On the Brittleness of Evaluation in NLP

Gabi Stanovsky

HUJI
NLP

NLP in the Wild

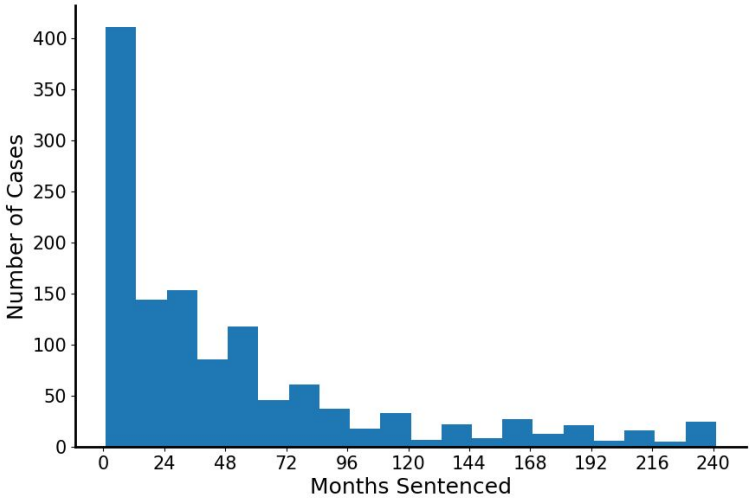
- We develop models models for **real-world settings**

NLP in the Wild

- We develop models models for **real-world settings**
  - **Historical texts:** (Lazar et al., EMNLP 2021)
    Multilingual LLMs can fill in gaps in Akkadian tablets

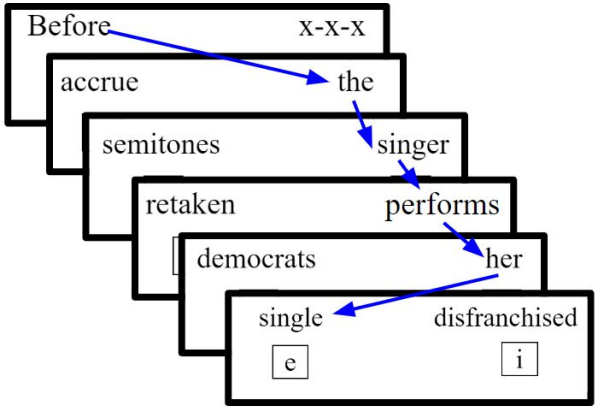
NLP in the Wild

- We develop models models for **real-world settings**
  - Historical texts, **legal analysis**: (Wegner et al., 2021; Habba et al., 2023)
    Examining sentencing patterns over 50 years of Israeli verdicts



NLP in the Wild

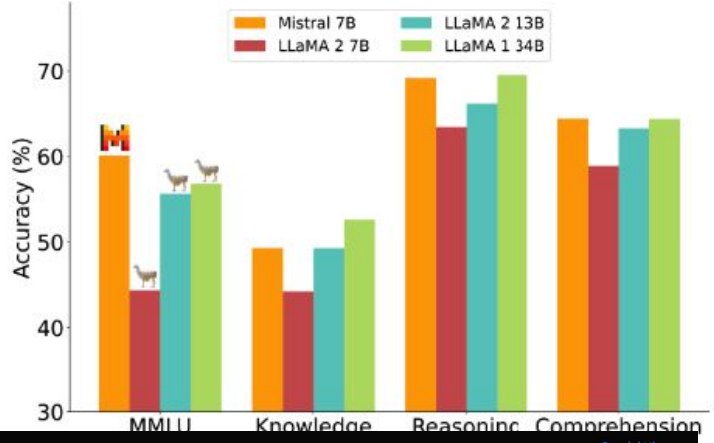
- We develop models models for **real-world settings**
  - Historical texts, legal analysis, **cognitive modelling**: (Lior et al., 2023)
    Human annotators exhibit biases comparable to models



Today

- **An explosion of LLMs**

Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4
TriviaQA	59.6	56.8	68.9	77.2			

