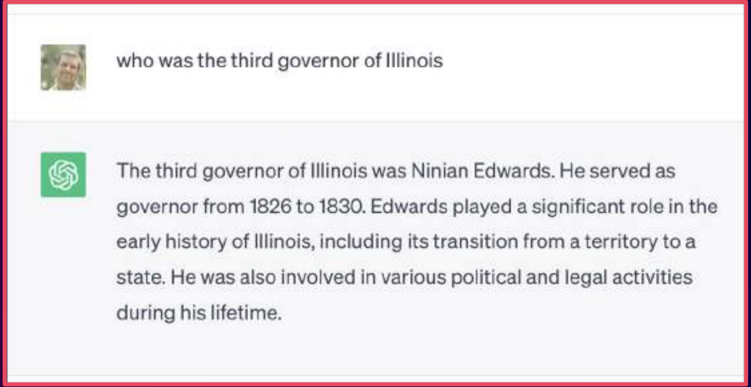
Evaluation Techniques for LLMs

https://github.com/rajshah4/LLM-Evaluation

Rajiv Shah
@rajistics
r.shah@snowflake.com

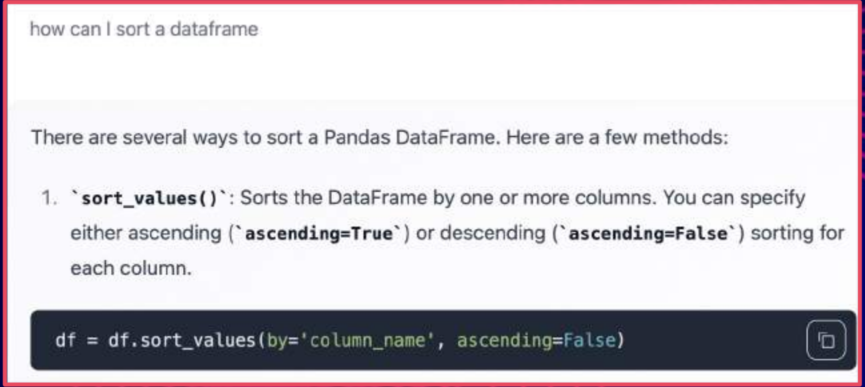
April 2024

Expectations from Generative AI



who was the third governor of Illinois

The third governor of Illinois was Ninian Edwards. He served as governor from 1826 to 1830. Edwards played a significant role in the early history of Illinois, including its transition from a territory to a state. He was also involved in various political and legal activities during his lifetime.

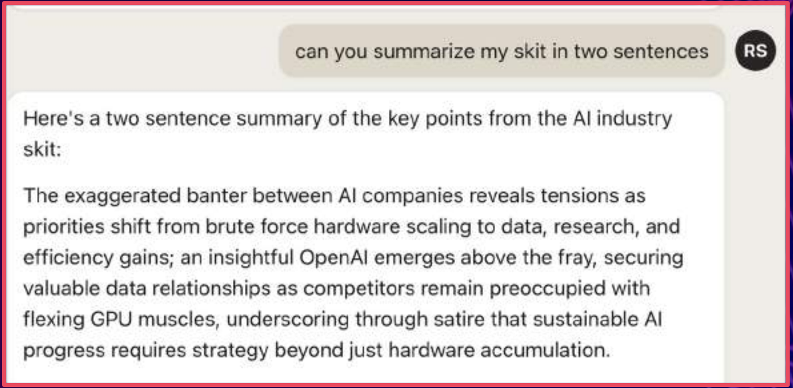


how can I sort a dataframe

There are several ways to sort a Pandas DataFrame. Here are a few methods:

1. `` `sort_values()` ``: Sorts the DataFrame by one or more columns. You can specify either ascending (`` `ascending=True` ``) or descending (`` `ascending=False` ``) sorting for each column.

```
df = df.sort_values(by='column_name', ascending=False)
```



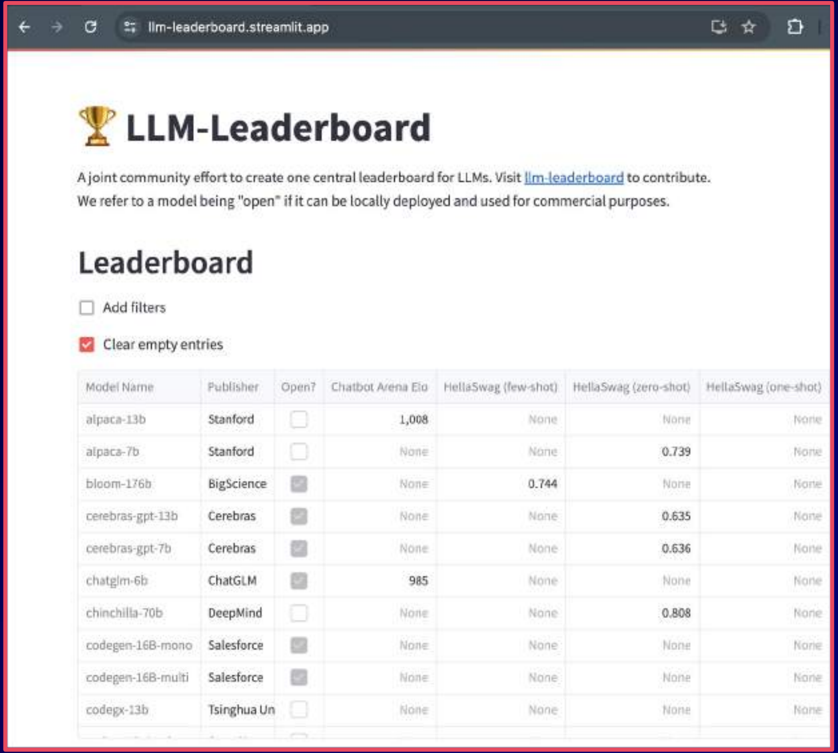
can you summarize my skit in two sentences

Here's a two sentence summary of the key points from the AI industry skit:

The exaggerated banter between AI companies reveals tensions as priorities shift from brute force hardware scaling to data, research, and efficiency gains; an insightful OpenAI emerges above the fray, securing valuable data relationships as competitors remain preoccupied with flexing GPU muscles, underscoring through satire that sustainable AI progress requires strategy beyond just hardware accumulation.

STABLE DIFFUSION

So many choices



🏆 LLM-Leaderboard

A joint community effort to create one central leaderboard for LLMs. Visit llm-leaderboard to contribute.
We refer to a model being "open" if it can be locally deployed and used for commercial purposes.

Leaderboard

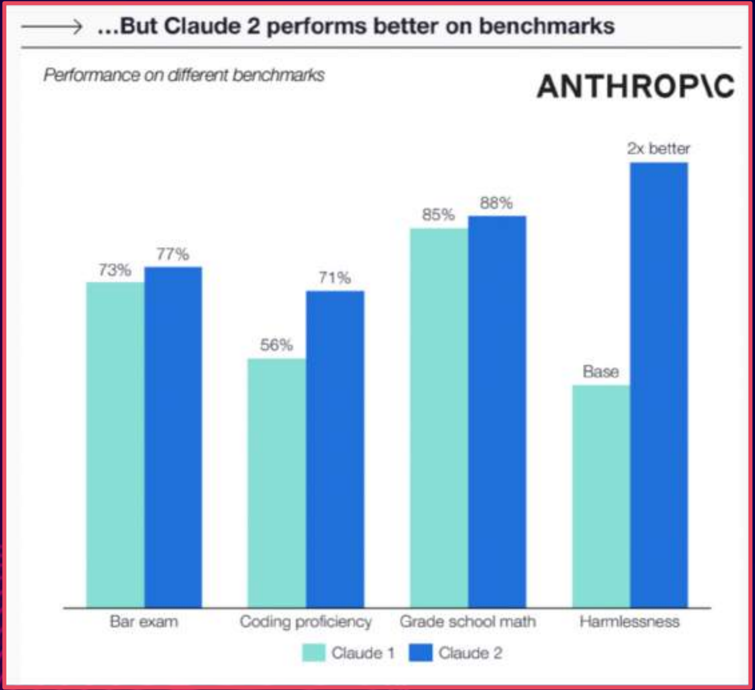
☐ Add filters

☑ Clear empty entries

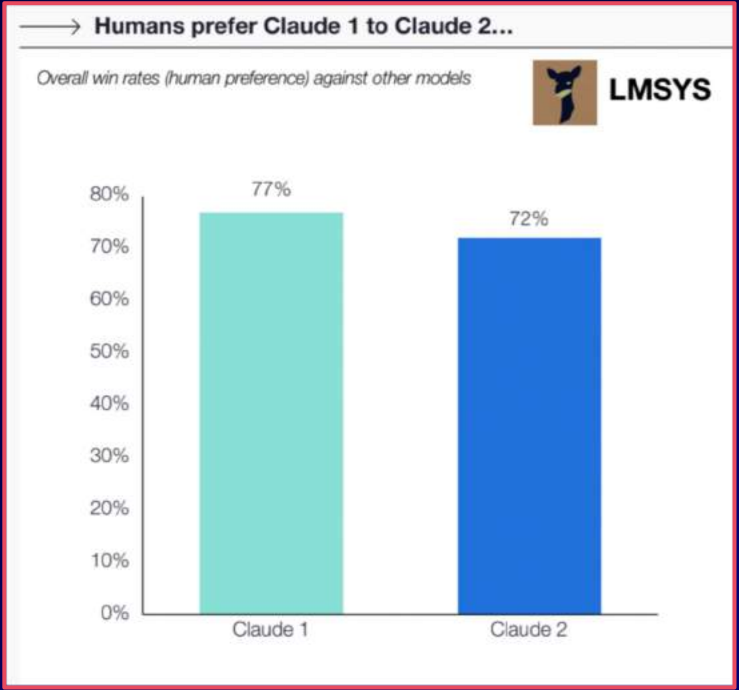
Model Name	Publisher	Open?	Chatbot Arena Elo	HellaSwag (few-shot)	HellaSwag (zero-shot)	HellaSwag (one-shot)
alpaca-13b	Stanford	☐	1,008	None	None	None
alpaca-7b	Stanford	☐	None	None	0.739	None
bloom-176b	BigScience	☑	None	0.744	None	None
cerebras-gpt-13b	Cerebras	☑	None	None	0.635	None
cerebras-gpt-7b	Cerebras	☑	None	None	0.636	None
chatglm-6b	ChatGLM	☑	985	None	None	None
chinchilla-70b	DeepMind	☐	None	None	0.808	None
codegen-16B-mono	Salesforce	☑	None	None	None	None
codegen-16B-multi	Salesforce	☑	None	None	None	None
codegx-13b	Tsinghua Un	☐	None	None	None	None

@rajistics

Not easy to evaluate


...But Claude 2 performs better on benchmarks

Performance on different benchmarks

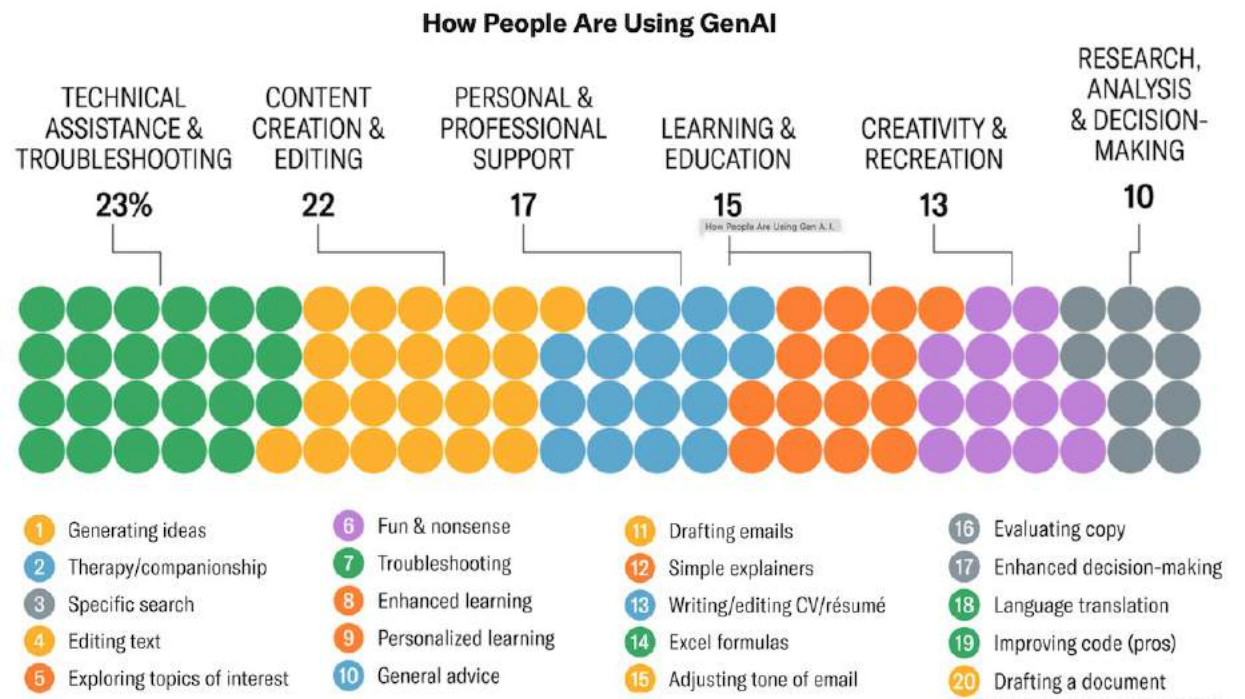

Humans prefer Claude 1 to Claude 2...

