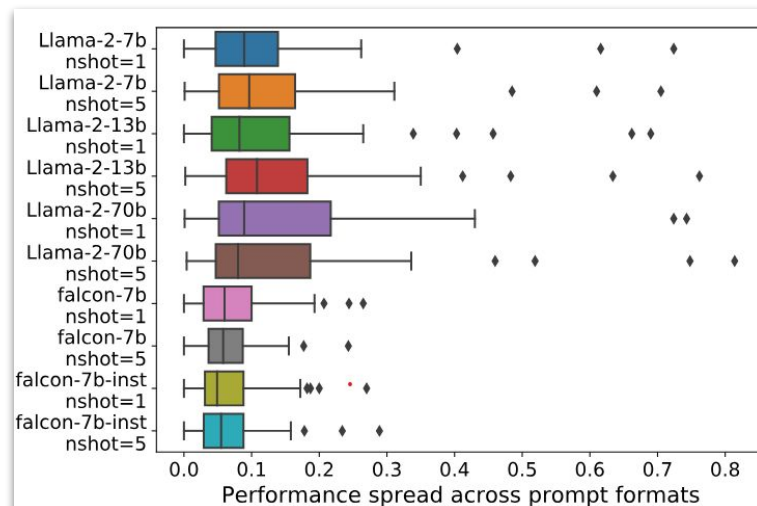
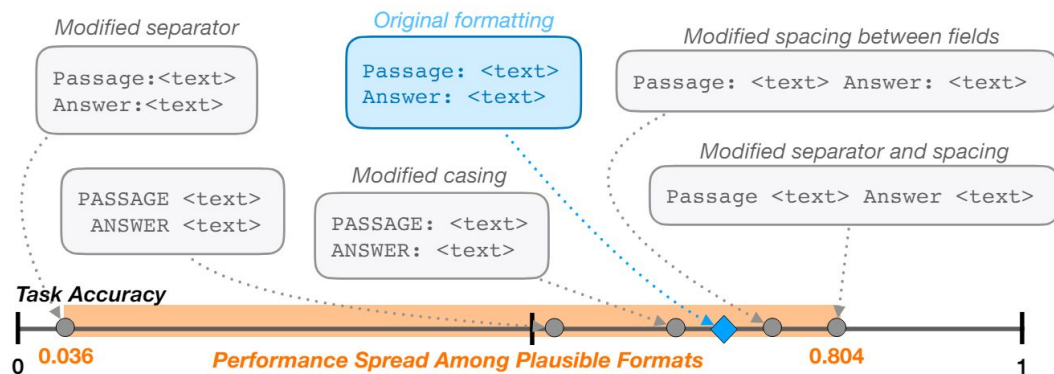
Output word:

- Paraphrasing instructions
- Choosing few-shot examples
- **Formatting the prompt**

Minor Prompt Formatting Matters

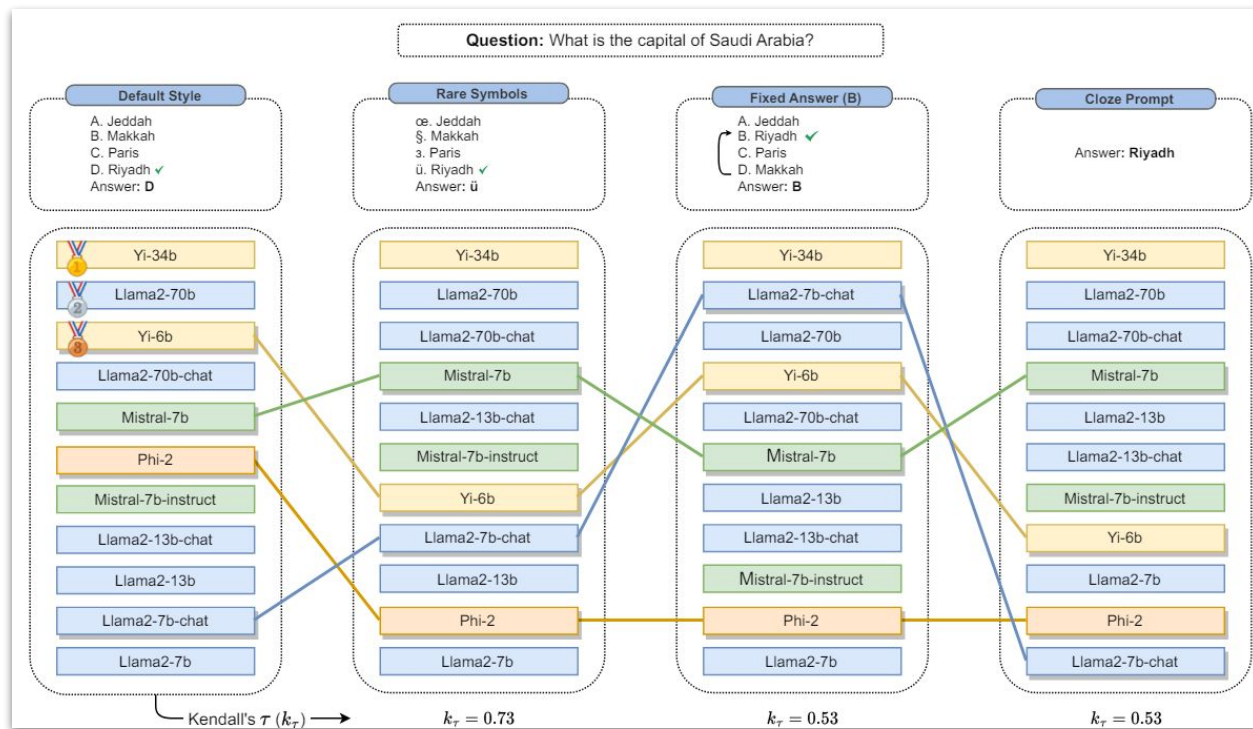
- Host of concurrent works notice the effect of **slight prompt edits**
- Provide large-scale evaluation and statistical analyses
- Seemingly **arbitrary changes** lead to **drastic changes**
- In both absolute performance as well as relative between models

Minor Prompt Formatting Matters



Sclar et al (2023)

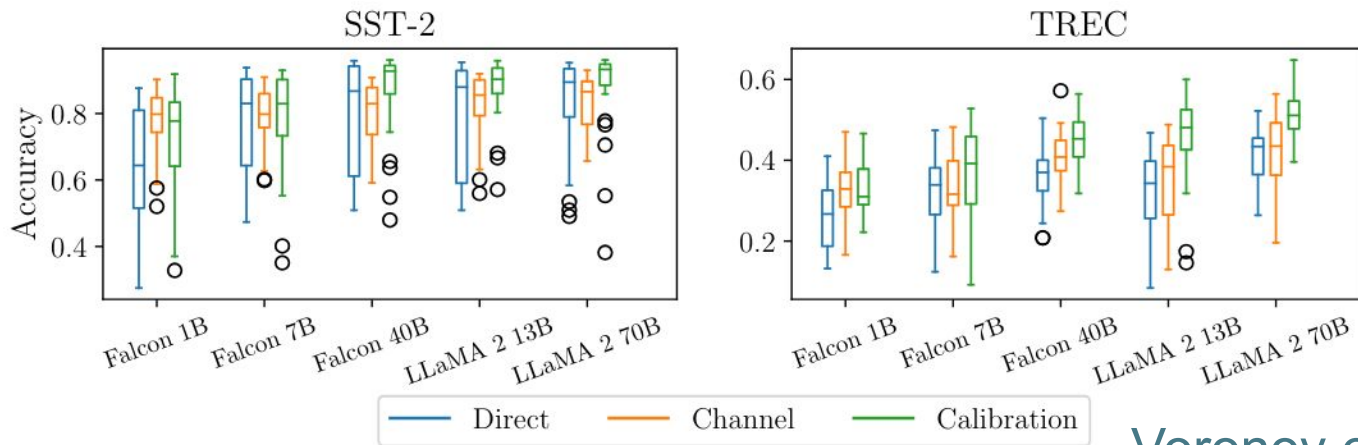
Minor Prompt Formatting Matters



(Alzahrani et al., 2024)

Minor Prompt Formatting Matters

Dataset	Input verbalizer	Output verbalizer	Intra-separator	Inter-separator
SST-2	"input: {}",	"output: {}", "target: {}", "label: {}", "emotion: {}", "sentiment: {}", "A {} one." "It was {}.", "All in all {}.", "A {} piece."		" "
DBPedia	"text: {}",		" ", "\n"	"\n",
AGNews	"sentence: {}",	"output: {}", "target: {}", "label: {}", "Topic: {}.", "Subject: {}.",		"\n\n"
TREC	"{}"	"This is about {}.", "It is about {}."		



Agenda

- Common practice in LLM evaluation: **single-prompt evaluation**
 - HELM, MMLU, BIG-Bench, etc.
- This leads to **very brittle evaluation results**
 - Subtle changes in various prompt elements lead to vast differences
- We'll discuss **mitigation strategies**
 - E.g., evaluating against a pool of prompts

Open Question: Why are LLMs so Sensitive?

- The behavior we've seen seems **paradoxical**
 - On the one hand, LLMs achieve SOTA on practically all tasks
 - Yet they can be stumped by an extra space or comma
- **Implicit biases introduced during training?**
 - Preprocessing over immense data may introduce unintended regularities
 - Other biases in e.g., RLHF or instruction tuning
- **Dataset contamination?**
 - LLMs may partly succeed due to dataset leakage and overfit their format
- **Finding the root cause seems very hard**
 - Training a strong LLM is an immense engineering challenge
 - With various of steps which are hard to control for independently
 - **But definitely worth exploring!**

Proposed Approaches are Expensive

- Mitigation strategies focus on either more **meaningful evaluation**
 - [Mizrahi et al \(2024\)](#) evaluate over a pool of prompts
 - [Sclar et al \(2023\)](#) calculate confidence intervals across prompt formats
- Or on more **robust models**
 - [Voronov et al \(2024\)](#) ensemble responses during inference
- These approaches add **many templates per sample**
- Highlighting the importance of **efficient evaluation**
 - **Which is the topic of our next section!**

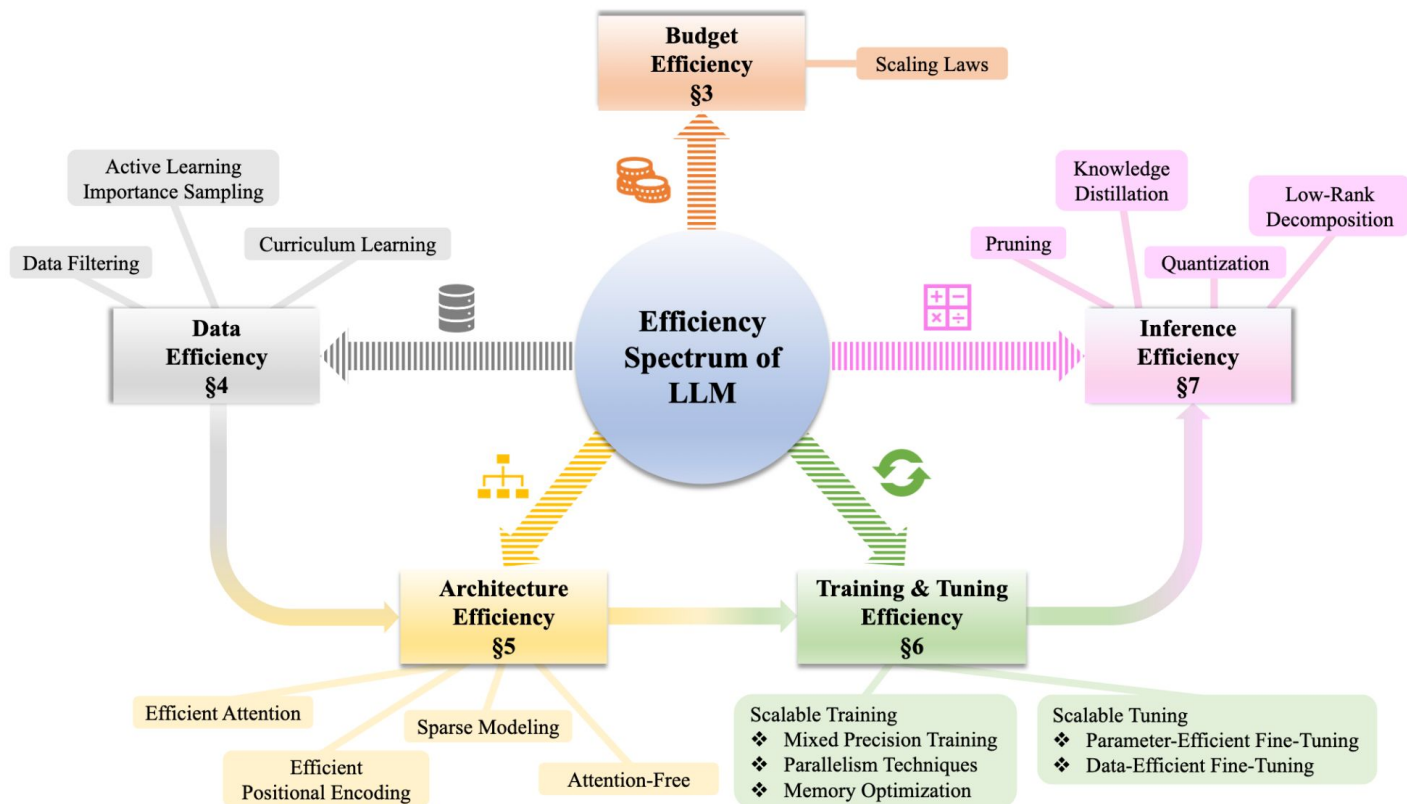
Part 4: Efficient Evaluation

<i>LLMs are...</i>	Expressive	Large	SOTA on everything
<i>Making their evaluation...</i>	Sensitive to instruction phrasing	Computationally demanding	Subjective Hard to automate
<i>Tutorial Sections</i>	Prompts	Efficiency	Metrics + Human evaluation