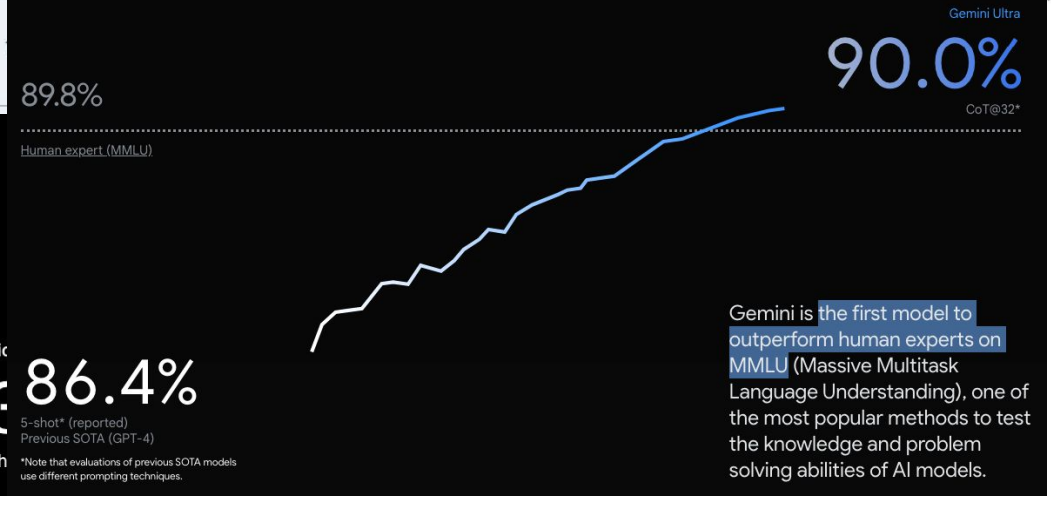


GPT-4 outperforms ChatGPT by scoring in higher approximate percentiles among test-takers.

Uniform Bar Exam

10th
ChatGPT

90th
GPT-4



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), …

Today

- Common practice in LLM evaluation: **Single-prompt benchmarks**
  - HELM, MMLU, BIG-Bench, etc.

- This leads to **very brittle evaluation results**
  - We evaluate 20 LLMs on more than 6.5M instances in 39 tasks
  - High variability in absolute performance and relative ranking

- **How should we evaluate LLMs?**
  - We propose a set of metrics for different downstream use-cases
  - A proposal for stochastic benchmarking of LLMs

Today

- **Common practice in LLM evaluation: Single-prompt benchmarks**
- Leads to **very brittle evaluation results**
- **How should we evaluate LLMs?**

Single-Prompt Evaluation

- Common practice: evaluate all models against a **single prompt template**

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

- 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

- 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

Few shot examples

Single-Prompt Evaluation

- 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*

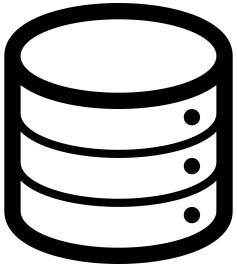
Word1: eight
Word2: mouth
Query word: ate
Output word:

Instruction

Few shot examples

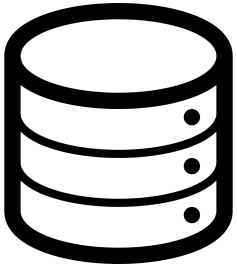
Input sample

Homophones test set **Prompt Construction** **Generation** **Evaluation**



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

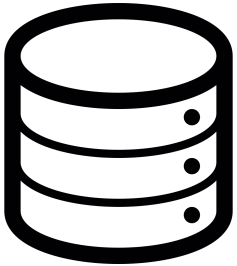
Homophones test set | **Prompt Construction** | **Generation** | **Evaluation**



Determine which of the two words is a homophone or sounds more like a query word.
Word1: eight
Word2: mouth
Query word: ate
Output word:

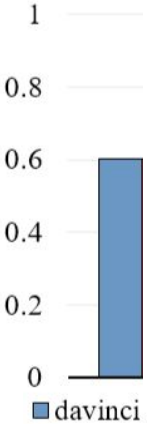
(ate, eight, mouth)
(flour, flower, oven)
…

Homophones test set **Prompt Construction** **Generation** **Evaluation**

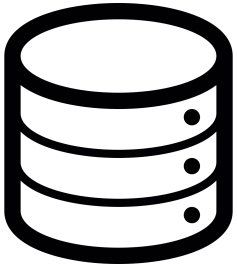


(ate, eight, mouth)
(flour, flower, oven)
…

Determine which of the two words is a homophone or sounds more like a query word.
Word1: oven
Word2: flower
Query word: flour
Output word:



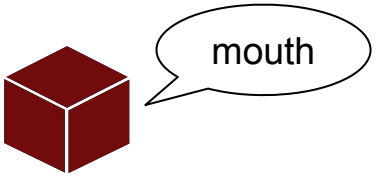
Homophones test set **Prompt Construction** **Generation** **Evaluation**

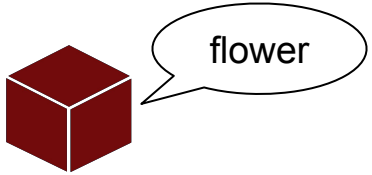


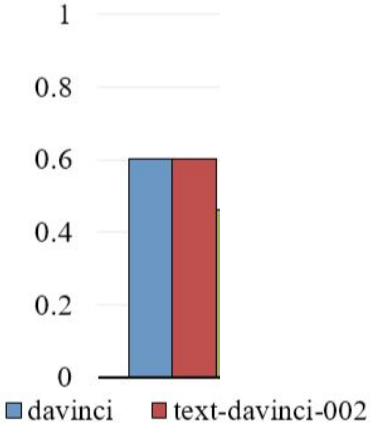
(ate, eight, mouth)
(flour, flower, oven)
…