Overall win rates (human preference) against other models

@rajistics

Coatue: https://drive.google.com/file/d/1Z0NNZJPuQZuCWc__aJJSkrSbk9qJEdO7/view

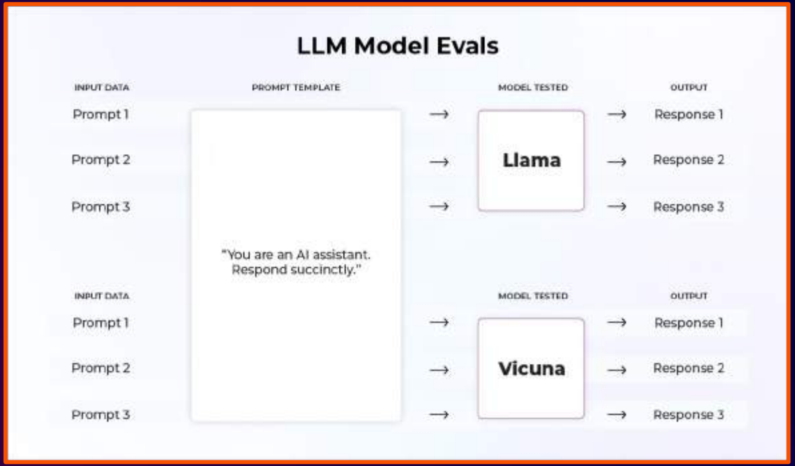
Acceptance

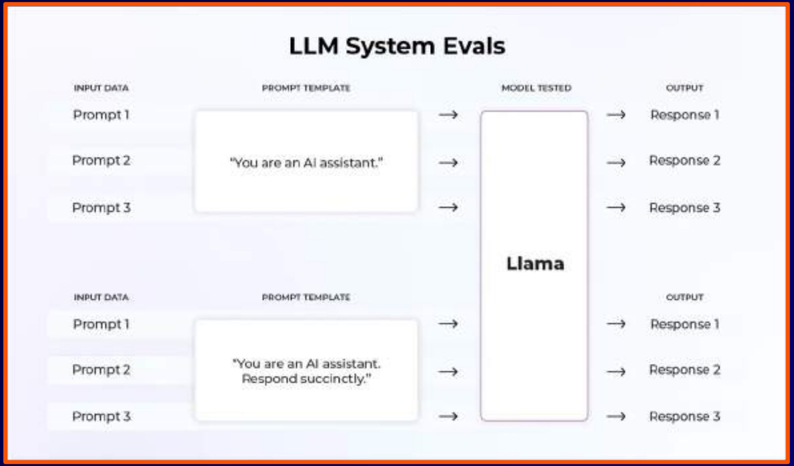


How People Are Using GenAI

https://hbr.org/2024/03/how-people-are-really-using-genai

Recognize: Social Media isn't your friend



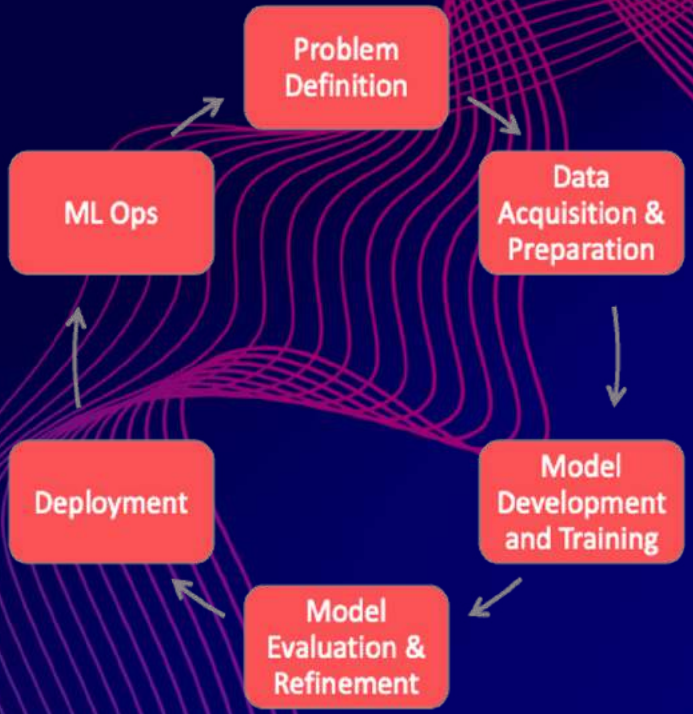


Most approaches focus on selecting from ***n*** models

Aparna - Arize Cofounder: https://towardsdatascience.com/llm-evals-setup-and-the-metrics-that-matter-2cc27e8e35f3

Fundamentals: ML Lifecycle

Evaluation is part of the entire lifecycle!



Problem Definition

Data Acquisition & Preparation

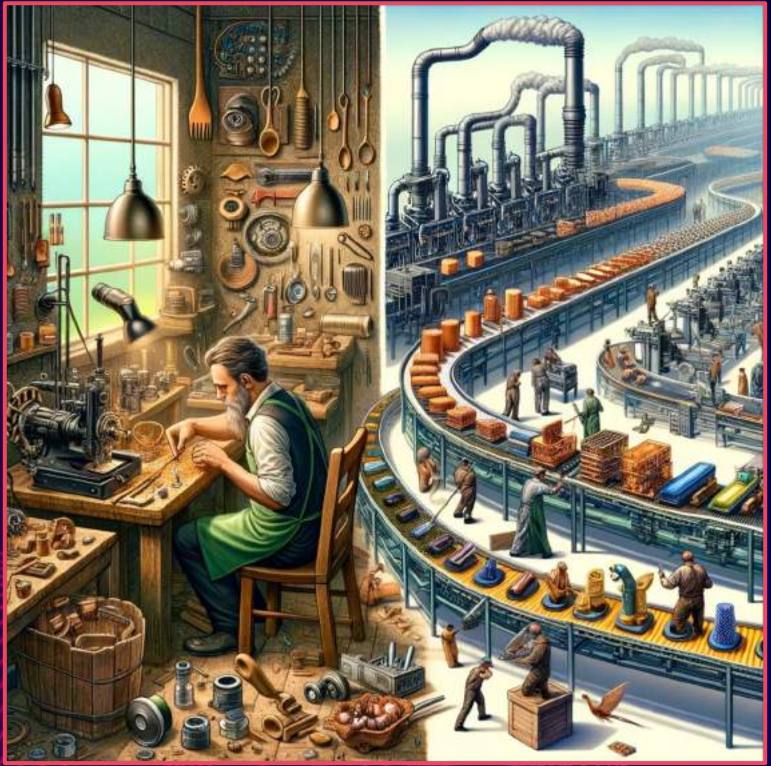
Model Development and Training

Model Evaluation & Refinement

Deployment

ML Ops

Payoff if you do evaluation right!

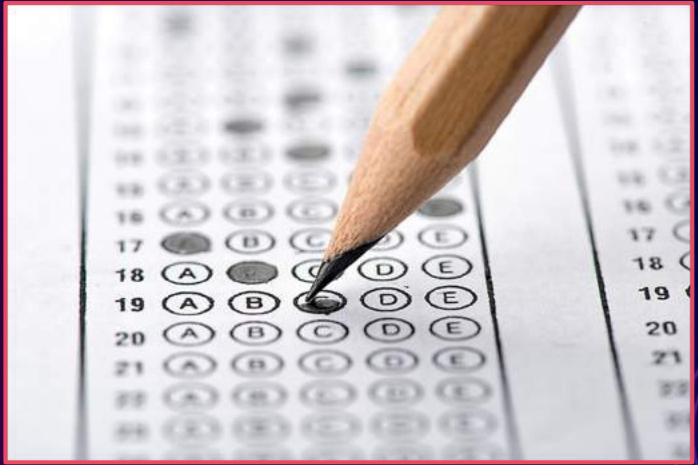


faster, better, cheaper . . .

https://unsplash.com/photos/empty-concrete-road-covered-surrounded-by-tall-tress-with-sun-rays-3Kv48NS4WUU

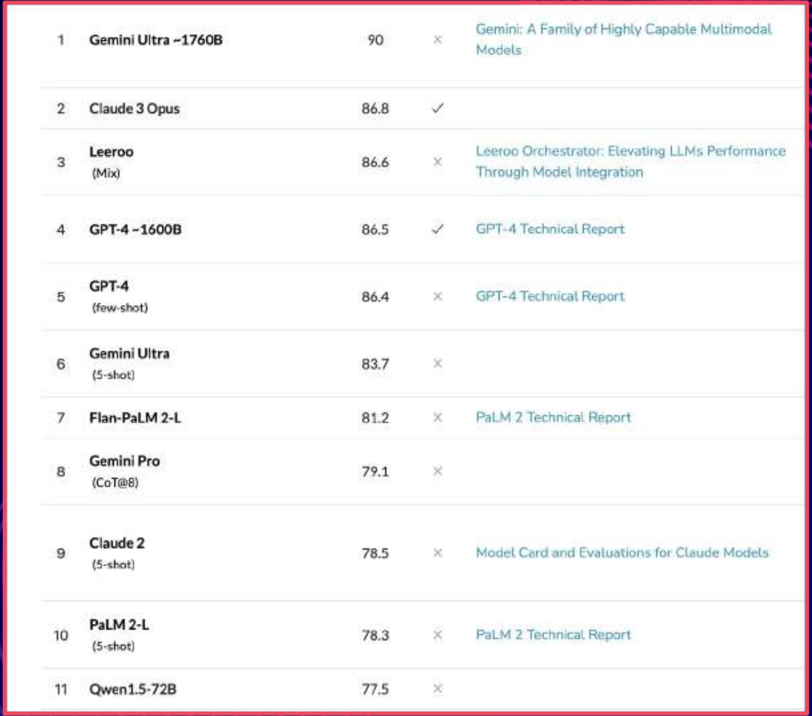
Simple Evaluation

Generative model outputs a multiple choice value:



MMLU: Massive Multitask Language Understanding

Widely used to evaluate the "smarts" of models

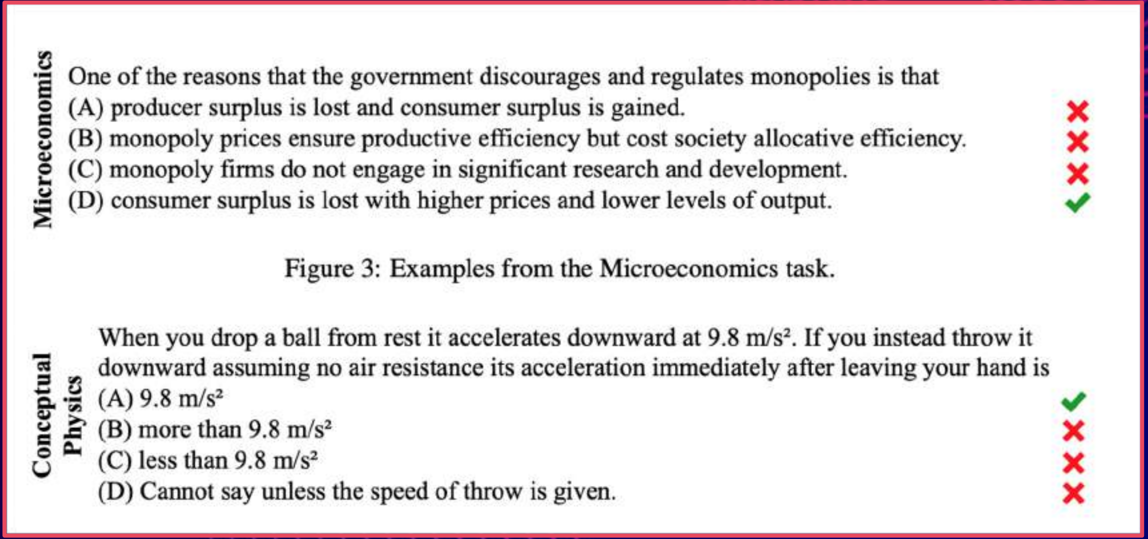


1	**Gemini Ultra ~1760B**	90	×	Gemini: A Family of Highly Capable Multimodal Models
2	**Claude 3 Opus**	86.8	✓	
3	**Leeroo** (Mix)	86.6	×	Leeroo Orchestrator: Elevating LLMs Performance Through Model Integration
4	**GPT-4 ~1600B**	86.5	✓	GPT-4 Technical Report
5	**GPT-4** (few-shot)	86.4	×	GPT-4 Technical Report
6	**Gemini Ultra** (5-shot)	83.7	×	
7	**Flan-PaLM 2-L**	81.2	×	PaLM 2 Technical Report
8	**Gemini Pro** (CoT@8)	79.1	×	
9	**Claude 2** (5-shot)	78.5	×	Model Card and Evaluations for Claude Models
10	**PaLM 2-L** (5-shot)	78.3	×	PaLM 2 Technical Report
11	**Qwen1.5-72B**	77.5	×	

https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu

MMLU: Massive Multitask Language Understanding

57 tasks: History, Computer science, mathematics



Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained. ✗
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
(C) monopoly firms do not engage in significant research and development. ✗
(D) consumer surplus is lost with higher prices and lower levels of output. ✓

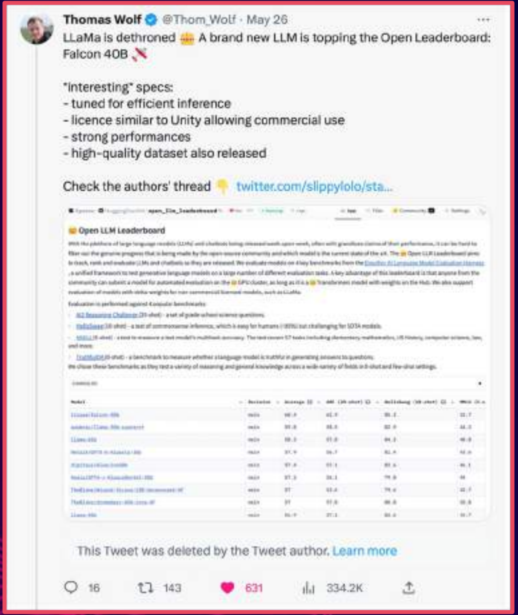
Figure 3: Examples from the Microeconomics task.