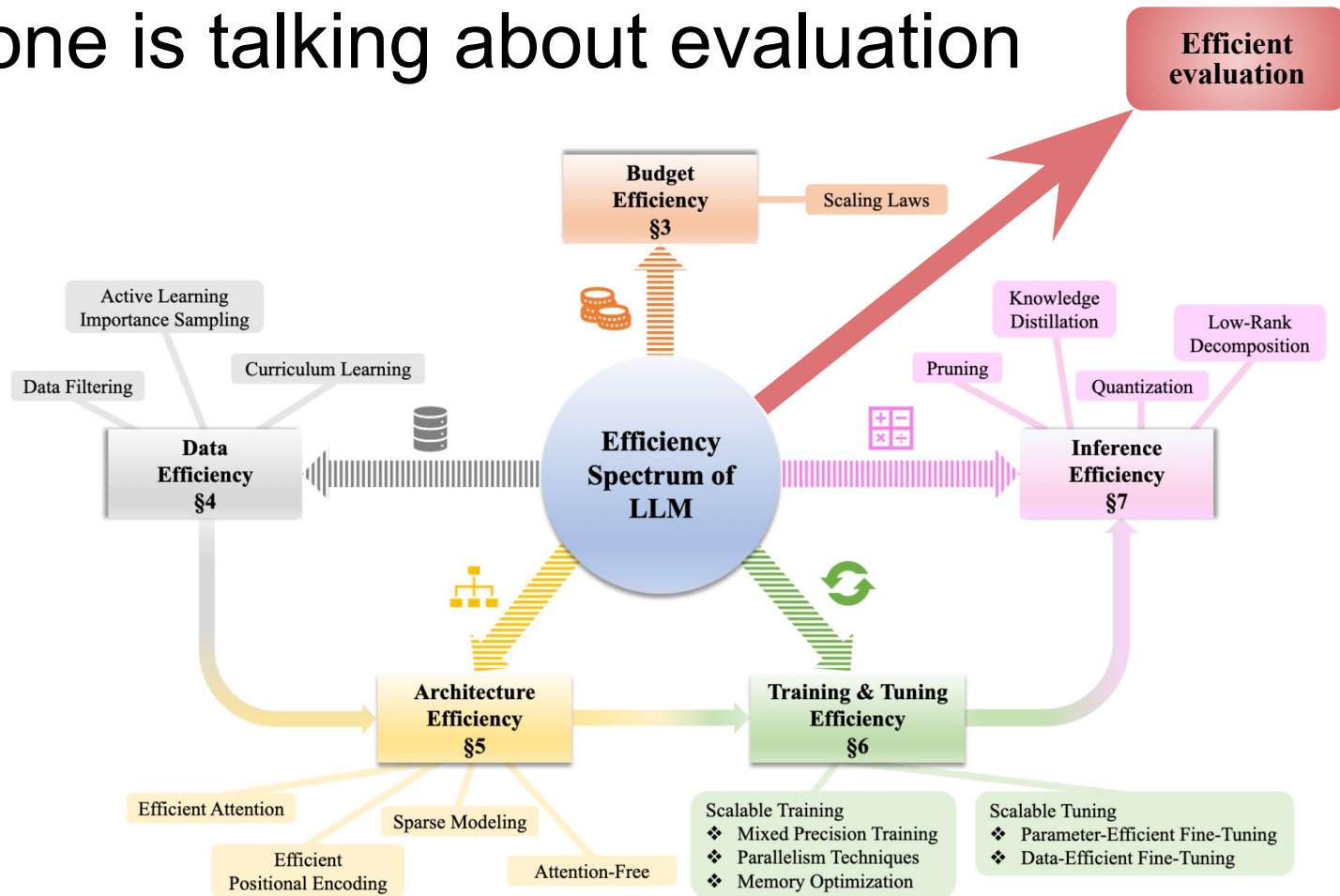
Agenda

- Zooming out
 - Evaluation is just a part of the efficiency spectrum
- Properties of evaluation
 - Define Objective, Validity, Reliability
- Challenges facing evaluations
 - Reliability Vs Compute (\$\$\$)
- Solutions - efficient methods
 - A naive (random) take
 - More sophisticated methods

LLMs efficiency spectrum



No one is talking about evaluation



Agenda: Efficient evaluation

- Zooming out
 - Evaluation is a part of the efficiency spectrum
- Properties of evaluation
 - Define Objective, Validity, Reliability
- Challenges
 - LLMs are large and expressive -> \$\$\$
- Solutions
 - A naive (and effective) take
 - More sophisticated methods

Properties of Evaluation

Objective: what do I want my eval to tell me? (score?, rank? etc)

Validity: Do I measure the thing I want?

- Hard to measure

Reliability: Are results consistent under **arbitrary changes**?

- Can be measured

Properties of Evaluation

Objective: what do I want my eval to tell me? (score?, rank? etc)

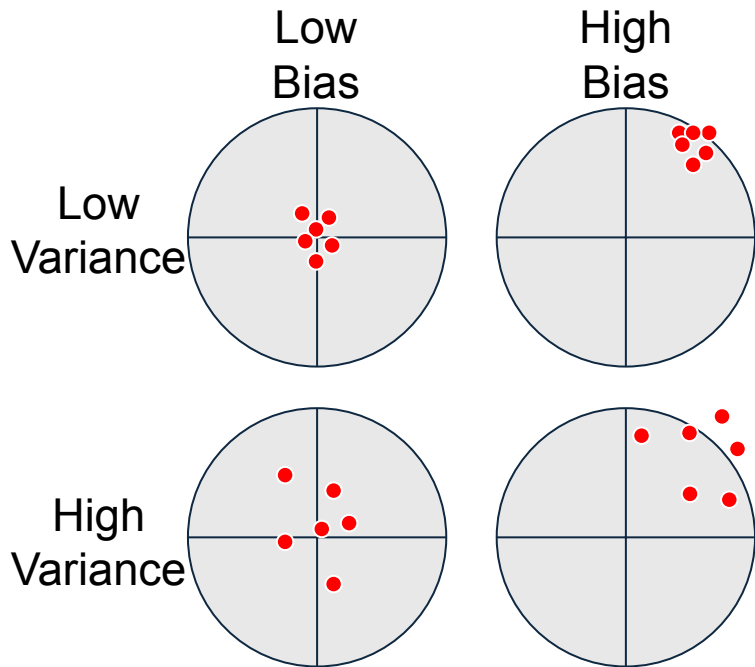
Validity: Do I measure the thing I want?

- Hard to measure

Reliability: Are results consistent under **arbitrary changes**?

- Can be measured

Validity $\leftrightarrow$ Bias
Reliability $\leftrightarrow$ Variance



Agenda: Efficient evaluation

- Zooming out
 - Evaluation is a part of the efficiency spectrum
- Properties of evaluation
 - Validity, Reliability and Cost
- **Challenges**
 - Reliability Vs Compute (\$\$\$)
- Solutions
 - A naive (and effective) take
 - More sophisticated methods

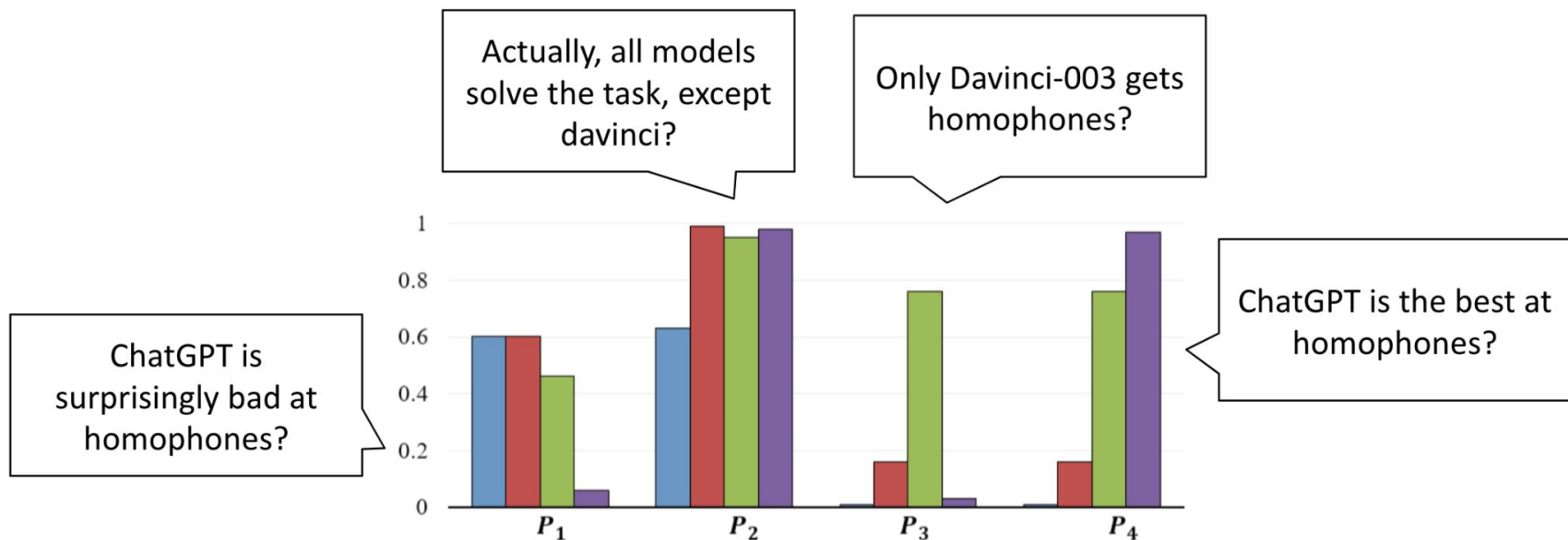
Challenge: Reliability

Every evaluation makes arbitrary decisions:

...what specific **datasets**, **examples** or **prompts** to use

When arbitrary decisions impact evaluation results, **reliability is compromised**, and conclusions become shaky

For example: The Arbitrary choice prompts impacts results



- [Mizrahi et al \(2024\)](#)

Challenge: Reliability

To improve reliability, authors avoid decisions and scale up their evaluations.

More prompts, more examples, more configurations, more datasets...

But what about compute?

The state of reliability - A test case



Holistic Evaluation of Language Models

The benchmark has tens of tasks that spans tens of thousands of examples per model.