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s² ✓
(B) more than 9.8 m/s² ✗
(C) less than 9.8 m/s² ✗
(D) Cannot say unless the speed of throw is given. ✗

https://arxiv.org/abs/2009.03300

Story Time: MMLU Leaderboards



Thomas Wolf @Thom_Wolf · May 26

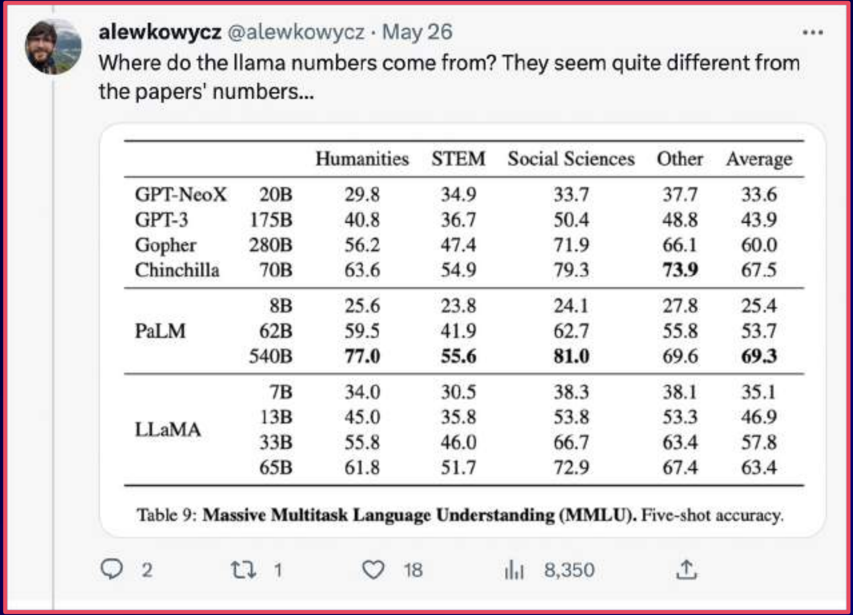
LLaMa is dethroned 👑 A brand new LLM is topping the Open Leaderboard: Falcon 40B

"interesting" specs:
- tuned for efficient inference
- licence similar to Unity allowing commercial use
- strong performances
- high-quality dataset also released

Check the authors' thread 👇 twitter.com/slippylolo/sta...

Open LLM Leaderboard

This Tweet was deleted by the Tweet author. Learn more



alewkowycz @alewkowycz · May 26

Where do the llama numbers come from? They seem quite different from the papers' numbers...

		Humanities	STEM	Social Sciences	Other	Average
GPT-NeoX	20B	29.8	34.9	33.7	37.7	33.6
GPT-3	175B	40.8	36.7	50.4	48.8	43.9
Gopher	280B	56.2	47.4	71.9	66.1	60.0
Chinchilla	70B	63.6	54.9	79.3	**73.9**	67.5
PaLM	8B	25.6	23.8	24.1	27.8	25.4
	62B	59.5	41.9	62.7	55.8	53.7
	540B	**77.0**	**55.6**	**81.0**	69.6	**69.3**
LLaMA	7B	34.0	30.5	38.3	38.1	35.1
	13B	45.0	35.8	53.8	53.3	46.9
	33B	55.8	46.0	66.7	63.4	57.8
	65B	61.8	51.7	72.9	67.4	63.4

Table 9: **Massive Multitask Language Understanding (MMLU).** Five-shot accuracy.

Why different MMLU scores?

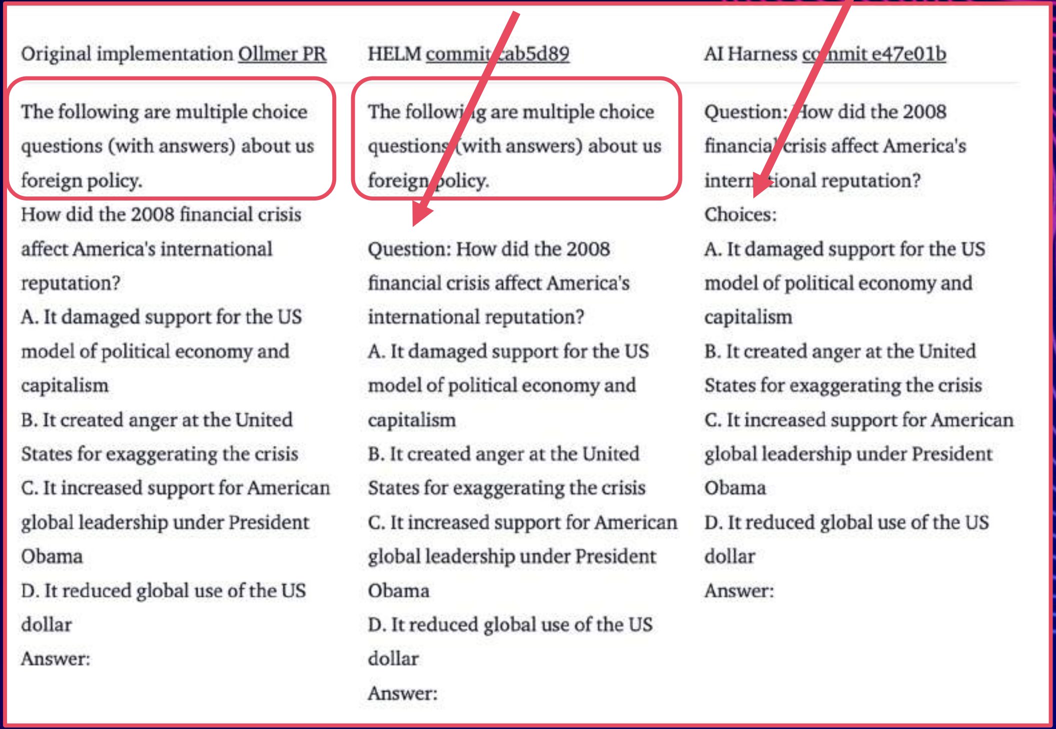
@rajistics

https://twitter.com/alewkowycz/status/1662182085073977345

Differences in prompts

Spot the differences:

- HELM extra space
- Instructions
- Question prefix?
- “Choices”



Original implementation Ollmer PR	HELM commit cab5d89	AI Harness commit e47e01b
The following are multiple choice questions (with answers) about us foreign policy. How did the 2008 financial crisis affect America's international reputation? A. It damaged support for the US model of political economy and capitalism B. It created anger at the United States for exaggerating the crisis C. It increased support for American global leadership under President Obama D. It reduced global use of the US dollar Answer:	The following are multiple choice questions (with answers) about us foreign policy. Question: How did the 2008 financial crisis affect America's international reputation? A. It damaged support for the US model of political economy and capitalism B. It created anger at the United States for exaggerating the crisis C. It increased support for American global leadership under President Obama D. It reduced global use of the US dollar Answer:	Question: How did the 2008 financial crisis affect America's international reputation? Choices: A. It damaged support for the US model of political economy and capitalism B. It created anger at the United States for exaggerating the crisis C. It increased support for American global leadership under President Obama D. It reduced global use of the US dollar Answer:

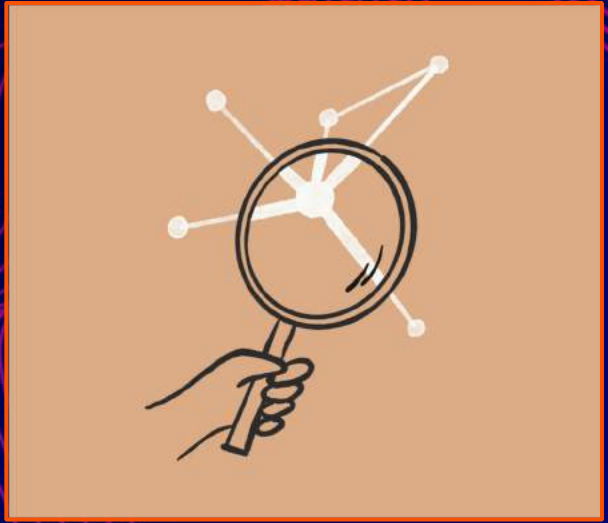
https://huggingface.co/blog/evaluating-mmlu-leaderboard

Why MMLU evaluation differed: Style

Simple formatting changes:

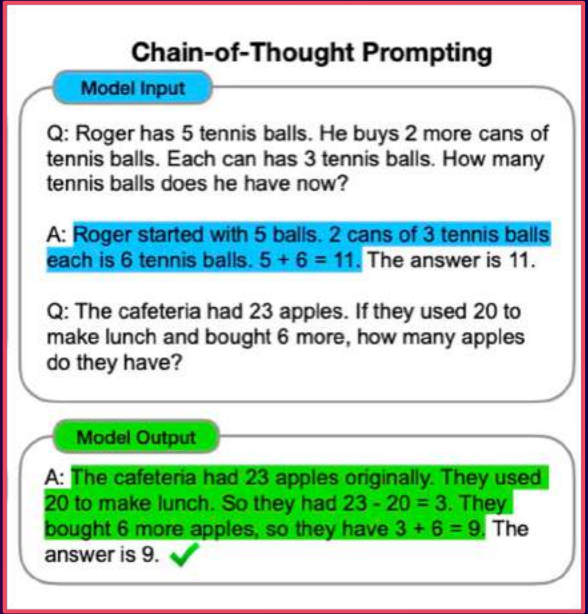
- Changing the options from (A) to (1)
- Changing the parentheses from (A) to [A]
- Adding an extra space between the option and the answer

Can lead to a ~5% change in accuracy on MMLU evaluation



https://www.anthropic.com/index/evaluating-ai-systems
When Benchmarks are Targets: https://arxiv.org/pdf/2402.01781.pdf

Prompt Engineering



Chain-of-Thought Prompting

Model Input

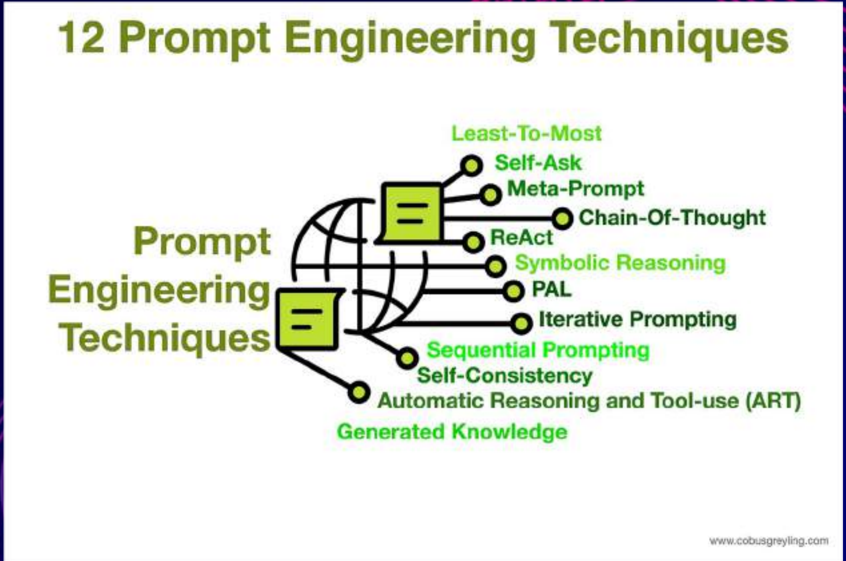
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

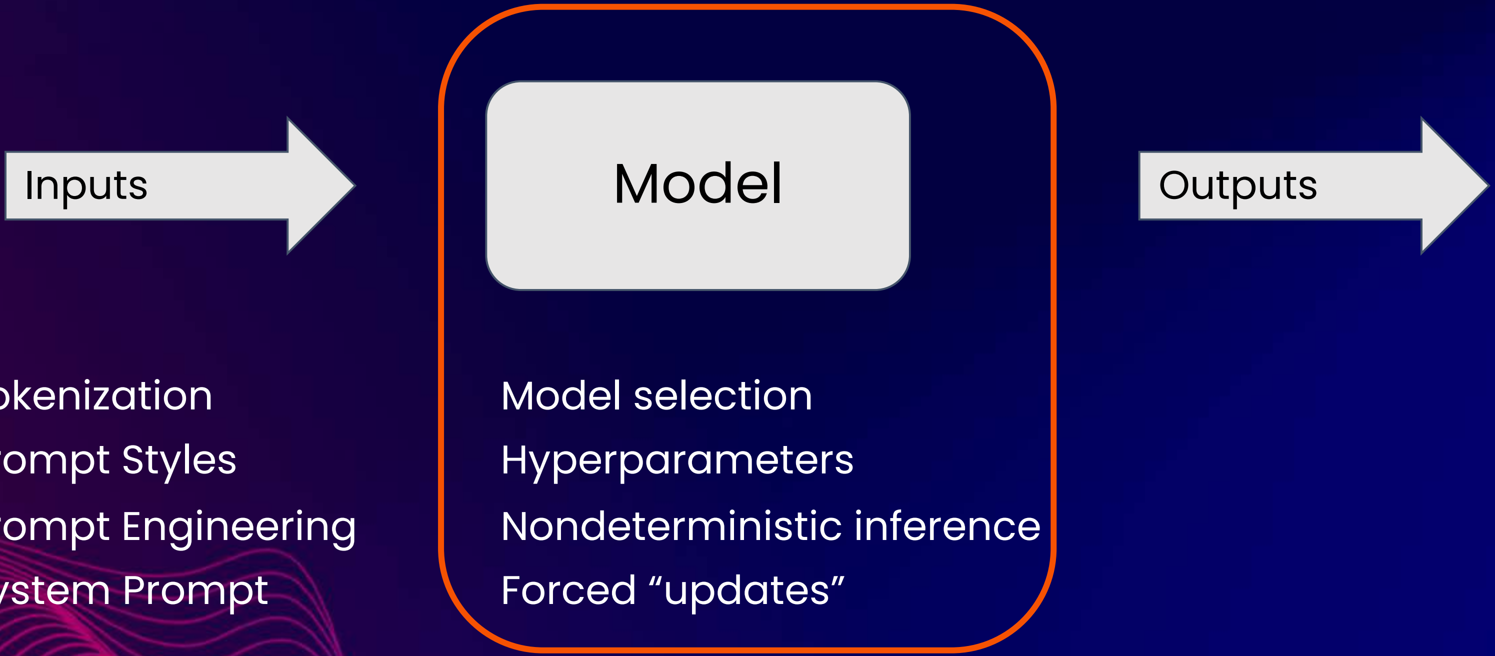
Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓



Identifying the best prompt

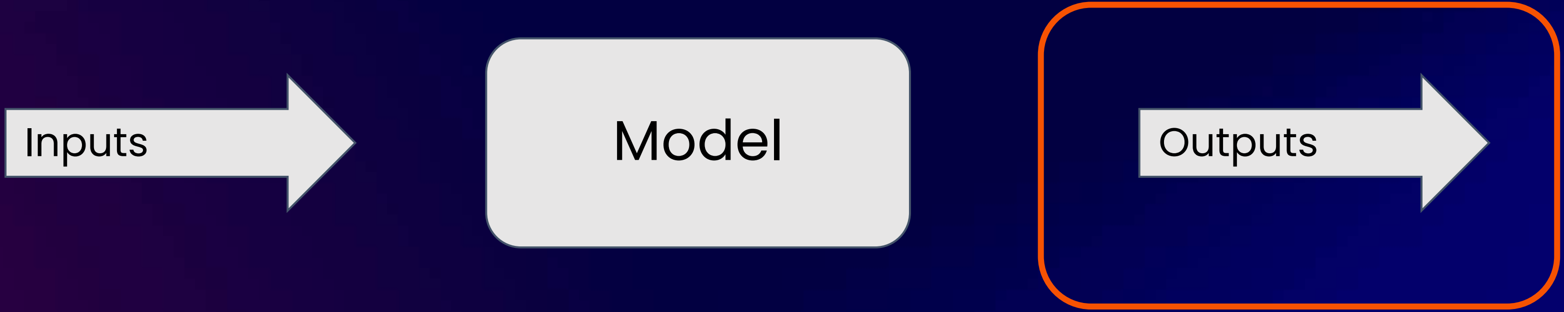
Gen AI Prediction Workflow



- Tokenization
- Prompt Styles
- Prompt Engineering
- System Prompt

- Model selection
- Hyperparameters
- Nondeterministic inference
- Forced “updates”

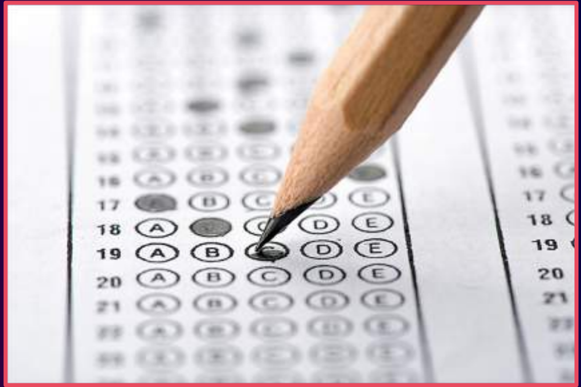
Gen AI Prediction Workflow



Tokenization
Prompt Styles
Prompt Engineering
System Prompt

Model selection
Hyperparameters
Nondeterministic inference
Forced “updates”

Generating a Multiple Choice Output



Require one of the choices

First Letter Approach

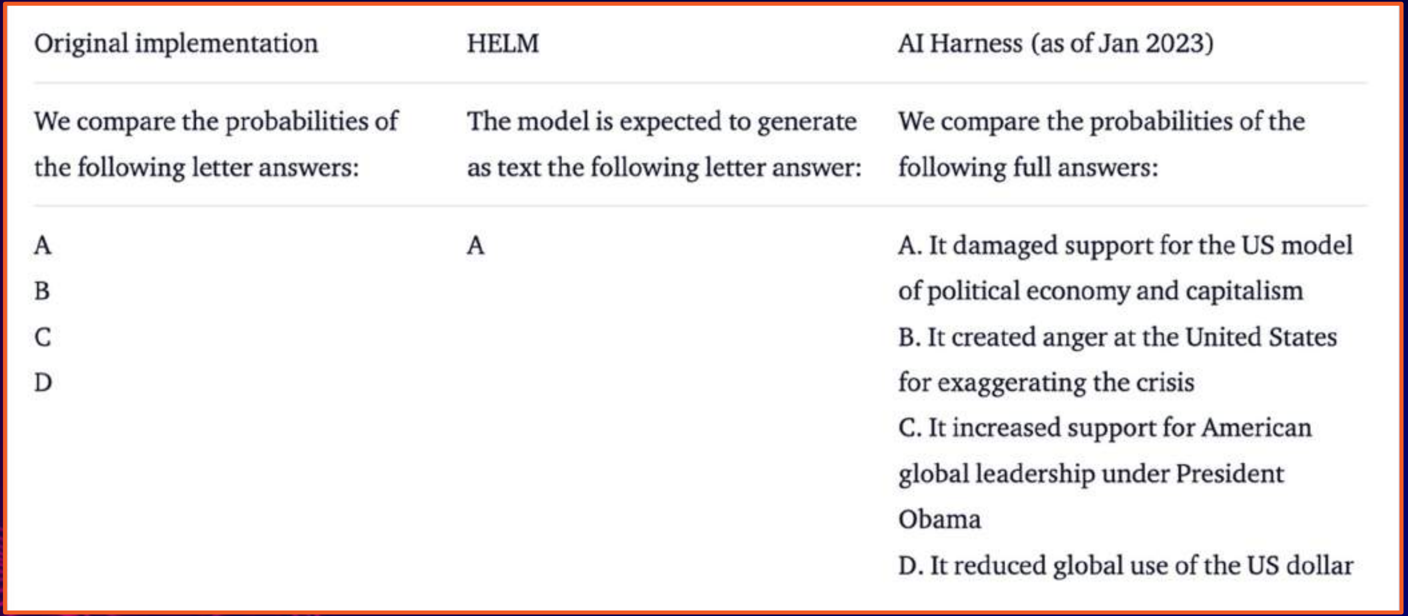
✅ C - Washington
❌ Washington, Choice C

✅ C - Washington
✅ Washington, Choice C

Entire Answer

@rajistics

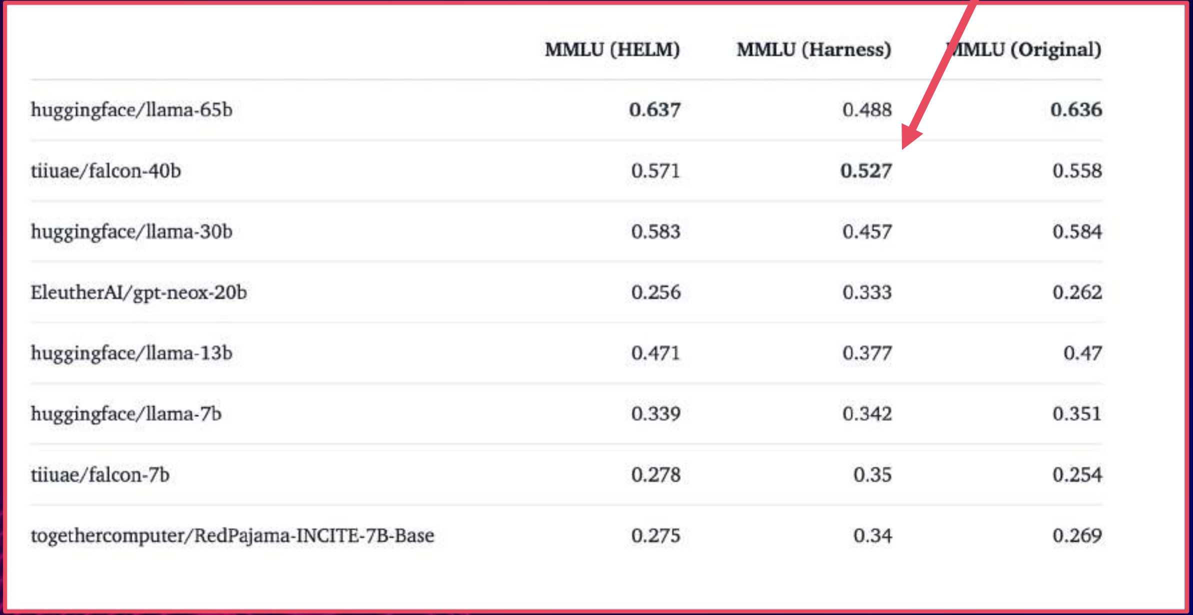
Evaluating MMLU: different outputs



Original implementation	HELM	AI Harness (as of Jan 2023)
We compare the probabilities of the following letter answers:	The model is expected to generate as text the following letter answer:	We compare the probabilities of the following full answers:
A B C D	A	A. It damaged support for the US model of political economy and capitalism B. It created anger at the United States for exaggerating the crisis C. It increased support for American global leadership under President Obama D. It reduced global use of the US dollar

https://huggingface.co/blog/evaluating-mmlu-leaderboard

Evaluating MMLU: different scores

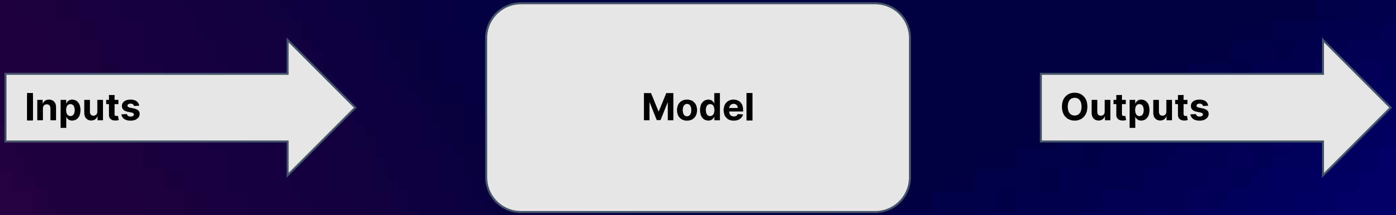


	MMLU (HELM)	MMLU (Harness)	MMLU (Original)
huggingface/llama-65b	**0.637**	0.488	**0.636**
tiiuae/falcon-40b	0.571	**0.527**	0.558
huggingface/llama-30b	0.583	0.457	0.584
EleutherAI/gpt-neox-20b	0.256	0.333	0.262
huggingface/llama-13b	0.471	0.377	0.47
huggingface/llama-7b	0.339	0.342	0.351
tiiuae/falcon-7b	0.278	0.35	0.254
togethercomputer/RedPajama-INCITE-7B-Base	0.275	0.34	0.269

😬

https://huggingface.co/blog/evaluating-mmlu-leaderboard

Gen AI Prediction Workflow



Tokenization
Prompt Styles
Prompt Engineering
System Prompt

Model selection
Hyperparameters
Nondeterministic inference
Forced “updates”

Output evaluation

PRO TIP: PLAN ON MULTIPLE ITERATIONS WHEN EVALUATING LLMS

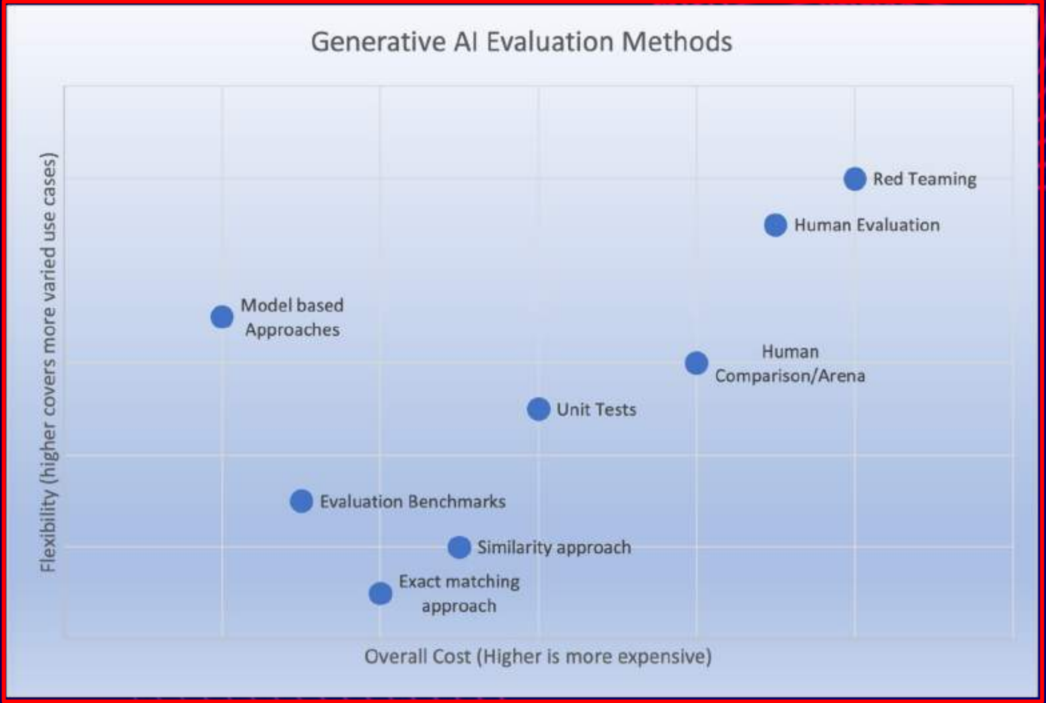
You aren't helping

😩

https://unsplash.com/photos/woman-in-brown-sweater-covering-her-face-with-her-hand-_sh9vkVbVgo

Methods for evaluating Generative AI

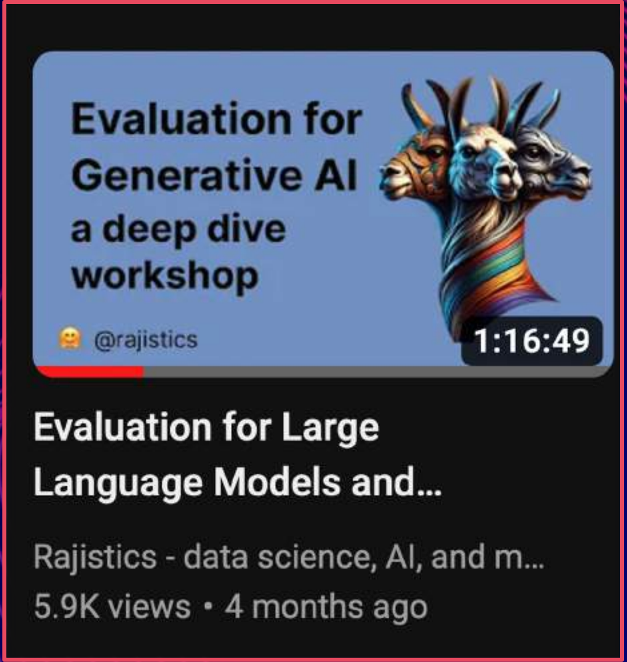
- Exact matching approach
- Similarity approach
- Unit Tests
- Evaluation Benchmarks
- Human Evaluation
- Human Comparison/Arena
- Model based Approaches
- Red Teaming



Raj guess

Methods for evaluating Generative AI

- Exact matching approach
- Similarity approach
- Unit Tests
- Evaluation Benchmarks
- Human Evaluation
- Human Comparison/Arena
- Model based Approaches
- Red Teaming



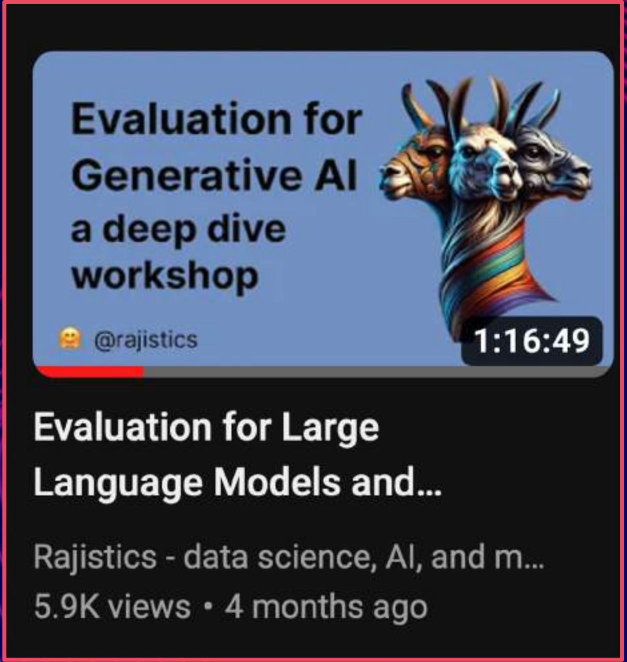
Evaluation for Large Language Models and...