Accuracy

Model/adaptor	Mean win rate ↑ [sort]	MMLU - EM ↑ [sort]	BoolQ - EM ↑ [sort]	NarrativeQA - F1 ↑ [sort]	NaturalQuestions (closed-book) - F1 ↑ [sort]	NaturalQuestions (open-book) - F1 ↑ [sort]	QuAC - F1 ↑ [sort]	HellaSwag - EM ↑ [sort]	OpenbookQA - EM ↑ [sort]	TruthfulQA - EM ↑ [sort]	MS MARCO (regular) - RR@10 ↑ [sort]	MS MARCO (TREC) - NDCG@10 ↑ [sort]	CNN - RO [sort]
Llama 2 (70B)	0.943	0.582	0.886	0.77	0.458	0.674	0.484	-	-	0.554	-	-	-
LLaMA (65B)	0.912	0.584	0.871	0.755	0.431	0.672	0.401	-	-	0.508	-	-	-
text-davinci-002	0.904	0.568	0.877	0.727	0.383	0.713	0.445	0.815	0.594	0.61	0.421	0.664	0.15:
Cohere Command beta (52.4B)	0.88	0.452	0.856	0.752	0.372	0.76	0.432	0.811	0.582	0.269	0.472	0.762	0.16
text-davinci-003	0.869	0.569	0.881	0.727	0.406	0.77	0.525	0.822	0.646	0.593	0.368	0.644	0.15:
Llama 2 (13B)	0.839	0.507	0.811	0.744	0.376	0.637	0.424	-	-	0.33	-	-	-
Jurassic-2 Jumbo (178B)	0.83	0.48	0.829	0.733	0.385	0.669	0.435	0.788	0.558	0.437	0.398	0.661	0.14:
TNLG v2 (530B)	0.8	0.469	0.809	0.722	0.384	0.642	0.39	0.799	0.562	0.251	0.377	0.643	0.16:
gpt-3.5-turbo-0613	0.798	0.391	0.87	0.625	0.348	0.675	0.485	-	-	0.339	-	-	-
LLaMA (30B)	0.795	0.531	0.861	0.752	0.408	0.666	0.39	-	-	0.344	-	-	-
Anthropic-LM v4-s3 (52B)	0.791	0.481	0.815	0.728	0.288	0.686	0.431	0.807	0.558	0.368	-	-	0.15:
gpt-3.5-turbo-0301	0.767	0.59	0.74	0.663	0.39	0.624	0.512	-	-	0.609	-	-	-

Many more models...

Many more scenarios...

Why does efficiency matter?

Model	Model Creator	Modality	# Parameters	Tokenizer	Window Size	Access	Total Tokens	Total Queries	Total Cost
J1-Jumbo v1 (178B)	AI21 Labs	Text	178B	AI21	2047	limited	327,443,515	591,384	\$10,926
J1-Grande v1 (17B)	AI21 Labs	Text	17B	AI21	2047	limited	326,815,150	591,384	\$2,973
J1-Large v1 (7.5B)	AI21 Labs	Text	7.5B	AI21	2047	limited	342,616,800	601,560	\$1,128
Anthropic-LM v4-s3 (52B)	Anthropic	Text	52B	GPT-2	8192	closed	767,856,111	842,195	-
BLOOM (176B)	BigScience	Text	176B	BLOOM	2048	open	581,384,088	849,303	4,200 GPU hours
T0++ (11B)	BigScience	Text	11B	T0	1024	open	305,488,229	406,072	1,250 GPU hours
Cohere xlarge v20220609 (52.4B)	Cohere	Text	52.4B	Cohere	2047	limited	397,920,975	597,252	\$1,743
Cohere large v20220720 (13.1B) ⁵⁸	Cohere	Text	13.1B	Cohere	2047	limited	398,293,651	597,252	\$1,743
Cohere medium v20220720 (6.1B)	Cohere	Text	6.1B	Cohere	2047	limited	398,036,367	597,252	\$1,743
Cohere small v20220720 (410M) ⁵⁹	Cohere	Text	410M	Cohere	2047	limited	399,114,309	597,252	\$1,743
T5 (11B)	Google	Text	11B	T5	512	open	199,017,126	406,072	1,380 GPU hours
UL2 (20B)	Google	Text	20B	UL2	512	open	199,539,380	406,072	1,570 GPU hours
OPT (66B)	Meta	Text	66B	OPT	2048	open	612,752,867	851,178	2,000 GPU hours
OPT (175B)	Meta	Text	175B	OPT	2048	open	610,436,798	851,178	3,400 GPU hours

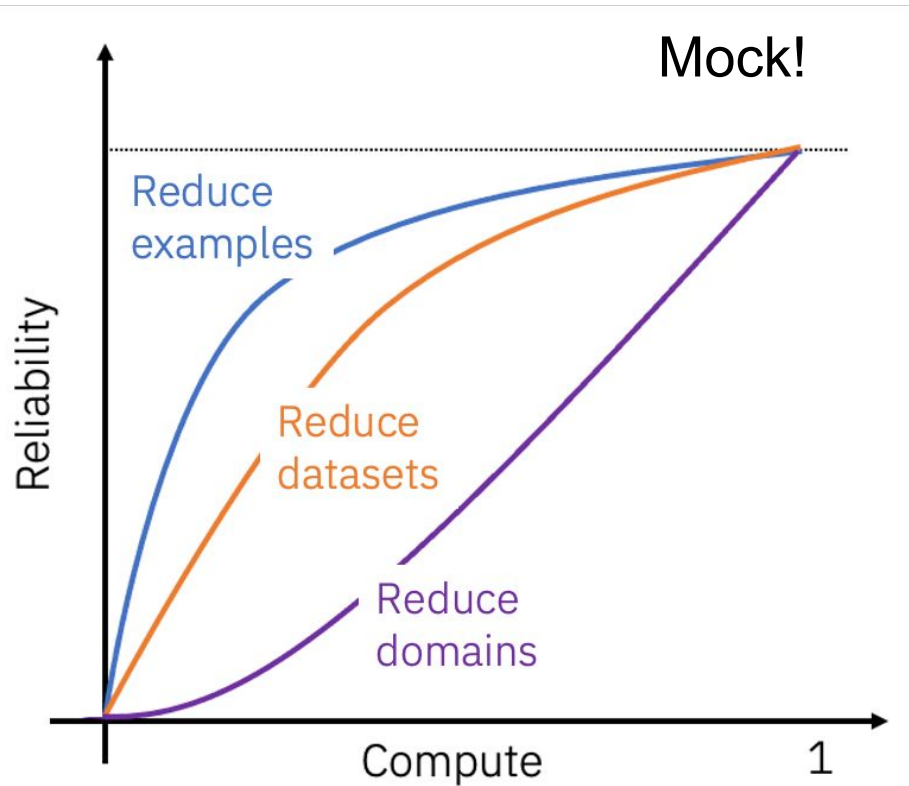
[From Liang et al \(HELM paper\)](#)

The Compute - Reliability Tradeoff

Reliability is a monotonic function of compute

However, Efficient evaluations can give better trade-offs

delivering more reliable evaluations for a smaller cost



Back to our Example

Agenda: Efficient evaluation

- Zooming out
 - Evaluation is a part of the efficiency spectrum
- Properties of evaluation
 - Validity, Reliability and Cost
- Challenges
 - Reliability and Compute (\$\$\$)
- **Solutions**
 - A naive (and effective) take
 - More sophisticated methods

Running Example



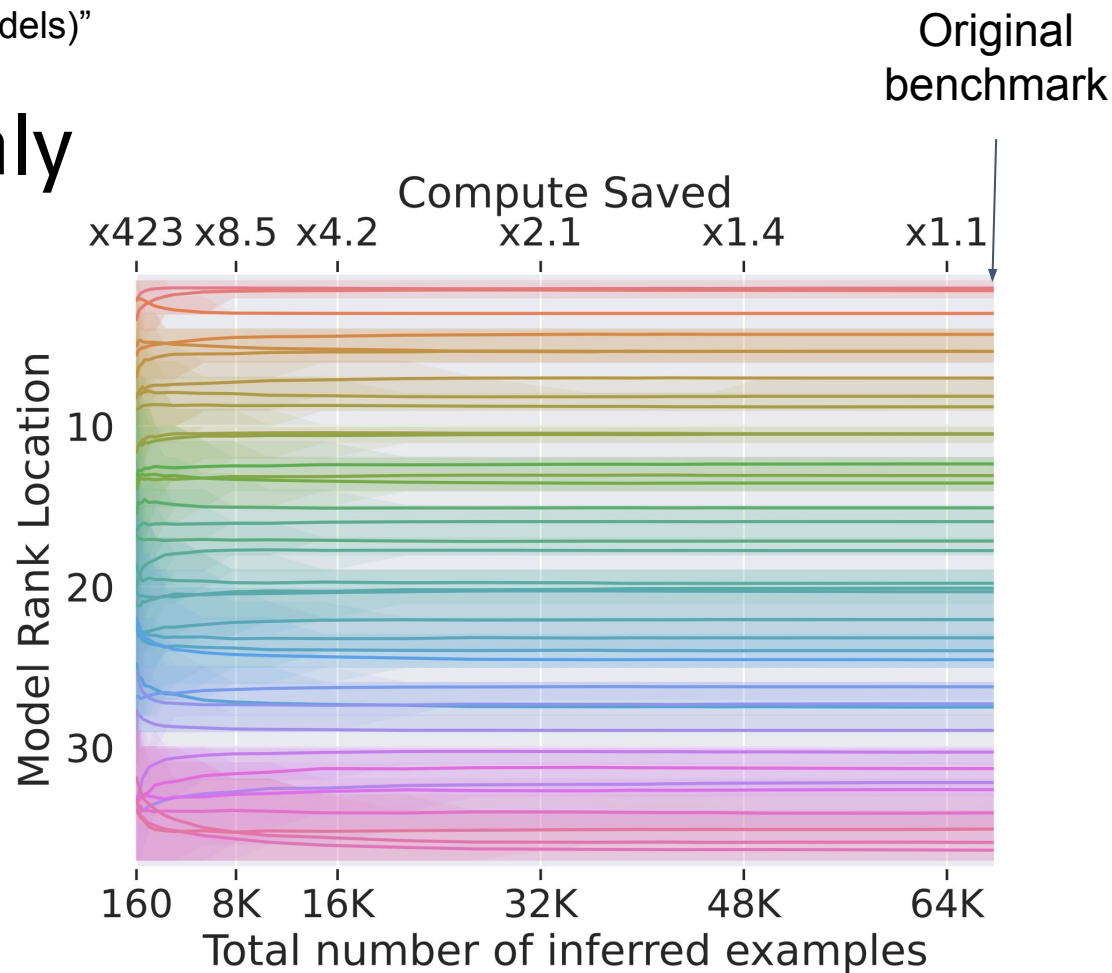
Holistic Evaluation of Language Models

HELM's main objective: Rank models

Sampling randomly

Model ranks are extremely stable even when compute drops dramatically:

- $\times 8.5$ produces nearly the same results as the full benchmark
- $\times 400$ preserves clusters

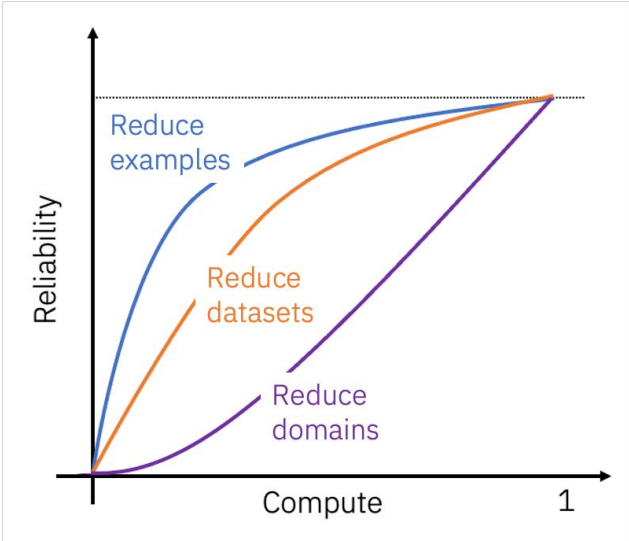
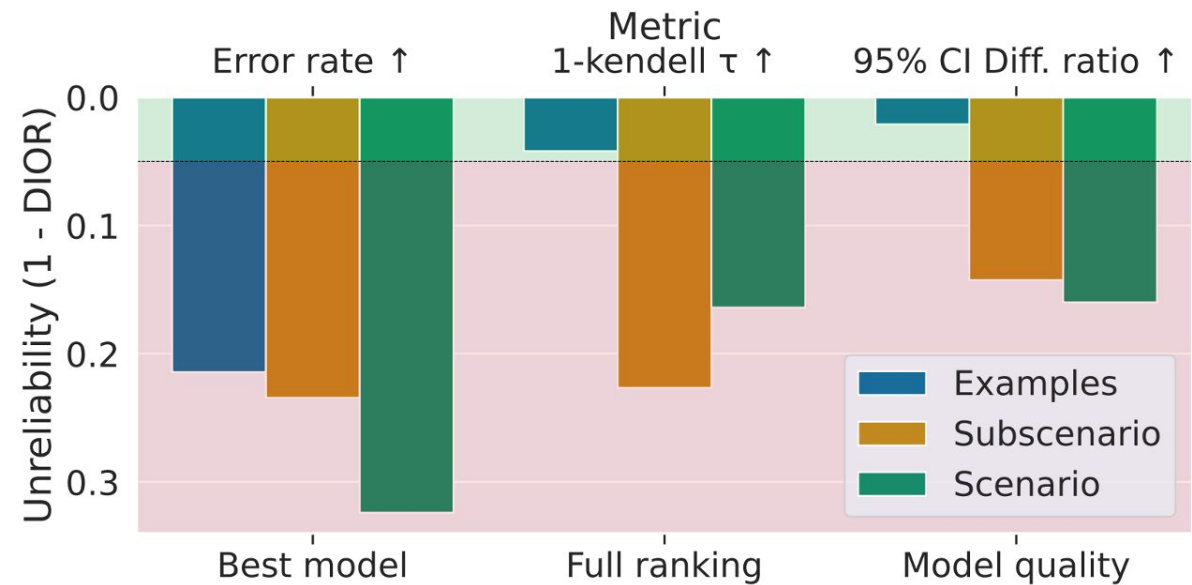


Two key takeaways

1. Some objectives **cannot be achieved**
2. Others, can be with **much less compute**



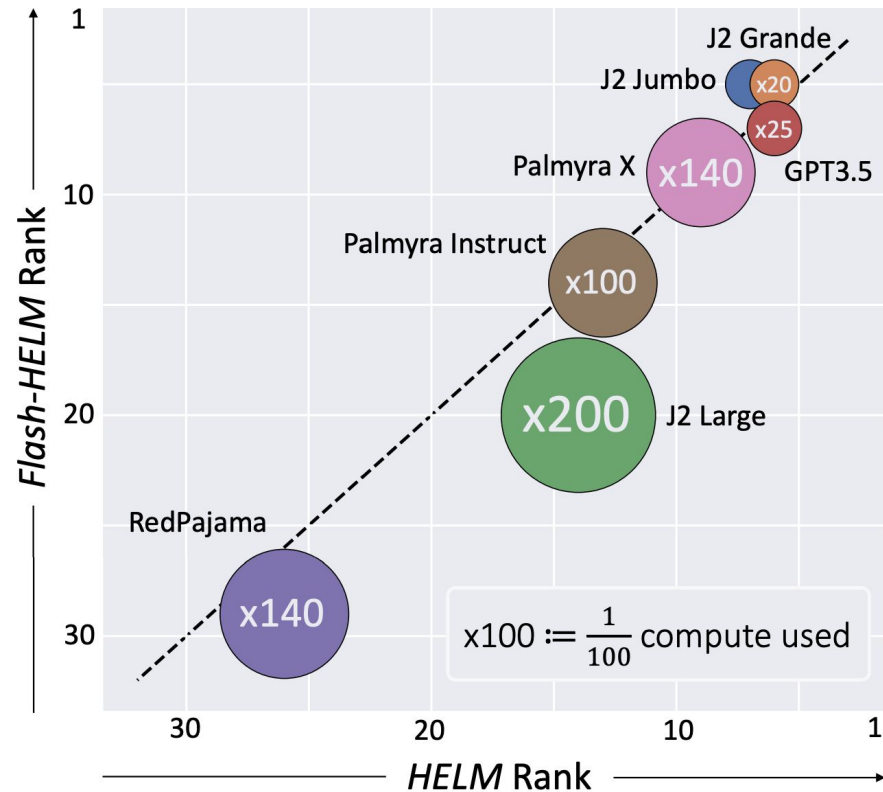
Subset examples – not datasets



Application

Algorithm 1 Efficient ‘coarse-to-fine’ tournament

```
 $M \leftarrow$  The evaluated model
for Sample size  $S \in [20, 50, 200, 1000, Max]$ 
do
     $Rank(M, S) \leftarrow$  Evaluated model  $M$  using
    sub-sample of size  $S$ .
    if  $Rank(M, S) - Res(S) \geq TierRank(S)$ 
    then
        stop;
    end if
end for
report  $Rank(M, S)$ ;
```



Efficient Benchmark Building Checklist

✓ **Report Benchmark compute costs (§1)**

Benchmarks often have heavy compute requirements, report required compute to increase usability.

✓ **Verify your design decisions with *DIoR* (§2)**

Quantify your benchmark's reliability-compute trade-off across your different decisions. For example: did you use enough examples/scenarios/prompts/seeds? perhaps too many?

✓ **Compute matters - Suggest an Efficient benchmark version (§5.3 and §6)**

In addition to the full benchmark, provide the user with efficient compute-saving alternatives with varying degrees of reliability, e.g., by reducing the number of examples.

✓ **Reliability matters - Report where it is lacking (§5.3)**

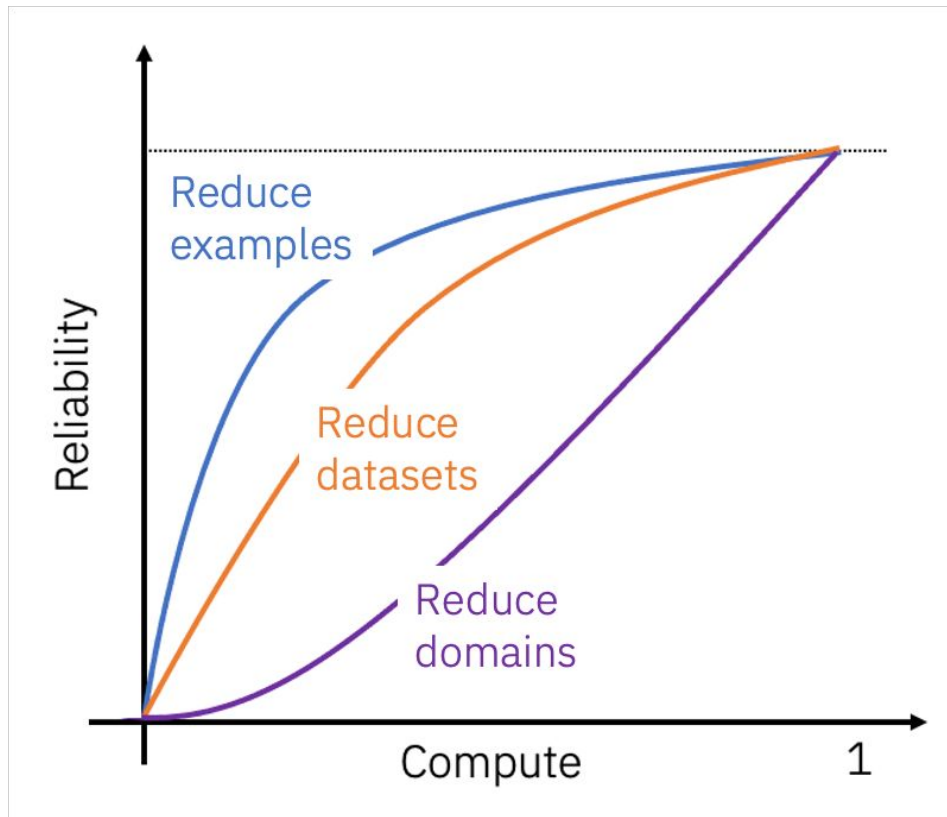
Identify reliability issues, such as distinguishing between top models for which we found HELM to be unreliable. Transparently report these limitations to avoid over-interpreting unreliable results.

✓ **Maximize data-points variability to improve reliability (§5.4)**

When sampling from multiple sources of variation (e.g., prompts, examples), maximize the coverage of each source, rather than exhausting all cross-product combinations of a few sources.

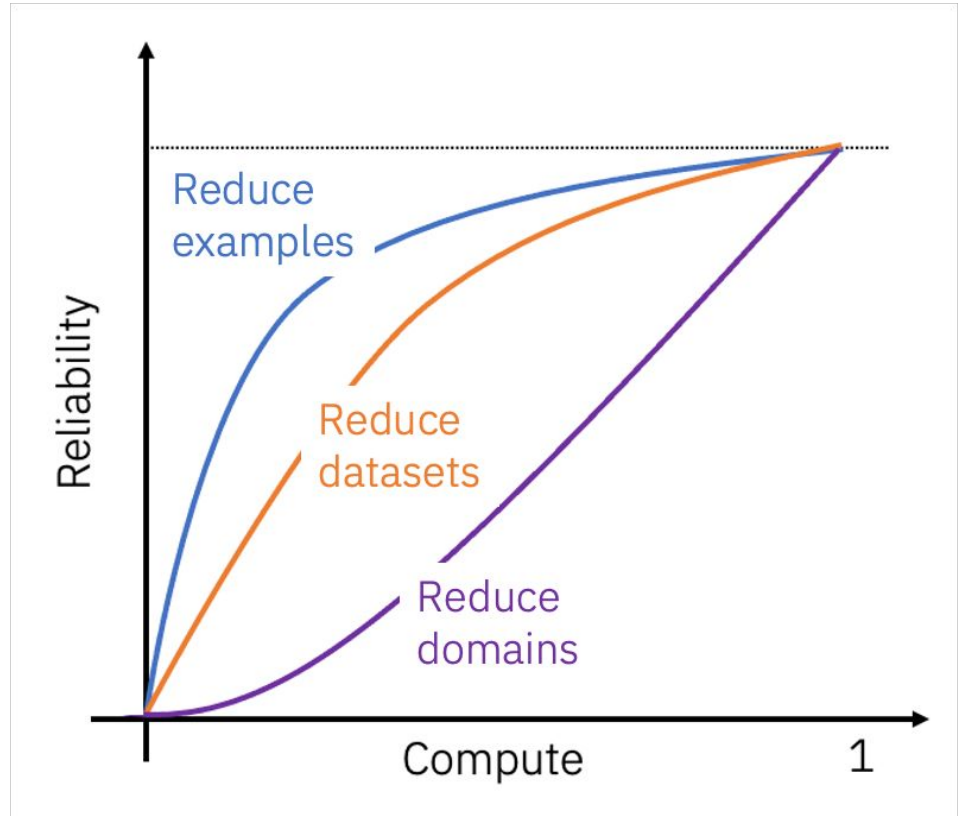
We have shown
that:

Simply sampling
randomly works
quite well



Next (obvious)
question:

Can we get a
better tradeoff
with smarter
sampling?



Agenda: Efficient evaluation

- Zooming out
 - Evaluation is a part of the efficiency spectrum
- Properties of evaluation
 - Validity, Reliability and Cost
- Challenges
 - Reliability and Compute (\$\$\$)
- **Solutions**
 - A naive (and effective) take
 - More sophisticated methods

“Smart sampling” methods

Example Clustering by model predictions

[Anchor Points: Benchmarking Models with Much Fewer Examples](#)

Item response theory (IRT)

[tinyBenchmarks: evaluating LLMs with fewer examples](#)

Full prediction

[How Predictable Are Large Language Model Capabilities? A Case Study on BIG-bench](#)

“Smart sampling” methods

Will describe the mechanisms of these methods, but will not elaborate on their effectiveness...

Generally, as a rule of thumb, even if methods are effective, the effects will be present only with very low number of examples and usually disappears when there are more than 100 examples.