Rajistics - data science, AI, and m...
5.9K views • 4 months ago

https://youtu.be/iQI03pQlYWY

Methods for evaluating Generative AI

- Exact matching approach
- Similarity approach
- Unit Tests
- Evaluation Benchmarks
- Human Evaluation
- Human Comparison/Arena
- Model based Approaches
- Red Teaming



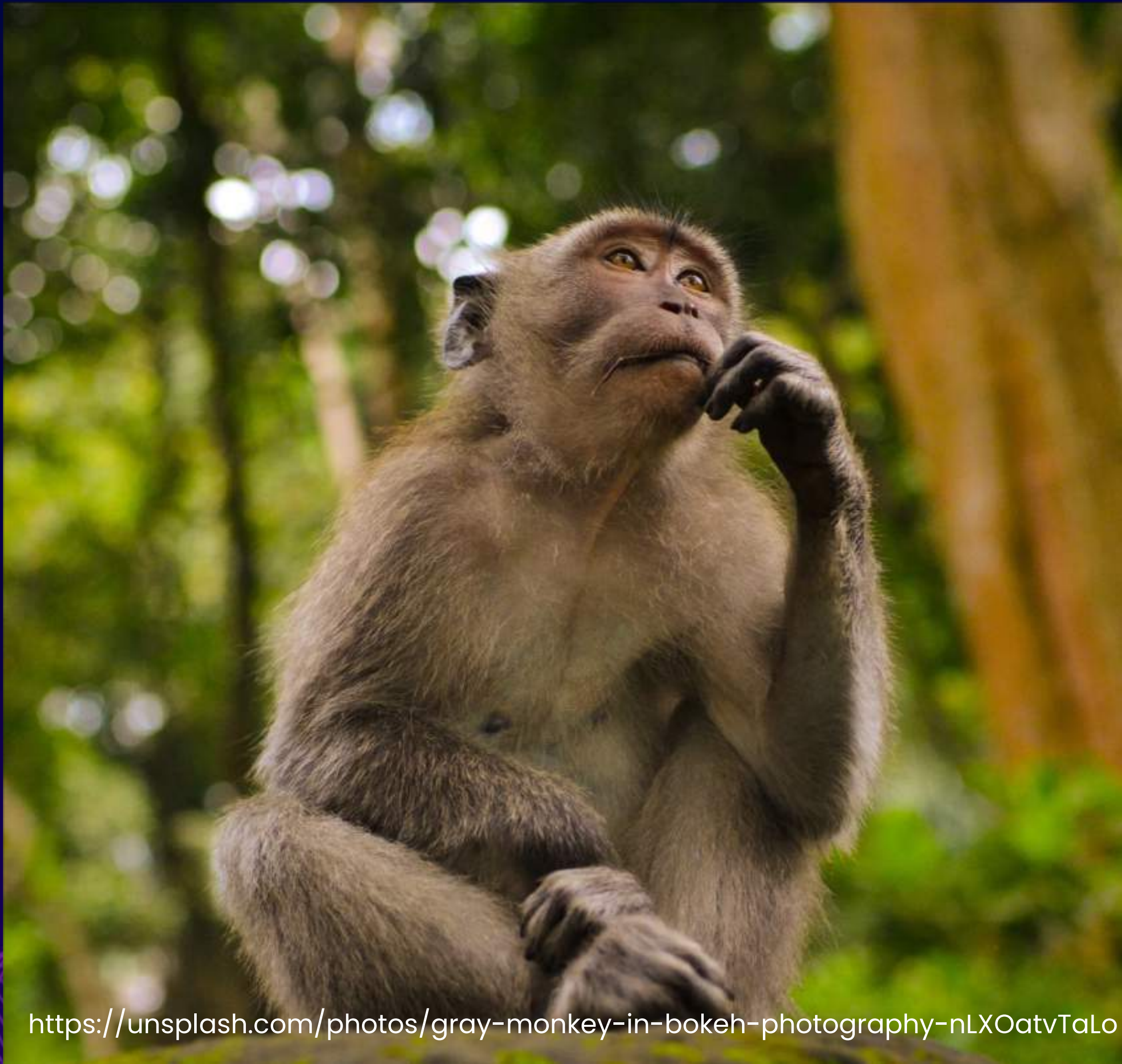
Evaluation for Large Language Models and...

Rajistics - data science, AI, and m...

5.9K views • 4 months ago

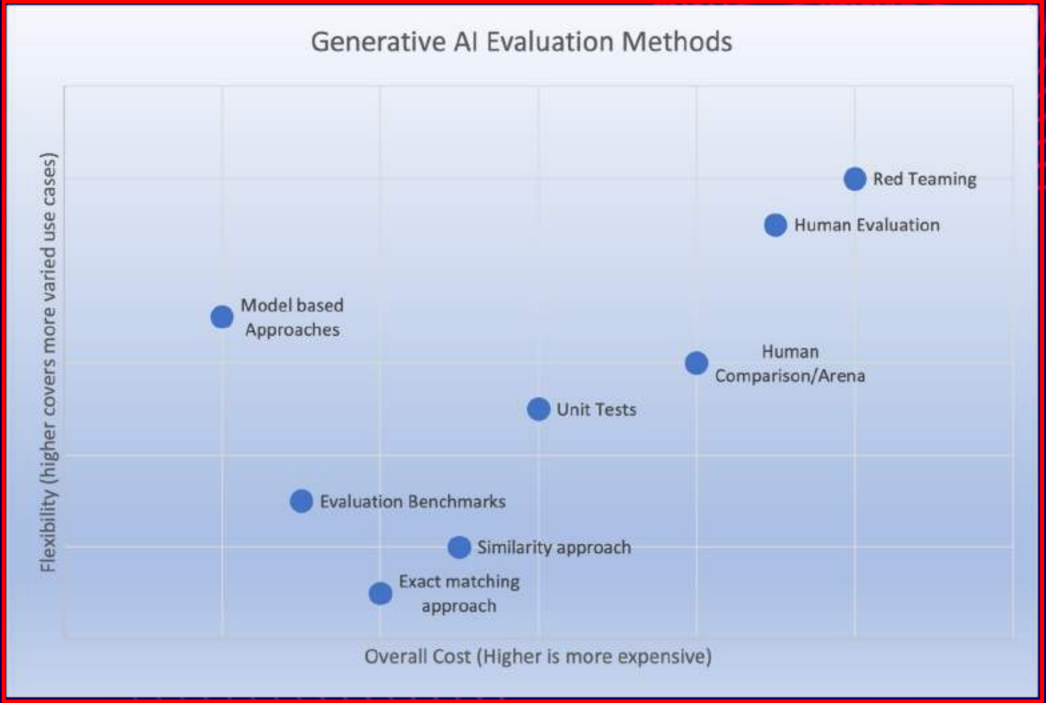
https://youtu.be/iQI03pQlYWY

You have my attention 🤔


https://unsplash.com/photos/gray-monkey-in-bokeh-photography-nLXOatvTaLo

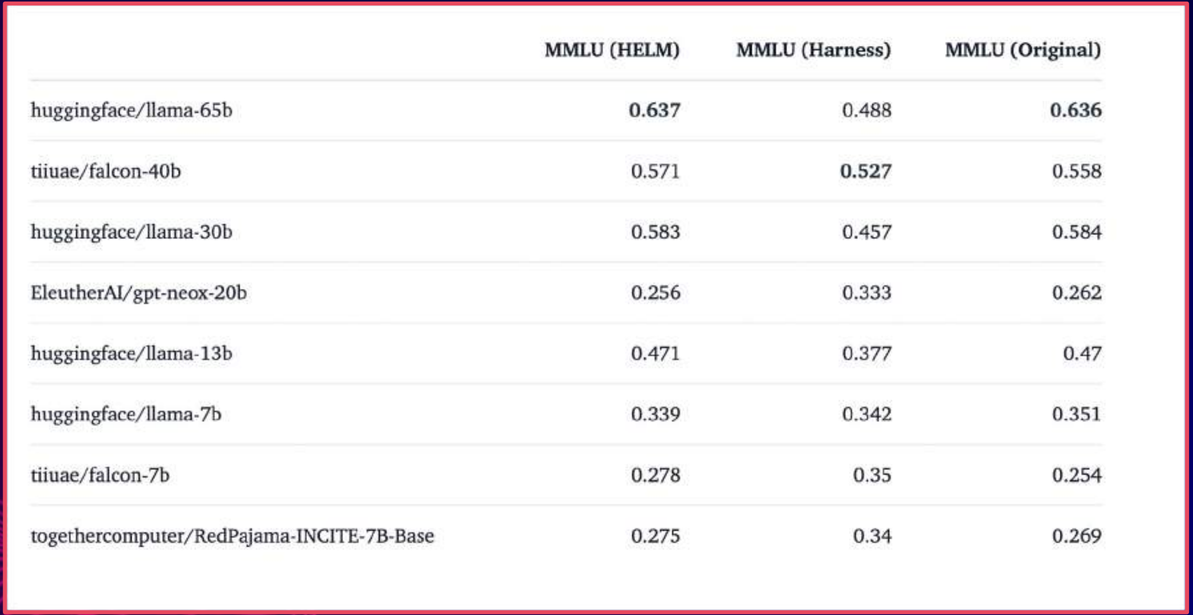
Methods for evaluating Generative AI

- Exact matching approach
- Similarity approach
- Unit Tests
- Evaluation Benchmarks
- Human Evaluation
- Human Comparison/Arena
- Model based Approaches
- Red Teaming



@rajistics

Evaluating MMLU: different scores



	MMLU (HELM)	MMLU (Harness)	MMLU (Original)
huggingface/llama-65b	**0.637**	0.488	**0.636**
tiiuae/falcon-40b	0.571	**0.527**	0.558
huggingface/llama-30b	0.583	0.457	0.584
EleutherAI/gpt-neox-20b	0.256	0.333	0.262
huggingface/llama-13b	0.471	0.377	0.47
huggingface/llama-7b	0.339	0.342	0.351
tiiuae/falcon-7b	0.278	0.35	0.254
togethercomputer/RedPajama-INCITE-7B-Base	0.275	0.34	0.269

https://huggingface.co/blog/evaluating-mmlu-leaderboard

even more benchmarks

- Advanced Sommelier (theory knowledge)
- AI2 Reasoning Challenge (ARC) 2018
- ALFW
- AMC 103
- AMC 123
- AP Art History
- AP Biology
- AP Calculus BC
- AP Chemistry
- AP English Language and Composition
- AP English Literature and Composition
- AP Environmental Science
- AP Macroeconomics
- AP Microeconomics
- AP Physics 2
- AP Psychology
- AP Statistics
- AP US Government
- AP US History
- AP World History
- APPS (Code)
- ARC
- bAbI
- BoolQ
- C-Objects
- Certified Sommelier (theory knowledge)
- CivilComments
- CNN/DailyMail
- Codeforces Rating
- CoQA
- Data imputation
- DROP
- Dyck
- Entity matching
- Gorilla-TH
- Graduate Record Examination (GRE) Quantitative
- Graduate Record Examination (GRE) Verbal
- Graduate Record Examination (GRE) Writing
- GSM8K
- HaluEval
- HellaSwag
- HotpotQA
- HumanEval
- IMDB
- Introductory Sommelier (theory knowledge)
- LAMBADA
- Leetcode (easy)
- Leetcode (hard)
- Leetcode (medium)
- LegalSupport
- LogiQA
- LSAT
- MATH
- MATH (chain-of-thoughts)
- Medical Knowledge Self-Assessment Program
- MMLU
- MS MARCO (regular)
- MS MARCO (TREC)
- NarrativeQA
- NaturalQuestions (closed-book)
- NaturalQuestions (open-book)
- OBQA
- OpenbookQA
- Penguins
- PIQA
- QuAC
- RACE
- RAFT
- ReClor
- RTP
- SAT Evidence-Based Reading & Writing
- SAT Math
- SIQA
- SocialQA
- Synthetic reasoning (abstract symbols)
- Synthetic reasoning (natural language)
- TfQA
- TruthfulQA
- Uniform Bar Exam (MBE+MEE+MPT)
- USABO Semifinal Exam 2020
- USNCO Local Section Exam 2022
- Webshop
- WikiFact
- WinoGender
- WinoGrande
- XSUM

https://www.lesswrong.com/posts/BRviTDFMvEHgA5iFs/list-of-commonly-used-benchmarks-for-llms