“Smart sampling” methods

Example Clustering by model predictions

[Anchor Points: Benchmarking Models with Much Fewer Examples](#)

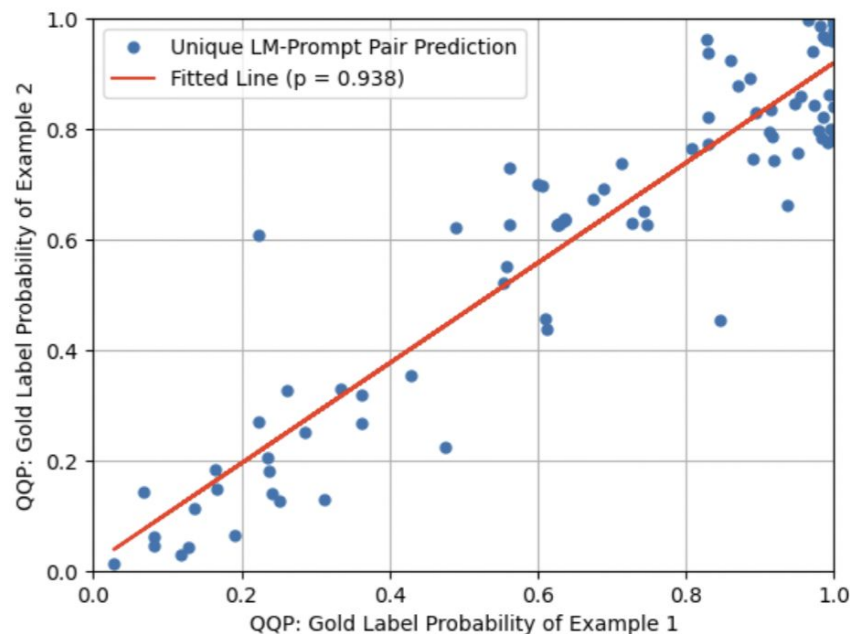
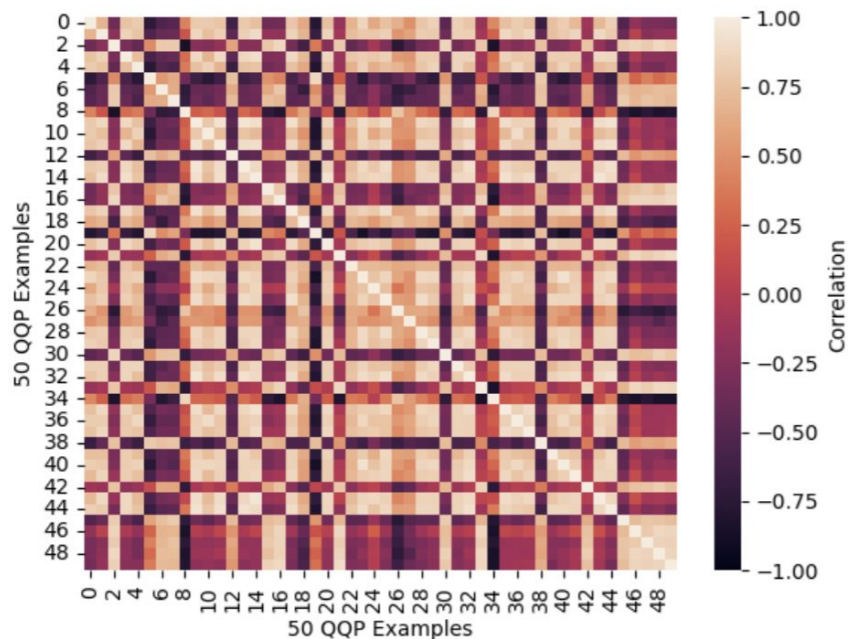
Item response theory (IRT)

[tinyBenchmarks: evaluating LLMs with fewer examples](#)

Full prediction

[How Predictable Are Large Language Model Capabilities? A Case Study on BIG-bench](#)

Found that examples are generally highly predictive to other ones in terms of model evaluation.



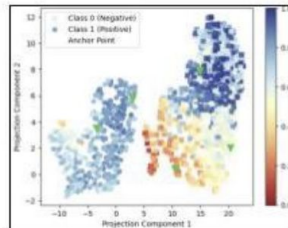
Algorithm: Train

1. Obtain scores for all examples with a few models
2. Cluster examples according to their evaluation results
3. Sample anchor points from centroids

Benchmarking:
Estimated Target
Model Rankings

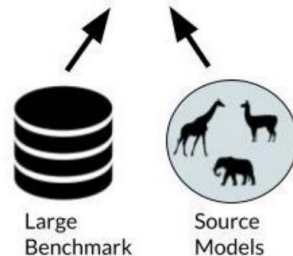
	model	score
1		0.94
2		0.91
3		0.87

Anchor Point Map:
Estimated Target Model Behavior
Over Entire Benchmark



Anchor Points

text	label	weight
"the jabs it employs are short"	1	0.05
"turned it... into drama"	0	0.48
"this is successful as a film"	1	0.47



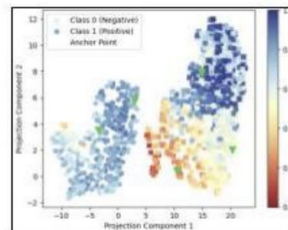
Algorithm: Test

1. Evaluate examples from anchor points
2. Predict full eval results from anchor points scores

Benchmarking:
Estimated Target
Model Rankings

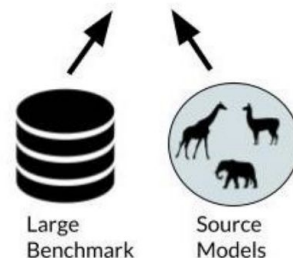
	model	score
1		0.94
2		0.91
3		0.87

Anchor Point Map:
Estimated Target Model Behavior
Over Entire Benchmark



↑ ↑
Anchor Points

text	label	weight
"the jabs it employs are short"	1	0.05
"turned it... into drama"	0	0.48
"this is successful as a film"	1	0.47



“Smart sampling” methods

Example Clustering by model predictions

Anchor Points: Benchmarking Models with Much Fewer Examples

Item response theory (IRT)

tinyBenchmarks: evaluating LLMs with fewer examples

Full prediction

How Predictable Are Large Language Model Capabilities? A Case Study on BIG-bench

Item Response Theory (IRT)

Each example is:

- A vector of required skills
- A difficulty constant

Each LLM as a vector of abilities

Train:

- Estimate α_i and β_i using a training set,
- Cluster (α_i, β_i) to find a "tiny" subset of points to evaluate

$$p_{il} \triangleq \frac{1}{1 + \exp(-\underbrace{\alpha_i^\top}_{\text{Required skills for example } i} \underbrace{\theta_l}_{\text{Abilities of model } l} + \underbrace{\beta_i}_{\text{Difficulty of example } i})}$$

Infer:

- Estimate θ_l for a new LLM using evaluations on the "tiny" subset
- Predict correctness on the remaining examples

Lastly:

- Combine observed and predicted correctness to estimate overall performance

$$p_{il} \triangleq \frac{1}{1 + \exp(-\alpha_i^\top \theta_l + \beta_i)}$$

**Difficulty of
example i**

**Required skills for
example i** **Abilities of model l**

Authors open-sourced their benchmarks

The authors produced a tiny version of a few common benchmarks that produce results that are highly consistent with the full datasets

<https://huggingface.co/tinyBenchmarks>

Next (obvious) question:

Can we evaluate without
inference?

Subsample tasks

Train an MLP to predict model score on a new (model, task) configuration.

Model Family	# param	Task	# shot	Perf.
GPT-3	3B	strategy_qa	0	0.48
BIG-G T=1	8B	elementary_math	3	0.19
PaLM	64B	code_line_desc	2	0.23
GPT-3	6B	elementary_math	1	?

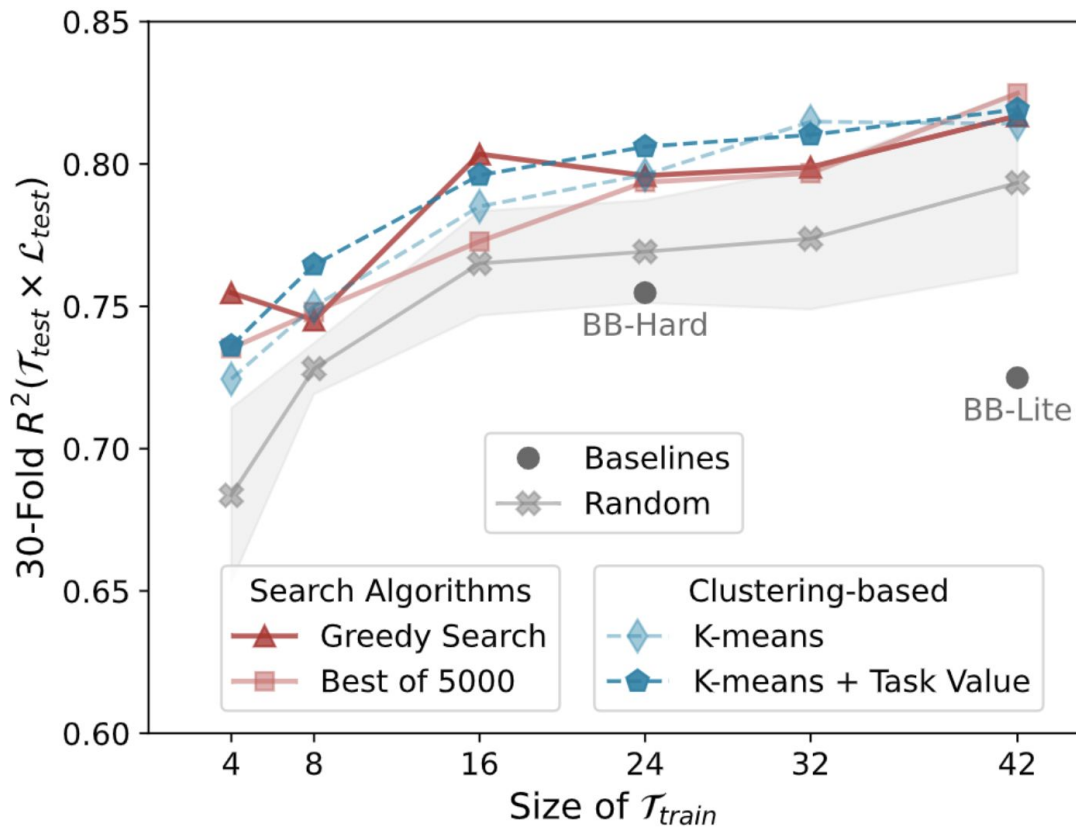
How *predictable* are LLM capabilities?



How to evaluate new models *within budget constraints*?

Small bench

Using search methods
to find the most
predictive subset of
tasks



Takeaways

- Compute matters
 - Reliability matters
 - There is a tradeoff
-
- If you know your objective
 - Efficient methods can produce a good tradeoff

Metrics



<i>LLMs are...</i>	Expressive	Large	SOTA on everything
<i>Making their evaluation...</i>	Sensitive to instruction phrasing	Computationally demanding	Subjective Hard to automate
<i>Tutorial Sections</i>	Prompts	Efficiency	Metrics + Human evaluation

What did we talk about

Now what?

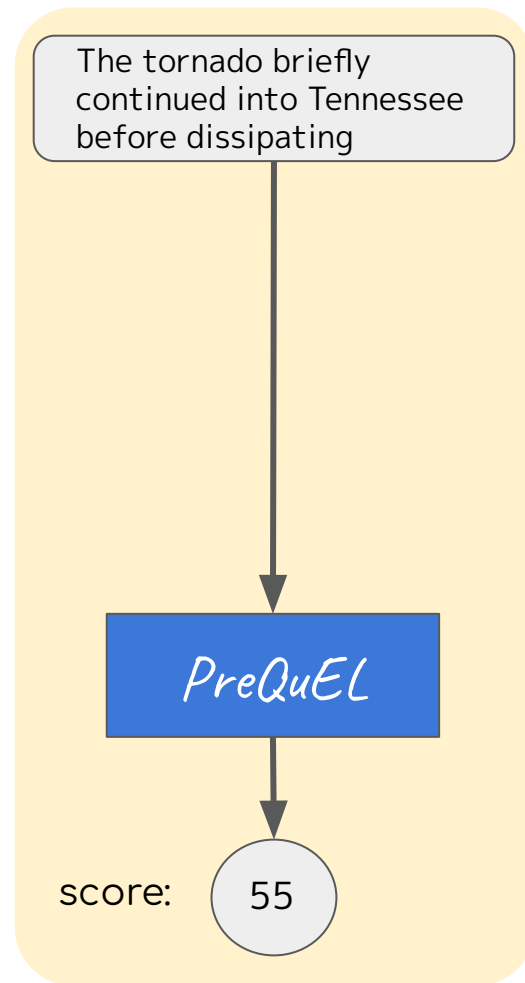
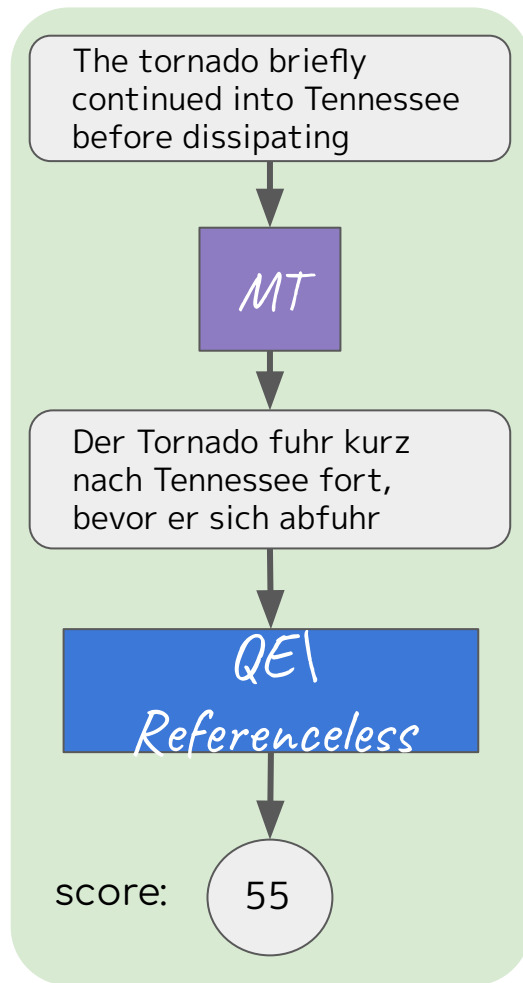
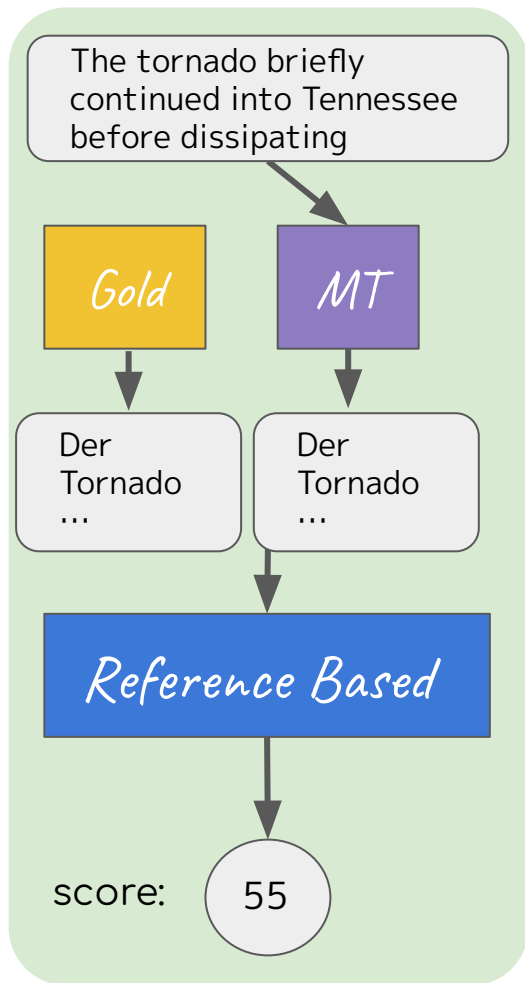
Booorin
g



Why metrics?

A metric quantifies the required trait
Usually by approximation





Many new metrics (Per field, specific characteristics...)

Adoption is rare

Real advantage

Replication of older work

Trust

We will focus on *some* of the mainstream ones

Overview the paradigmes

AMENLE: Robust Evaluation Metrics from Natural Language Inference

Yanran Chen^{1,2} and Steffen Eger²
¹Technische Universität Darmstadt, Germany
Natural Language Learning Group (NLLG) <https://nllg.github.io/>
chen@stud.tu-darmstadt.de, steffen.eger@uni-bielefeld.de

Abstract
BERT-based evaluation metrics perform well on standard benchmarks, but are vulnerable to adversarial attacks. Nonetheless, these novel metrics also have key limitations. For example, Sai and Kater et al. (2021) show that they are not robust to various adversarial attacks.

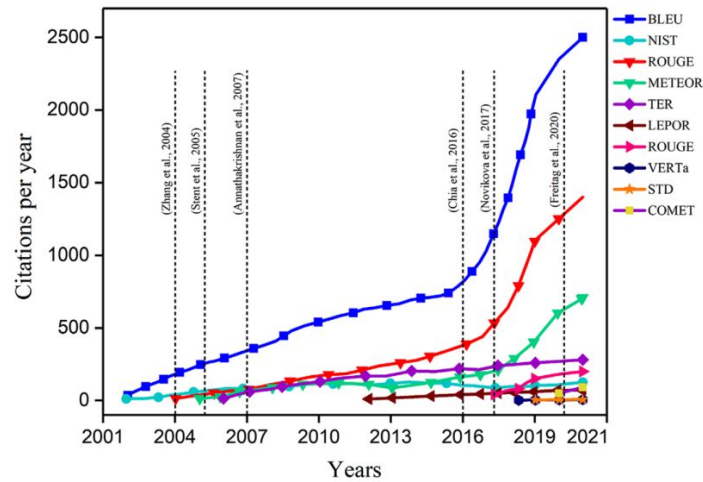
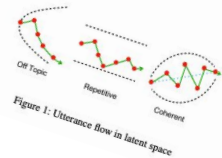
A PROPOSED NEW METRIC FOR THE CONCEPTUAL DIVERSITY OF A TEXT

İlkay Dönmez^a, Mehmet Hakkı İdris^b
^aBİLGEM, TÜBİTAK Kocaeli, Gebze, Türkiye, ilkay.donmez@tubitak.gov.tr
^bBİLGEM, TÜBİTAK Kocaeli, Gebze, Türkiye, mehmet.haklidr@tubitak.gov.tr

BBScore: A Brownian Bridge Based Metric for Assessing Text Coherence

Zhecheng Sheng¹, Tianhao Zhang², Chen Jiang², Dongyuep Kang²
¹University of Minnesota, Twin Cities
Minneapolis, Minnesota 55455
{sheng136, zhan7564, jiang649, dongyuep}@umn.edu

Abstract
Measuring the coherence of text is a vital aspect of evaluating the quality of written content. Recent advancements in neural coherence modeling have demonstrated their efficacy in capturing co-reference and discourse relations, thereby enhancing entity modeling and discourse analysis. However, many existing methods heavily depend on static embeddings, or focus narrowly on nearby context constraining their capacity to measure the overarching coherence of long texts. In this paper, we posit that coherent texts inherently manifest a sequential and cohesive interplay among sentences, effectively conveying the central theme, purpose, or storyline. To evaluate this abstract



Classic N-gram based metrics
Language Model based metrics
Reference-less Metrics
Language models as evaluators
Fine-grained and specialized metrics
Challenge sets, perturbation and data-based

Accuracy - exact match

Rouge (Bleu?)

F score

Precision Recall (what else RAG is currently using)

Perplexity - Cross entropy (bits per byte?)

Copy from the proposal

Title: A Comprehensive Analysis of the Effectiveness of Large Language Models
as Automatic Dialogue Evaluators

\\ (<https://arxiv.org/abs/2312.15407> , 1495kb)

Exact match\Acc vs. normalized acc (HF) or quasi (HELM) - the often ignored side of smoothing (and it always comes with a cost)

Metrics come up all the time for stuff like coherence or semantic diversity

Title: BBScore: A Brownian Bridge Based Metric for Assessing Text Coherence

\\ (<https://arxiv.org/abs/2312.16893> , 4000kb)

Title: A proposed new metric for the conceptual diversity of a text

Authors: İlknur Dönmez Phd, Mehmet Hakkı Doğan Phd

((<https://arxiv.org/abs/2312.16548> , 705kb)

System level evaluation - classic

But also modern methods try to do that. However entering the full evaluation into the model is too much, therefore one can use batches. <https://arxiv.org/pdf/2401.00437.pdf>

If you can answer a question, most likely you can evaluate others' responses. Do LLMs show similar behaviors?

We find that they do not! Models sometimes incorrectly evaluate the answer to the question they generate, and vice

versa. #EACL2024 #EACL2024SRW <https://t.co/dVL2vfs8KX>

(<https://x.com/juhyunohh/status/1770217898109215035?s=03>)

Probably unnecessary:

Evaluation by LLMs of dialogue (turn or full dialogue) fine-grained

<https://arxiv.org/pdf/2312.15407.pdf>

Classification

Accuracy - Percentage of correct responses
Aka Exact Match

Acc. 30%



Exact match\Acc
Normalized acc (HF)
Quasi acc (HELM)
Metrics are rarely simple
And most fixes are also not simple



Dr. Choshen, Leshem, Leshem Choshen
President Bush, President Bush 2nd {answer: Bush}

Classification

Accuracy - Percentage of correct responses
Aka Exact Match

Ranking

Precision

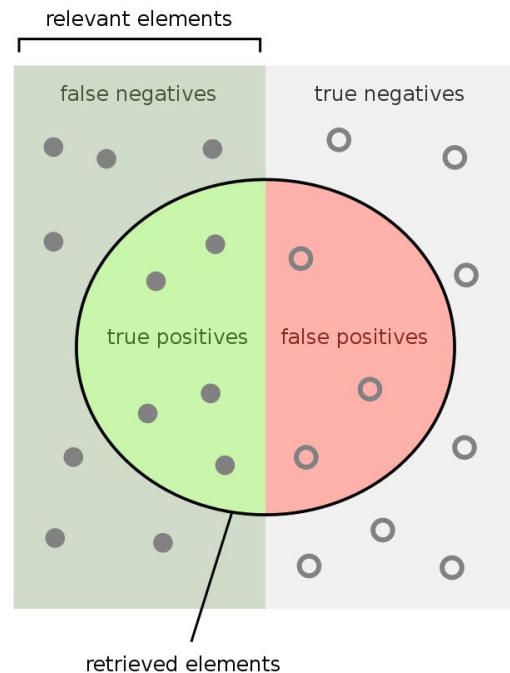
$$\frac{\text{Relevant retrieved instances}}{\text{All retrieved instances}}$$

Recall

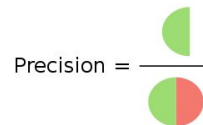
$$\frac{\text{Relevant retrieved instances}}{\text{All relevant instances}}$$

F

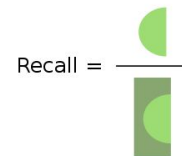
$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



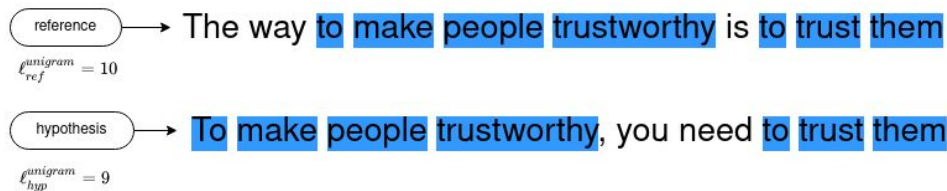
How many retrieved items are relevant?



How many relevant items are retrieved?



Ngrams



BLEU

$$p_n = \frac{\sum_{n\text{-gram} \in \text{hypothesis}} \text{Count}_{\text{match}}(n\text{-gram})}{\sum_{n\text{-gram} \in \text{hypothesis}} \text{Count}(n\text{-gram})}$$

$$BLEU_{N=4} = BP \cdot \exp \left(\sum_{n=1}^{N=4} \frac{1}{4} \log p_n \right)$$

ROUGE

$$ROUGE_1 = \frac{\sum_{unigram \in \text{reference}} \text{Count}_{\text{match}}(unigram)}{\sum_{unigram \in \text{reference}} \text{Count}(unigram)}$$

Ngrams

reference

$$\ell_{ref}^{unigram} = 10$$

hypothesis

$$\ell_{hyp}^{unigram} = 9$$

The way to make people trustworthy is to trust them

To make people trustworthy, you need to trust them

RougeL - longest sequence of consecutive words out of X

$$\begin{cases} R_{LCS} &= \frac{LCS(reference, hypothesis)}{\ell_{ref}^{unigram}} &= \frac{7}{10} \\ P_{LCS} &= \frac{LCS(reference, hypothesis)}{\ell_{hypothesis}^{unigram}} &= \frac{7}{9} \\ ROUGE_{LCS} &= \frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2P_{LCS}} &= \frac{(1+\beta^2)49}{70+\beta^263} \end{cases}$$

Write an email announcing a new tutorial at LREC conference
about evaluation and metrics in NLP.