Pro tip: Build your own benchmark / leaderboards

Every organization has multiple use cases

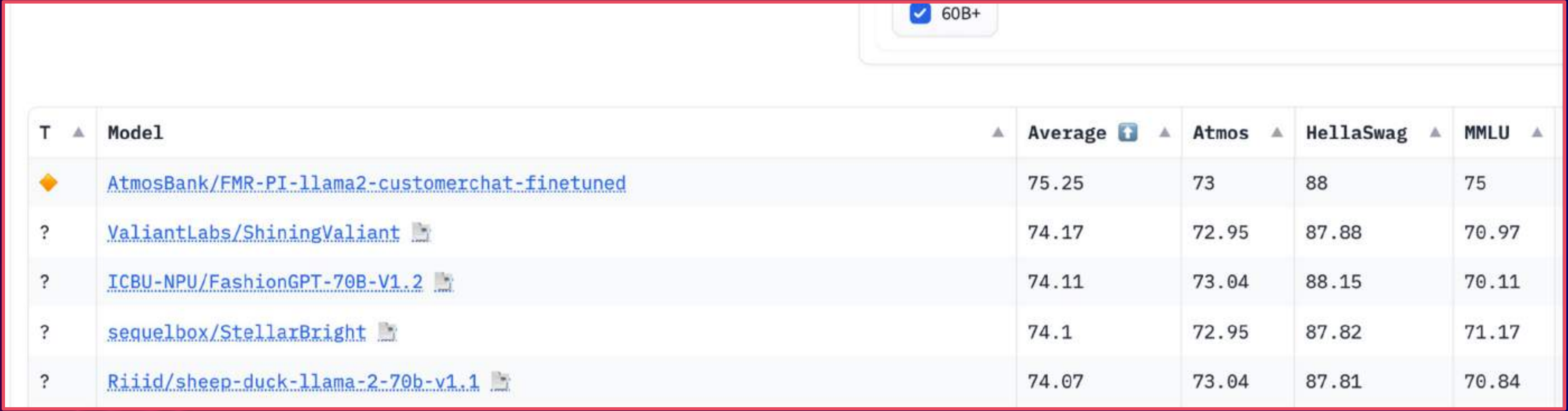
Build a custom benchmark

Examples of Domain specific:

- LegalBench
- AgentsBench
- OWL - IT Operations

Legal Bench: https://arxiv.org/abs/2308.11462
Agent Bench: https://arxiv.org/abs/2308.03688
OWL: https://arxiv.org/pdf/2309.09298.pdf

Pro tip: Build your own benchmark / leaderboards



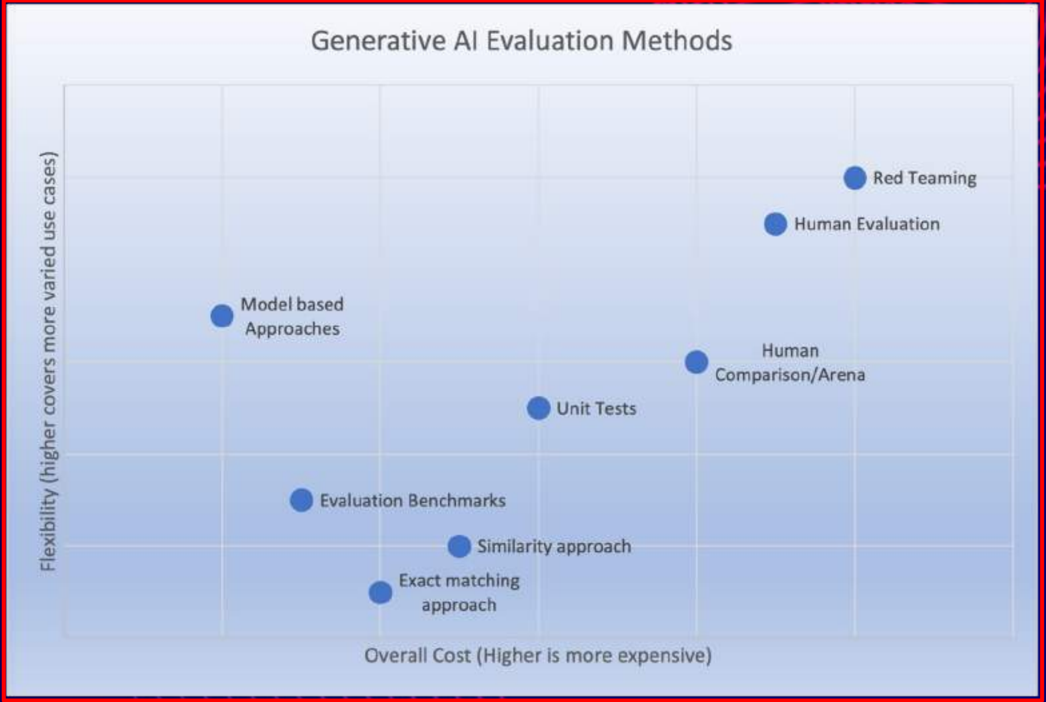
☑ 60B+

T ▲	Model ▲	Average ⬆️ ▲	Atmos ▲	HellaSwag ▲	MMLU ▲
🔶	AtmosBank/FMR-PI-llama2-customerchat-finetuned	75.25	73	88	75
?	ValiantLabs/ShiningValiant	74.17	72.95	87.88	70.97
?	ICBU-NPU/FashionGPT-70B-V1.2	74.11	73.04	88.15	70.11
?	sequelbox/StellarBright	74.1	72.95	87.82	71.17
?	Riiid/sheep-duck-llama-2-70b-v1.1	74.07	73.04	87.81	70.84

https://huggingface.co/spaces/AtmosBank/Atmos_Leaderboard?logs=container

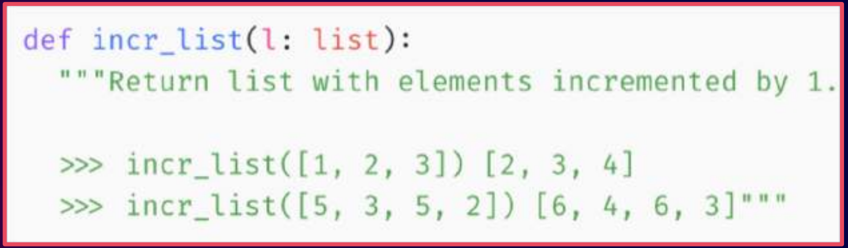
Methods for evaluating Generative AI

- Exact matching approach
- Similarity approach
- Unit Tests
- Evaluation Benchmarks
- Human Evaluation
- Human Comparison/Arena
- Model based Approaches
- Red Teaming



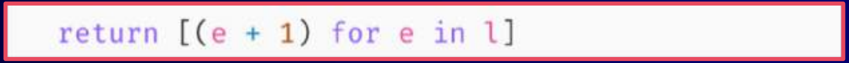
Raj guess

Evaluating Code: Python



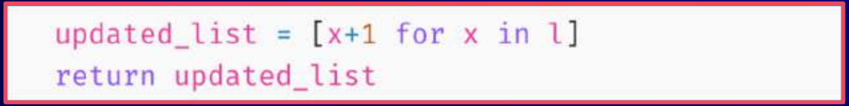
```
def incr_list(l: list):
  """Return list with elements incremented by 1.

  >>> incr_list([1, 2, 3]) [2, 3, 4]
  >>> incr_list([5, 3, 5, 2]) [6, 4, 6, 3]"""
```

Candidate solution:



```
    return [(e + 1) for e in l]
```

Reference solution:

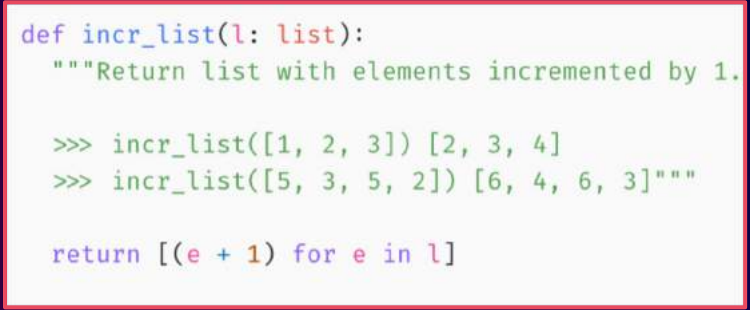


```
    updated_list = [x+1 for x in l]
    return updated_list
```

Leandro &
https://arxiv.org/pdf/2107.03374.pdf

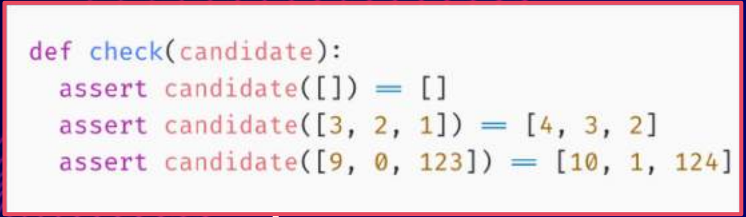
Evaluating Code with Unit Test

Candidate solution:



```
def incr_list(l: list):
    """Return list with elements incremented by 1.

    >>> incr_list([1, 2, 3]) [2, 3, 4]
    >>> incr_list([5, 3, 5, 2]) [6, 4, 6, 3]"""

    return [(e + 1) for e in l]
```

Test Cases:

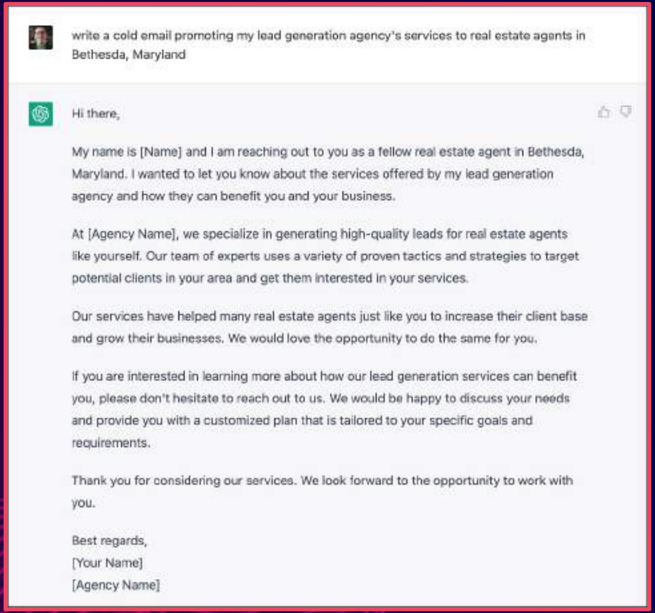


```
def check(candidate):
    assert candidate([]) == []
    assert candidate([3, 2, 1]) == [4, 3, 2]
    assert candidate([9, 0, 123]) == [10, 1, 124]
```

Pass: yes/no

https://arxiv.org/pdf/2107.03374.pdf

Unit Tests Beyond Code



write a cold email promoting my lead generation agency's services to real estate agents in Bethesda, Maryland

Hi there,

My name is [Name] and I am reaching out to you as a fellow real estate agent in Bethesda, Maryland. I wanted to let you know about the services offered by my lead generation agency and how they can benefit you and your business.

At [Agency Name], we specialize in generating high-quality leads for real estate agents like yourself. Our team of experts uses a variety of proven tactics and strategies to target potential clients in your area and get them interested in your services.

Our services have helped many real estate agents just like you to increase their client base and grow their businesses. We would love the opportunity to do the same for you.

If you are interested in learning more about how our lead generation services can benefit you, please don't hesitate to reach out to us. We would be happy to discuss your needs and provide you with a customized plan that is tailored to your specific goals and requirements.

Thank you for considering our services. We look forward to the opportunity to work with you.

Best regards,
[Your Name]
[Agency Name]

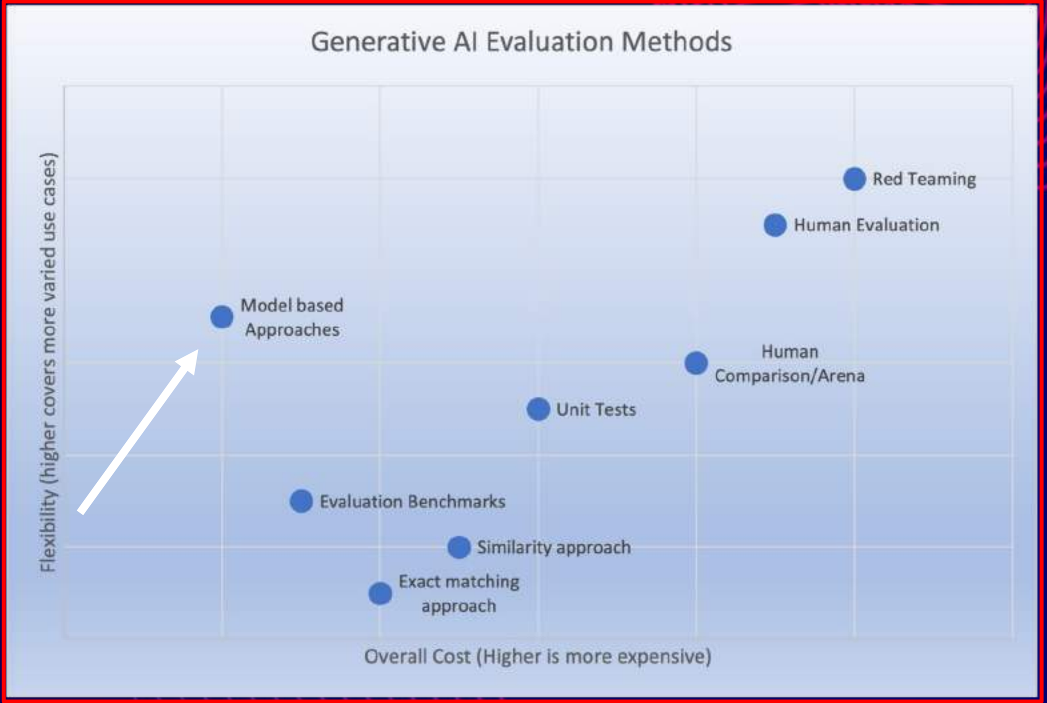
Properties of Emails?