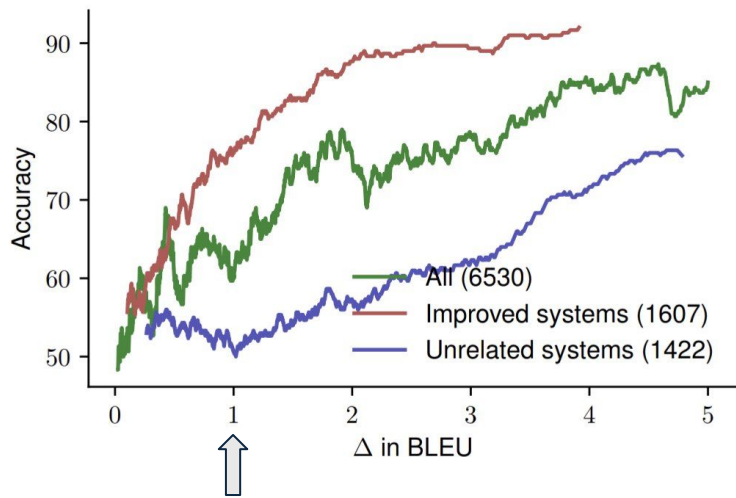
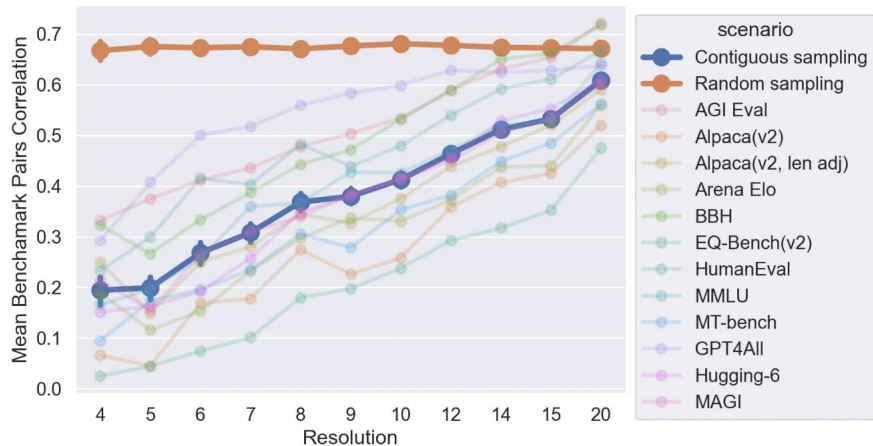
Dear LREC Conference Attendees,

...

Similarity to a reference is even harder in new open tasks
Model expressiveness is a bit challenging

Hard to differentiate similarly skilled models

Easier to assess progress than to compare models



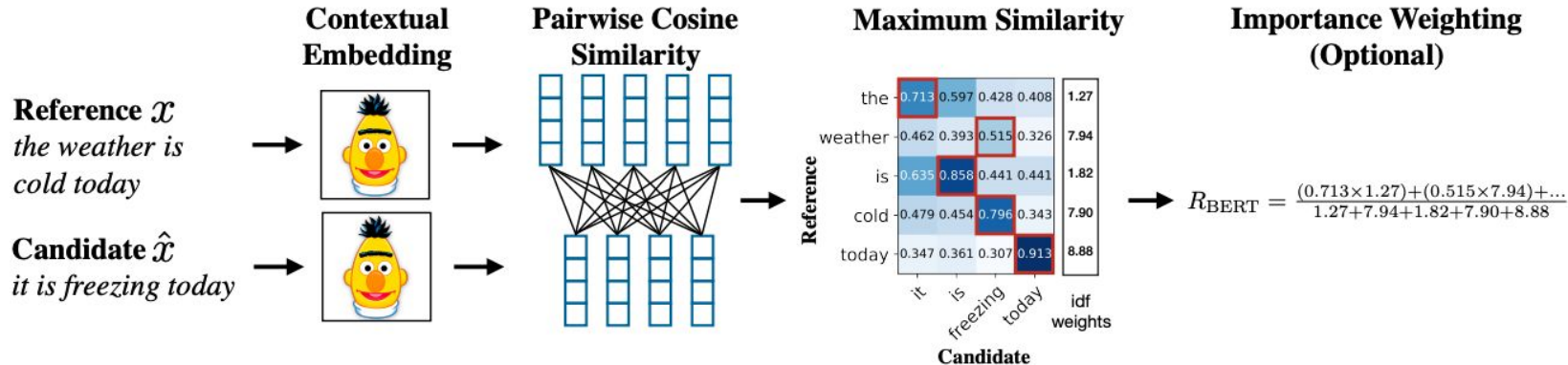
But we wanted to calculate
similarity of output to
reference...

Language Model based metrics

BERTscore - cosine similarity

Originally with BERT:

Stronger sentence encoders today



Quality Estimation

BLEURT
COMET

Why use only the pretrained
when we can train?

Better!
Stronger!

Ignores specifics,
like an actual year,
number or a year

What happens when you train

Training is on models evaluated by humans

Domain specific, task specific

Models change

Some tasks or languages don't have data etc.

Reliability introduced into metrics

Fine-tuned metrics have **lower** correlation on Biomedical domain than WMT

... despite Biomedical domain having **higher** correlation for other metrics

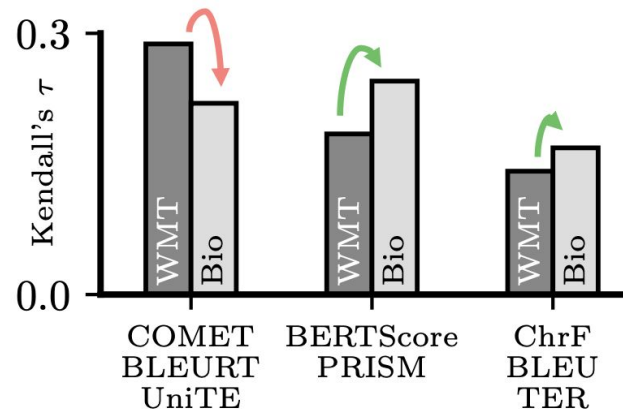


Figure 1: Automatic machine translation metric performance on the WMT and biomedical domains, averaged across metric types (see [Figure 2](#) for full results).

Reference-less Metrics

Do we have to use the reference?

For example,

if the back translation works - good translation

If no errors are detected - errors were corrected

LLM as a judge

Can new models just know it all?

Prompt a model to give a score to the other model (referenceful or more commonly referenceless)

Bias towards themselves (Prometheus)

Biases (gender, length etc.)

Is detection always simpler than generation?

Models sometimes incorrectly evaluate the answer to the question they generate, and vice versa.



There's more to evaluation than a single score

General: Reasoning - discourse - math

Linguistic: Syntax - semantics - morphology

Translation: Adequacy - fluency

...

Data no metric

Challenge sets - extract what hard phenomena

Adversarial - extract what models fail at (even if easy for humans)

Perturbation - Performance does not imply consistency

Data-based metrics - collect data that with a given metric tells us something (e.g. BLIMP)

EWOK (just announced):

C1: Ali treats people kindly.

C2: Ali treats people coldly.

T1: Ali is friendly.

T2: Ali is hostile.

Metric: $P(T1|C1) > P(T1|C2) \ \& \ P(T2|C2) > P(T2|C1)$

Challenges

Metrics:

Hard to adopt new ones

Hard to evaluate as the output is “richer”

Hard to evaluate as models are more similar
expressive

Part 6: Human Evaluation

Human evaluation

1. Definition

What do we refer to? (actually, many things)

2. Intrinsic issues

Assuming we need it, what should we be aware of?

3. Emerging trends

Nowadays, do we still need it?

Human evaluation - definition

While most evaluations have a human element to them (e.g., gold reference answers), we will focus on deliberate evaluation by humans

- **Goal-oriented human annotations for evaluating system outputs**
 - What qualities / goals are we measuring?
 - Who are the annotators?
 - How do they annotate - edits, qualitative feedback, absolute scores, preference...
- *Disclaimer: we will only scratch the surface*

**Twenty Years of Confusion in Human Evaluation:
NLG Needs Evaluation Sheets and Standardised Definitions**

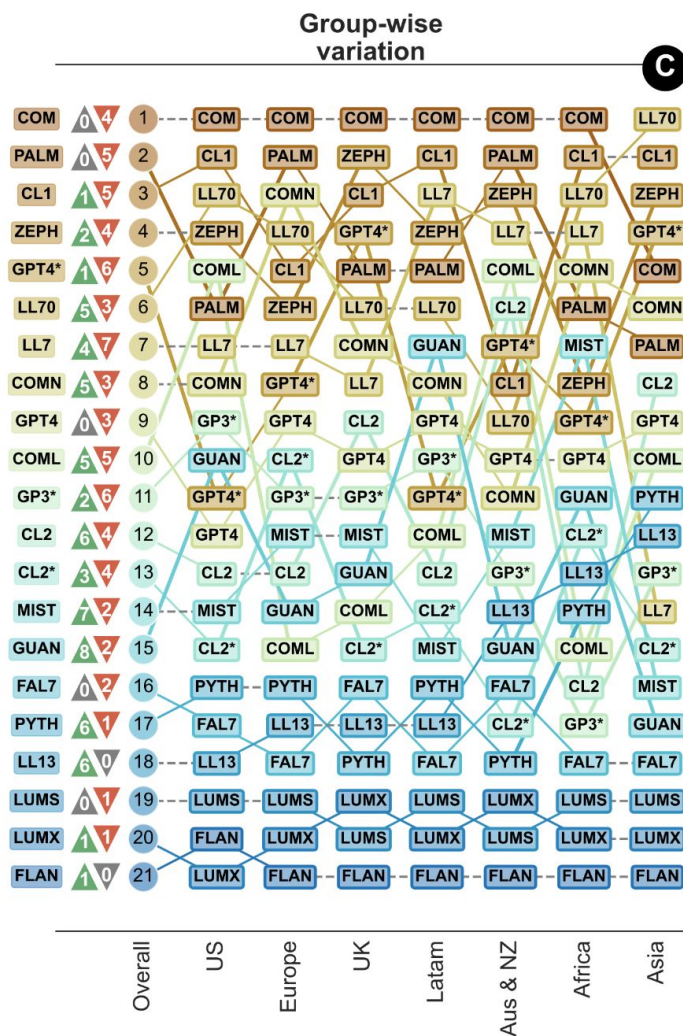
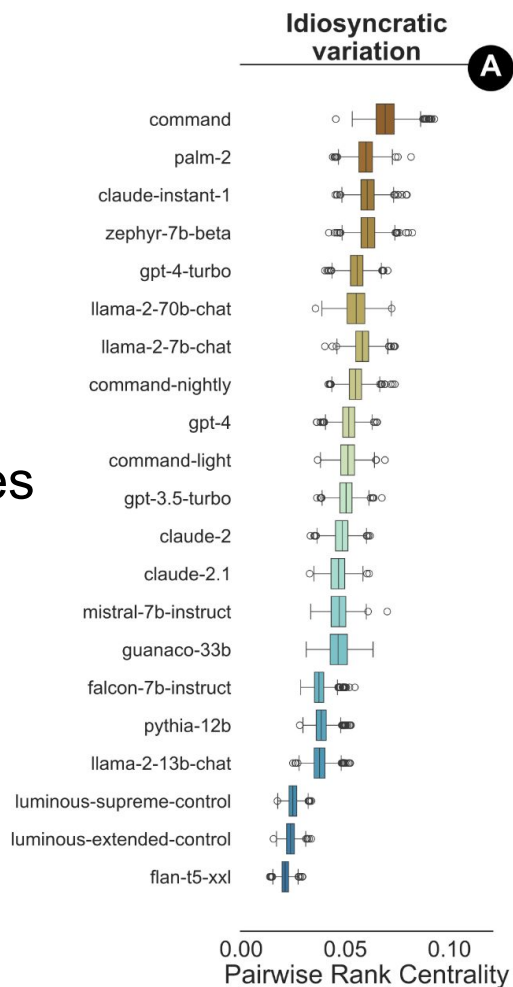
**Missing Information, Unresponsive Authors, Experimental Flaws:
The Impossibility of Assessing the Reproducibility
of Previous Human Evaluations in NLP**

**All That's 'Human' Is Not Gold:
Evaluating Human Evaluation of Generated Text**

Human evaluation - issues

- More often than not, annotation is **subjective**
 - A single gold answer, population “mean”, or diverse perspectives?
 - Inter-annotator agreement - *reliability* vs. *validity* [Amidei et al. 2019](#)

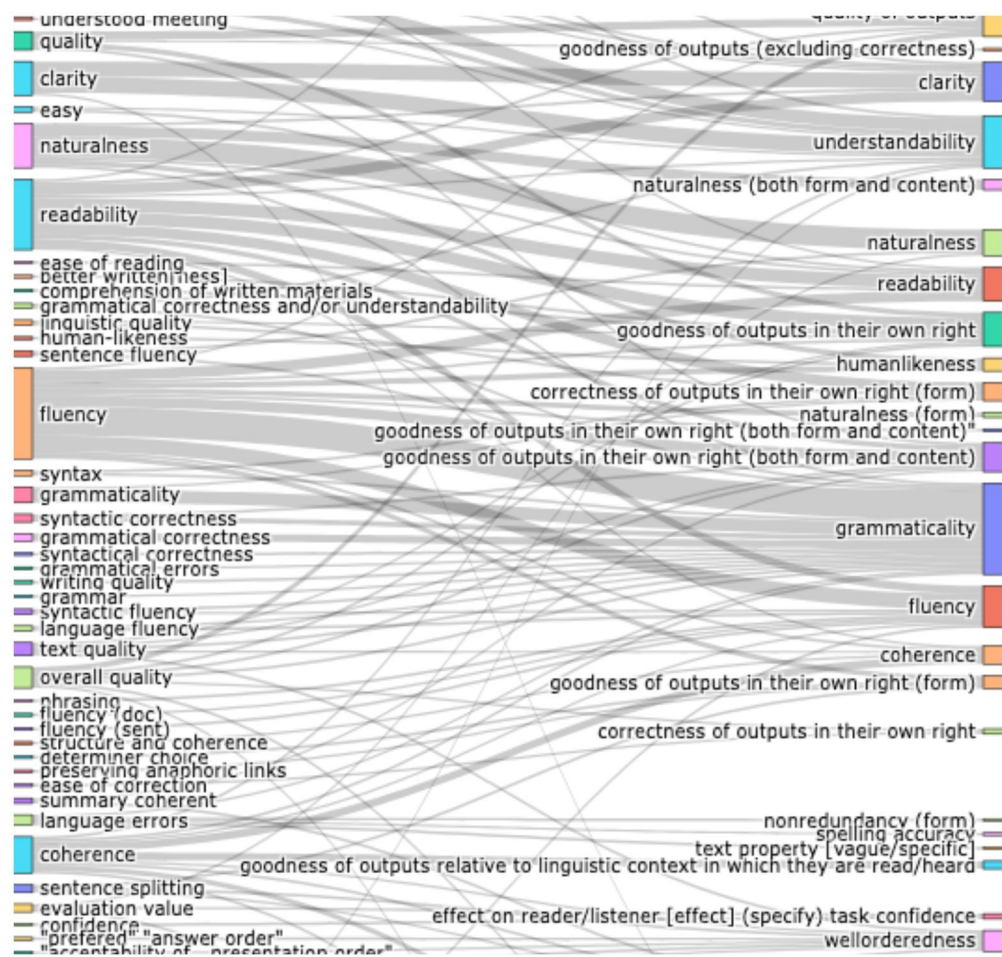
Ranking models by annotator preferences



Human evaluation - issues

- Choices of *rating scale* and of *aggregation* each make non-trivial assumptions about annotators [Ethayarajh & Jurafsky 2022](#)
- What quality criteria do the researchers try to measure?

Let's start with
agreeing on
names...



Reproducibility, reproducibility, reproducibility

- A (human evaluation) “reproducibility crisis” in NLP
- The first step is proper documentation
 - Many works do not exactly define what they measured, and how
 - Almost all do not have enough details to reproduce
- Reproducibility shared tasks: ReproHum/ReproGen/ReproNLP

Belz et al. [\[1\]](#) [\[2\]](#) [\[3\]](#) [\[4\]](#)

Human evaluation - issues

- **Annotations are expensive**
 - Experts
 - *Wisdom of the crowd*: quality control, incentives, “cheating”...
- Efficient (human) evaluation [Ashury-Tahan et al. 2024](#)

Human evaluation - emerging trends

As model capabilities improve, we require:

- More **nuanced** and **subjective** rating dimensions/scales
- More skilled **annotators** and/or more diverse viewpoints
- **More** annotations (increase statistical power) and/or **smarter** annotation (focus on harder/more informative examples)

Case study - Machine translation

Freitag et al. 2021

(a) English→German				
System	WMT↑	cSQM↑	pSQM↑	MQM↓
Human-B	(1)	(1)	(1)	(1)
Human-A	(4)	(2)	(2)	(2)
Human-P	(10)	(5)	(3)	(3)
Tohoku-AIP-NTT	(3)	(3)	(4)	(4)
OPPO	(2)	(6)	(5)	(5)
eTranslation	(9)	(7)	(7)	(6)
Tencent Translation	(6)	(4)	(6)	(7)
VolcTrans	(7)	(8)	(8)	(8)
Online-B	(5)	(9)	(9)	(9)
Online-A	(8)	(10)	(10)	(10)

← **Annotation methods**

← **System rankings**

Expert annotators

Human evaluation

1. Definition

What do we refer to? (actually, many things)

2. Intrinsic issues

Assuming we need it, what should we be aware of?

3. Emerging trends

Nowadays, do we still need it?

	Human evaluation	LLM as a judge
Varied and subjective quality criteria	✓	✓
Long and complex annotation instructions	✓	✓
Biases	✓	✓
Sensitivity to phrasing	✓	✓
Reproducibility issues	✓	✓
Specialized expert knowledge	✓	meh...
Diverse and relevant perspectives	✓	meh...
Interpretability	✓	meh...

There is always a need for human evaluation, but to really utilize its value we need to work pretty hard

Part 7: Conclusion

Modeling and evaluation trends are intertwined

- Specialized / feature-based models $\leftrightarrow$ Multiple quality criteria, single task
- General-purpose foundation models $\leftrightarrow$ Overall “goodness” across many tasks
- The “alignment” paradigm
 - Training with RLHF / RLAIIF / DPO etc. $\leftrightarrow$ Evaluation with paired preference annotations

Models

Large

Expressive

Human-level

Multi-task

Instructions

Open-ended
tasks

Dominant
measurement error

Evaluation

Expensive

Sensitive

Subjective

**Potentially
misleading**

Tutorial Sections

Efficiency

Prompts

Metrics + Human evaluation

Evaluation is harder than ever,
but there are things we can do about it

1. **First, understand your goals!**

2. **Transparency matters**

Report the exact setup, prompts, metric implementations, annotation guidelines...

3. **Reliability matters**

Do significance testing, explore robustness to different factors

4. **Many important questions to explore**

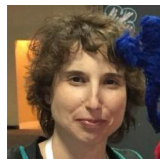
How to mitigate bias, what are the right ways to aggregate...

Reliability matters

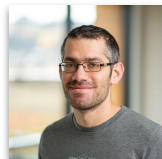
- **Absolute performance gap $\neq$ Conclusive finding**
- Significance testing
- To what extent are the results robust to:
 - Minor prompt variations?
 - Few-shot example choice?
 - Seed choice?
 - ...
- Better to sample uniformly across different sources of noise

Thank you!

Tutorial resources - <https://github.com/borgr/tutEval>



Michal
Shmueli-
Scheuer



Gabi
Stanovsky



Yotam
Perlitz



Leshem
Choshen



Ariel
Gera

References

References: Intro + Frameworks

1. A Survey on Evaluation of Large Language Models- <https://arxiv.org/abs/2307.03109>
2. Evaluating Large Language Models: A Comprehensive Survey- <https://arxiv.org/pdf/2310.19736.pdf>
3. Large Language Models: A Survey- <https://arxiv.org/pdf/2402.06196.pdf>
4. Pre-training paradigms: Recent advances in natural language processing via large pre-trained Language Models: A Survey (<https://arxiv.org/abs/2111.01243>)
5. A Survey on Fairness in Large Language Models (<https://arxiv.org/pdf/2308.10149.pdf>)
6. Holistic Evaluation of Language Models (HELM) (<https://crfm.stanford.edu/helm/latest/>)
7. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (<https://github.com/google/BIG-bench>)
8. A framework for few-shot language model evaluation (<https://github.com/EleutherAI/lm-evaluation-harness>)
9. Judging LLM-as-a-judge with mt-bench and chatbot arena (<https://arxiv.org/abs/2306.05685>)
10. AgentBench: Evaluating LLMs as Agents (<https://arxiv.org/pdf/2308.03688.pdf>)
11. Do Large Language Models Know What They Don't Know? (<https://github.com/yinzhangyue/SelfAware>)
12. Open LLM Leaderboard (https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)
13. OpenAI Evals (<https://github.com/openai/evals>)
14. Evaluating Large Language Models Trained on Code (<https://github.com/openai/human-eval>)
15. Program Synthesis with Large Language Models (<https://arxiv.org/abs/2108.07732>)
16. Dynabench: Rethinking Benchmarking in NLP (<https://arxiv.org/abs/2104.14337>)
17. PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts (<https://arxiv.org/abs/2306.04528>)
18. AlpacaEval : An Automatic Evaluator for Instruction-following Language Models (https://github.com/tatsu-lab/alpaca_eval)
19. Unitxt: Flexible, Shareable and Reusable Data Preparation and Evaluation for Generative AI (<https://arxiv.org/abs/2401.14019>)
20. Meta Comprehensive RAG Benchmark: KDD Cup 2024-
<https://www.aicrowd.com/challenges/meta-comprehensive-rag-benchmark-kdd-cup-2024/problems/end-to-end-retrieval-augmented-generation>

References: Prompts

21. State of what art - <https://arxiv.org/abs/2401.00595>
22. Selective Annotation Makes Language Models Better Few-Shot Learners- <https://arxiv.org/pdf/2209.01975>
23. The Dual Form of Neural Networks Revisited: Connecting Test Time Predictions to Training Patterns via Spotlights of Attention- <https://proceedings.mlr.press/v162/irie22a/irie22a.pdf>
24. Transformers learn in-context by gradient descent - <https://arxiv.org/abs/2212.07677>
25. Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers- <https://arxiv.org/abs/2212.10559>
26. Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation- <https://arxiv.org/abs/2305.16938>
27. Revisiting the Hypothesis: Do pretrained Transformers Learn In-Context by Gradient Descent?- <https://arxiv.org/abs/2310.08540>
28. In-Context Learning Creates Task Vectors- <https://arxiv.org/abs/2310.15916>
29. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity- <https://aclanthology.org/2022.acl-long.556/>
30. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting- <https://arxiv.org/abs/2310.11324>
31. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards- <https://arxiv.org/pdf/2402.01781>
32. Mind Your Format: Towards Consistent Evaluation of In-Context Learning Improvements - <https://arxiv.org/pdf/2401.06766>

References: Efficient evaluations

- 33. Efficient Benchmarking (of Language Models) <https://arxiv.org/pdf/2308.11696>
- 34. The Efficiency Spectrum of Large Language Models: An Algorithmic Survey. <https://arxiv.org/abs/2312.00678>
- 35. Holistic Evaluation of Language Models <https://arxiv.org/abs/2211.09110>
- 36. Anchor Points: Benchmarking Models with Much Fewer Examples <https://arxiv.org/pdf/2309.08638>
- 37. tinyBenchmarks: evaluating LLMs with fewer examples <https://arxiv.org/pdf/2402.14992>
- 38. How Predictable Are Large Language Model Capabilities? A Case Study on BIG-bench <https://arxiv.org/abs/2305.14947>

References: metrics

1. Fine-Tuned Machine Translation Metrics Struggle in Unseen Domains
2. A Comprehensive Analysis of the Effectiveness of Large Language Models as Automatic Dialogue Evaluators
3. BBScore: A Brownian Bridge Based Metric for Assessing Text Coherence
4. PreQuEL: Quality Estimation of Machine Translation Outputs in Advance
5. Inherent Biases in Reference-based Evaluation for Grammatical Error Correction and Text Simplification
6. ROUGE and BLEU scores for NLP model evaluation (Blog)
7. Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies
8. BERTScore: Evaluating Text Generation with BERT
9. Reference-less Measure of Faithfulness for Grammatical Error Correction
10. PROMETHEUS 2: An Open Source Language Model Specialized in Evaluating Other Language Models
11. Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models
12. A Call for Clarity in Reporting BLEU Scores

References: Human evaluation

1. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions - <https://aclanthology.org/2020.inlg-1.23>
2. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text - <https://arxiv.org/abs/2202.06935>
3. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text - <https://aclanthology.org/2021.acl-long.565/>
4. Agreement is overrated: A plea for correlation to assess human evaluation reliability - <https://aclanthology.org/W19-8642/>
5. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models - <https://arxiv.org/abs/2404.16019>
6. The Authenticity Gap in Human Evaluation - <https://aclanthology.org/2022.emnlp-main.406>
7. A Needle in a Haystack: An Analysis of High-Agreement Workers on MTurk for Summarization - <https://aclanthology.org/2023.acl-long.835/>
8. The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation - <https://aclanthology.org/2021.emnlp-main.97/>
9. Label-Efficient Model Selection for Text Generation - <https://arxiv.org/abs/2402.07891>
10. Missing Information, Unresponsive Authors, Experimental Flaws: The Impossibility of Assessing the Reproducibility of Previous Human Evaluations in NLP - <https://aclanthology.org/2023.insights-1.1/>
11. Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP - <https://aclanthology.org/2023.findings-acl.226/>
12. The 2022 ReproGen Shared Task on Reproducibility of Evaluations in NLG: Overview and Results - <https://aclanthology.org/2022.inlg-genchal.8/>
13. The 2023 ReproNLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results - <https://aclanthology.org/2023.humeval-1.4/>
14. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation - <https://aclanthology.org/2021.emnlp-main.97/>