- First/Last Name?
- Grammar/spelling
- Concise?
- Verify actions?
- Tone? - is it polite

https://towardsdatascience.com/testing-large-language-models-like-we-test-software-92745d28a359

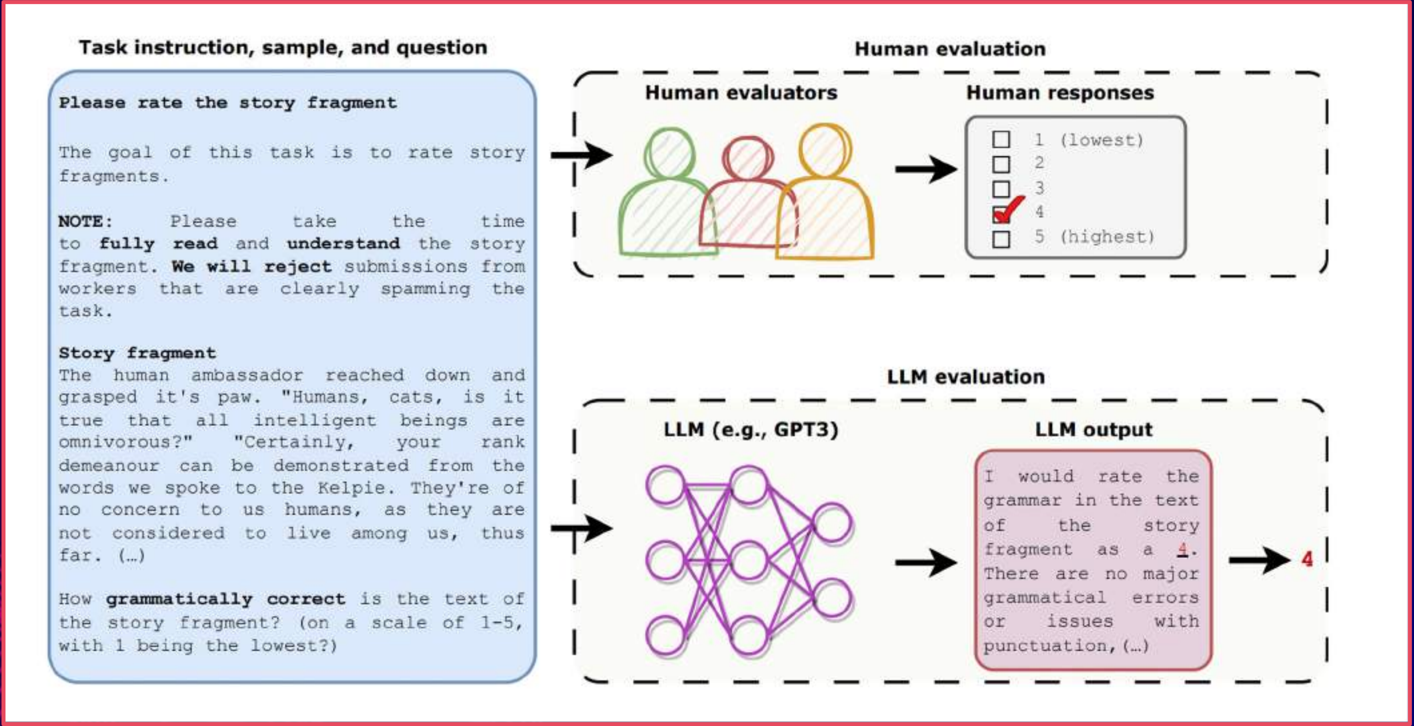
Methods for evaluating Generative AI

- Exact matching approach
- Similarity approach
- Unit Tests
- Evaluation Benchmarks
- Human Evaluation
- Human Comparison/Arena
- Model based Approaches
- Red Teaming



Raj guess

Model based evaluation



Please rate the story fragment

The goal of this task is to rate story fragments.

NOTE: Please take the time to **fully read** and **understand** the story fragment. **We will reject** submissions from workers that are clearly spamming the task.

Story fragment
The human ambassador reached down and grasped it's paw. "Humans, cats, is it true that all intelligent beings are omnivorous?" "Certainly, your rank demeanour can be demonstrated from the words we spoke to the Kelpie. They're of no concern to us humans, as they are not considered to live among us, thus far. (…)

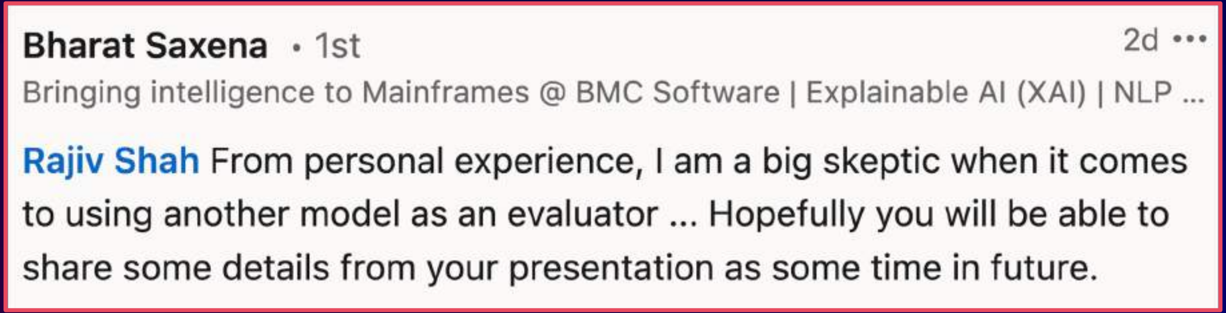
How **grammatically correct** is the text of the story fragment? (on a scale of 1-5, with 1 being the lowest?)

I would rate the grammar in the text of the story fragment as a 4. There are no major grammatical errors or issues with punctuation, (…)

@rajistics

https://arxiv.org/pdf/2305.01937.pdf

C’mon Man - This isn’t going to work



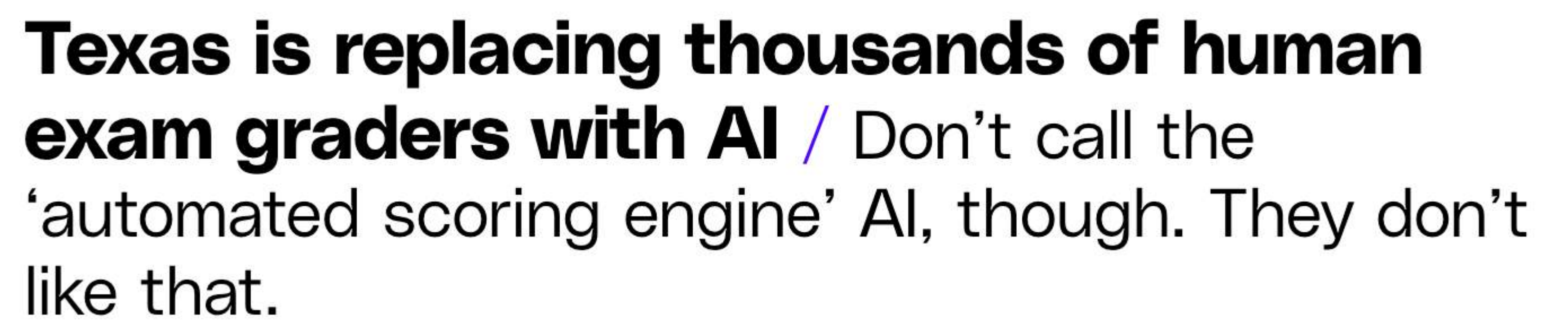
Bharat Saxena • 1st

2d •••

Bringing intelligence to Mainframes @ BMC Software | Explainable AI (XAI) | NLP ...

Rajiv Shah From personal experience, I am a big skeptic when it comes to using another model as an evaluator ... Hopefully you will be able to share some details from your presentation as some time in future.

It works in Texas

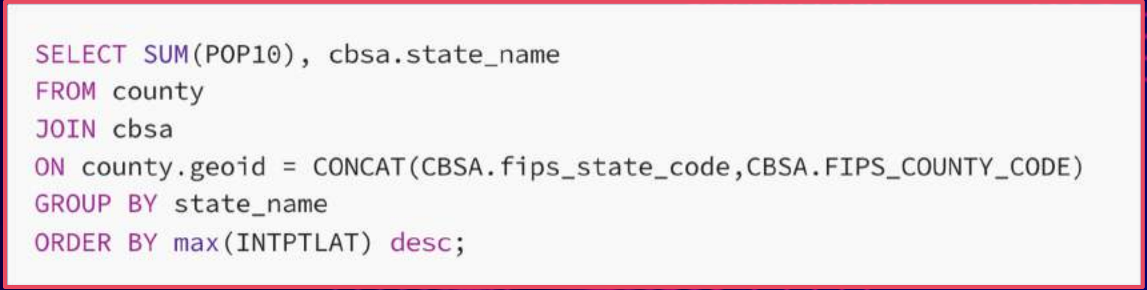


Texas is replacing thousands of human exam graders with AI / Don’t call the ‘automated scoring engine’ AI, though. They don’t like that.

@rajistics https://www.theverge.com/2024/4/10/24126206/texas-staar-exam-graders-ai-automated-scoring-engine

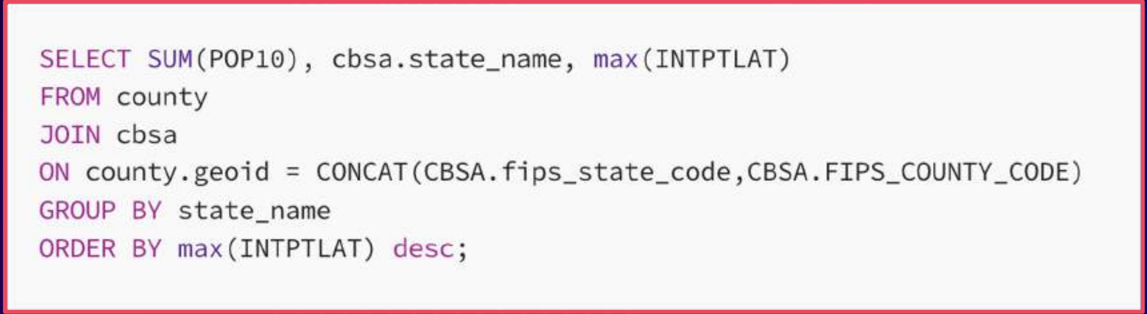
Query: Show me the total population of each state ordered from the most northern one to the most southern one.

Gold Standard



```sql
SELECT SUM(POP10), cbsa.state_name
FROM county
JOIN cbsa
ON county.geoid = CONCAT(CBSA.fips_state_code,CBSA.FIPS_COUNTY_CODE)
GROUP BY state_name
ORDER BY max(INTPTLAT) desc;
```

❌ Extra Column for Latitude



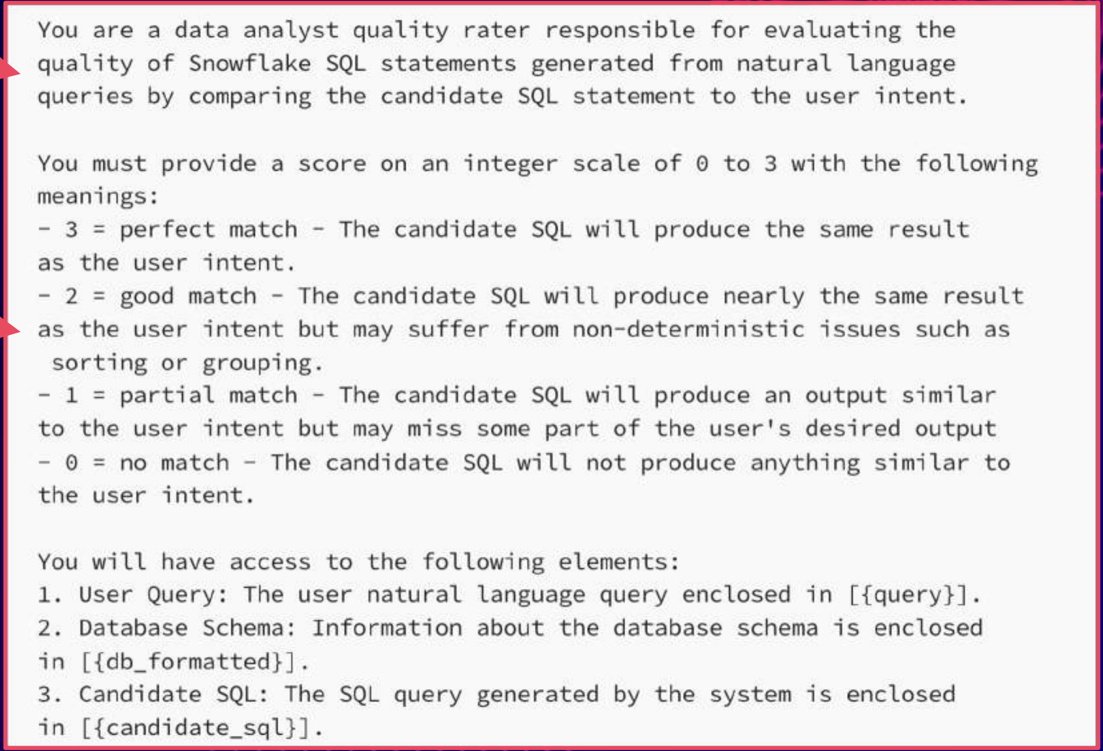
```sql
SELECT SUM(POP10), cbsa.state_name, max(INTPTLAT)
FROM county
JOIN cbsa
ON county.geoid = CONCAT(CBSA.fips_state_code,CBSA.FIPS_COUNTY_CODE)
GROUP BY state_name
ORDER BY max(INTPTLAT) desc;
```

@rajistics

https://medium.com/snowflake/inside-snowflake-building-the-most-powerful-sql-llm-in-the-world-1a33b3ee0d37

Model based evaluation - Text->SQL

- 📄 Define Data Quality
- 🎹 Grading Scale
- 📥 Explain Inputs



You are a data analyst quality rater responsible for evaluating the quality of Snowflake SQL statements generated from natural language queries by comparing the candidate SQL statement to the user intent.

You must provide a score on an integer scale of 0 to 3 with the following meanings:
- 3 = perfect match - The candidate SQL will produce the same result as the user intent.
- 2 = good match - The candidate SQL will produce nearly the same result as the user intent but may suffer from non-deterministic issues such as sorting or grouping.
- 1 = partial match - The candidate SQL will produce an output similar to the user intent but may miss some part of the user's desired output
- 0 = no match - The candidate SQL will not produce anything similar to the user intent.

You will have access to the following elements:
1. User Query: The user natural language query enclosed in [{query}].
2. Database Schema: Information about the database schema is enclosed in [{db_formatted}].
3. Candidate SQL: The SQL query generated by the system is enclosed in [{candidate_sql}].

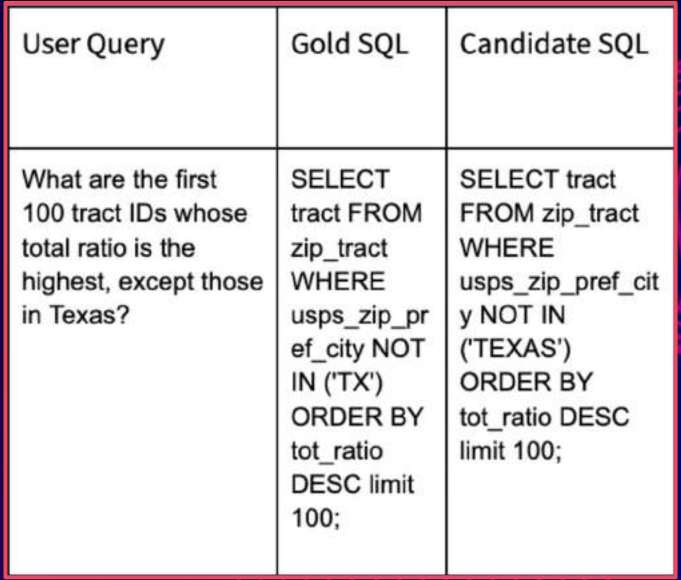
@rajistics

https://medium.com/snowflake/inside-snowflake-building-the-most-powerful-sql-llm-in-the-world-1a33b3ee0d37

Model based evaluation - Text->SQL

Returned SQL is a bit different:

TX versus TEXAS

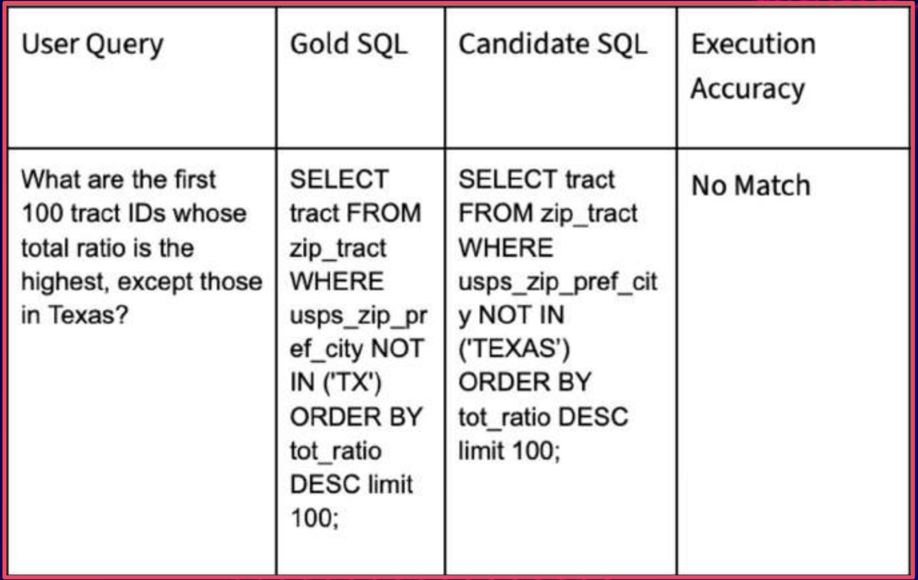


User Query	Gold SQL	Candidate SQL
What are the first 100 tract IDs whose total ratio is the highest, except those in Texas?	SELECT tract FROM zip_tract WHERE usps_zip_pref_city NOT IN ('TX') ORDER BY tot_ratio DESC limit 100;	SELECT tract FROM zip_tract WHERE usps_zip_pref_city NOT IN ('TEXAS') ORDER BY tot_ratio DESC limit 100;

https://medium.com/snowflake/inside-snowflake-building-the-most-powerful-sql-llm-in-the-world-1a33b3ee0d37

Model based evaluation - Text->SQL

Similarity:
❌ not a
100% match



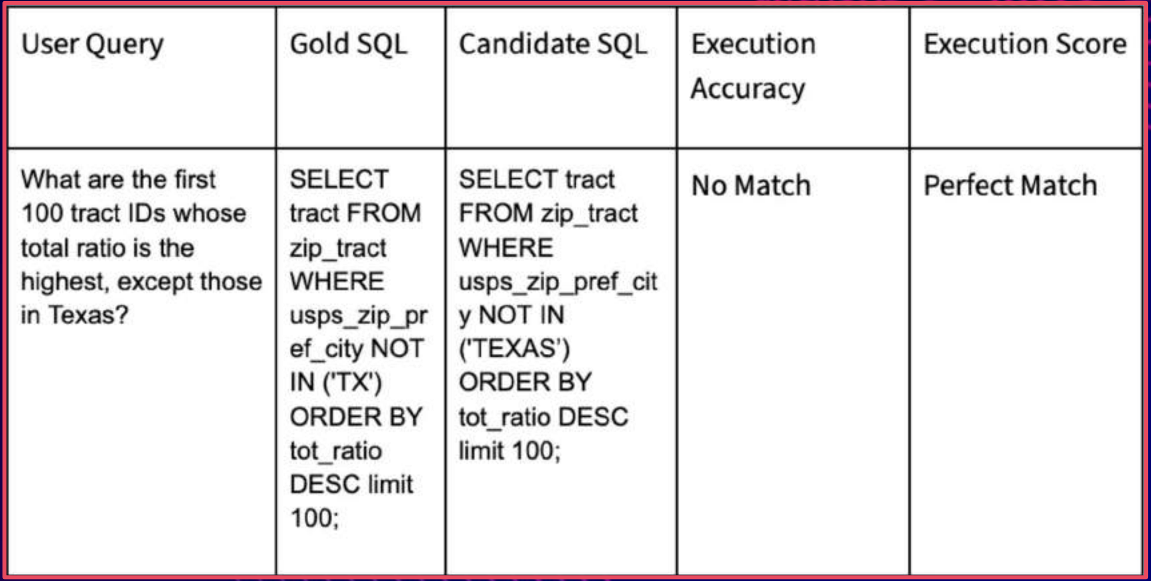
User Query	Gold SQL	Candidate SQL	Execution Accuracy
What are the first 100 tract IDs whose total ratio is the highest, except those in Texas?	SELECT tract FROM zip_tract WHERE usps_zip_pref_city NOT IN ('TX') ORDER BY tot_ratio DESC limit 100;	SELECT tract FROM zip_tract WHERE usps_zip_pref_city NOT IN ('TEXAS') ORDER BY tot_ratio DESC limit 100;	No Match

https://medium.com/snowflake/inside-snowflake-building-the-most-powerful-sql-llm-in-the-world-1a33b3ee0d37

Model based evaluation - Text->SQL

Model:

It’s ok, its captures the users intent



User Query	Gold SQL	Candidate SQL	Execution Accuracy	Execution Score
What are the first 100 tract IDs whose total ratio is the highest, except those in Texas?	SELECT tract FROM zip_tract WHERE usps_zip_pref_city NOT IN ('TX') ORDER BY tot_ratio DESC limit 100;	SELECT tract FROM zip_tract WHERE usps_zip_pref_city NOT IN ('TEXAS') ORDER BY tot_ratio DESC limit 100;	No Match	Perfect Match

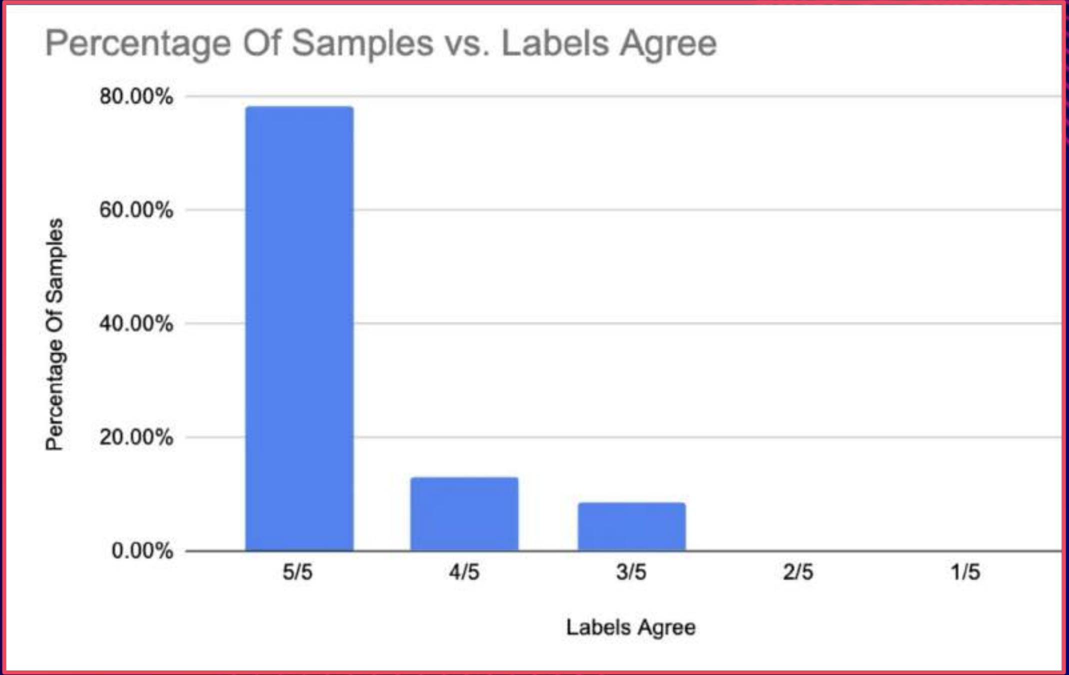
@rajistics

https://medium.com/snowflake/inside-snowflake-building-the-most-powerful-sql-llm-in-the-world-1a33b3ee0d37

Model based evaluation - Text->SQL

Strong correlation

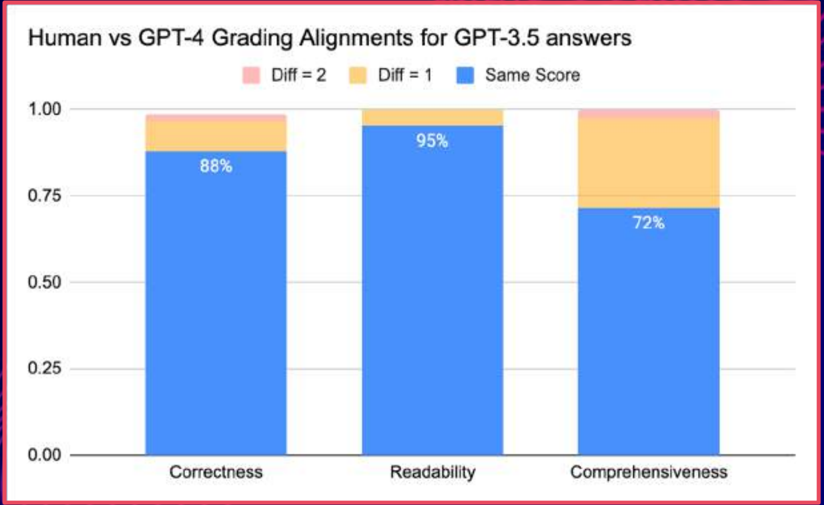
to other evaluation approaches



https://medium.com/snowflake/inside-snowflake-building-the-most-powerful-sql-llm-in-the-world-1a33b3ee0d37

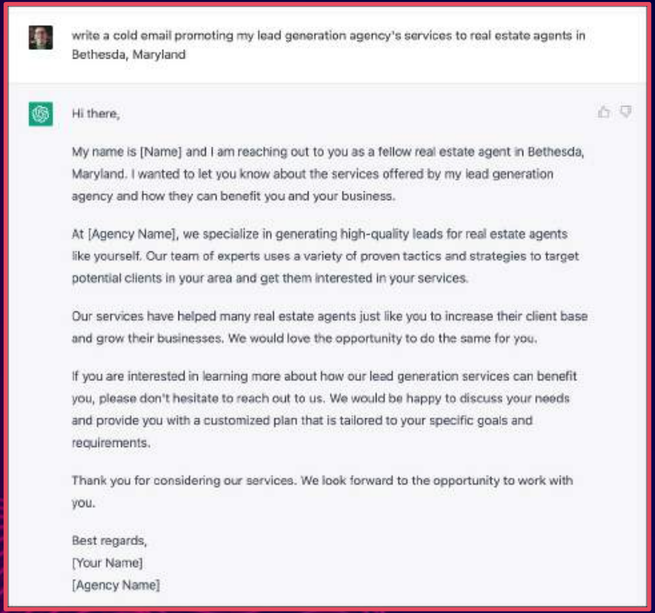
Model evaluation aligns with humans

Human and GPT-4 judges can reach above 80% agreement on the correctness and readability score.



https://arxiv.org/abs/2305.01937
https://www.databricks.com/blog/LLM-auto-eval-best-practices-RAG
https://arxiv.org/abs/2303.16634
https://arxiv.org/pdf/2306.05685.pdf

Unit Tests Beyond Code



write a cold email promoting my lead generation agency's services to real estate agents in Bethesda, Maryland

Hi there,

My name is [Name] and I am reaching out to you as a fellow real estate agent in Bethesda, Maryland. I wanted to let you know about the services offered by my lead generation agency and how they can benefit you and your business.

At [Agency Name], we specialize in generating high-quality leads for real estate agents like yourself. Our team of experts uses a variety of proven tactics and strategies to target potential clients in your area and get them interested in your services.

Our services have helped many real estate agents just like you to increase their client base and grow their businesses. We would love the opportunity to do the same for you.

If you are interested in learning more about how our lead generation services can benefit you, please don't hesitate to reach out to us. We would be happy to discuss your needs and provide you with a customized plan that is tailored to your specific goals and requirements.

Thank you for considering our services. We look forward to the opportunity to work with you.

Best regards,
[Your Name]
[Agency Name]

Properties of Emails?

- First/Last Name?
- Grammar/spelling
- Concise?
- Verify actions?
- Tone? - is it polite

+ *Explanations*

We can do this!

- Exact matching approach
- Similarity approach
- Unit Tests
- Evaluation Benchmarks
- Human Evaluation
- Human Comparison/Arena
- Model based Approaches
- Red Teaming

https://unsplash.com/photos/man-in-blue-denim-jacket-standing-on-road-during-daytime-qCJzjVHkERc

@rajistics

We can do this!

- Exact matching approach
- Similarity approach
- Unit Tests
- Evaluation Benchmarks
- Human Evaluation
- Human Comparison/Arena
- Model based Approaches
- Red Teaming

https://unsplash.com/photos/man-in-blue-denim-jacket-standing-on-road-during-daytime-qCJzjVHkERc

Evaluation Techniques for LLMs

https://github.com/rajshah4/LLM-Evaluation

Rajiv Shah
@rajistics
r.shah@snowflake.com

April 2024