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

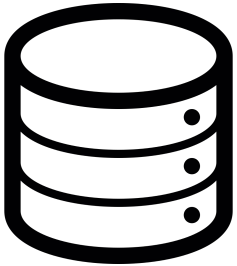
Determine which of the two words is a homophone or sounds more like a query word.
Word1: oven
Word2: flower
Query word: flour
Output word:







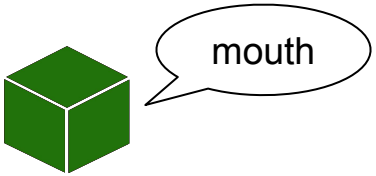
Homophones test set **Prompt Construction** **Generation** **Evaluation**

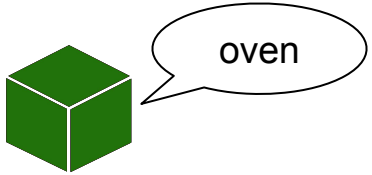


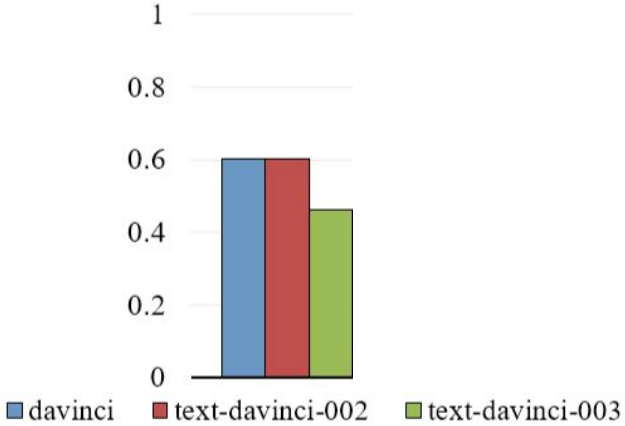
(ate, eight, mouth)
(flour, flower, oven)
…

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

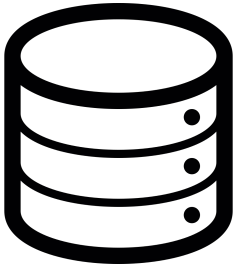
Determine which of the two words is a homophone or sounds more like a query word.
Word1: oven
Word2: flower
Query word: flour
Output word:







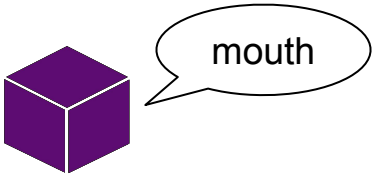
Homophones test set **Prompt Construction** **Generation** **Evaluation**

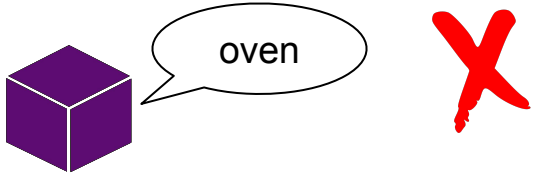


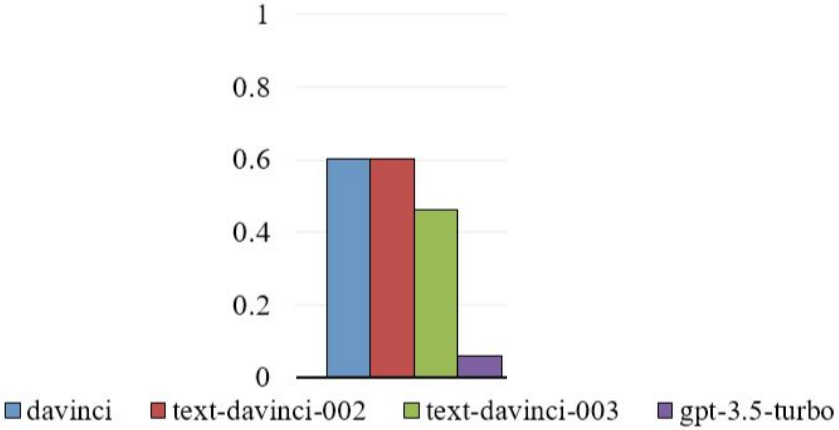
(ate, eight, mouth)
(flour, flower, oven)
…

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

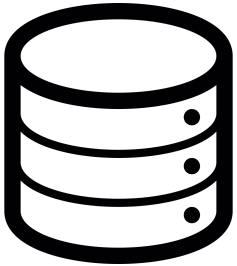
Determine which of the two words is a homophone or sounds more like a query word.
Word1: oven
Word2: flower
Query word: flour
Output word:







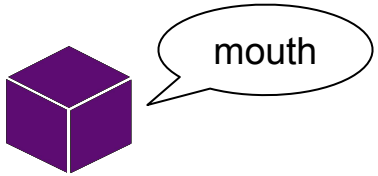
Homophones test set | **Prompt Construction** | **Generation** | **Evaluation**

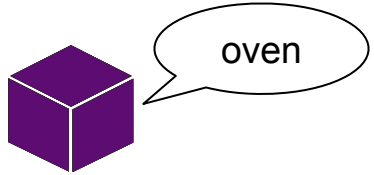


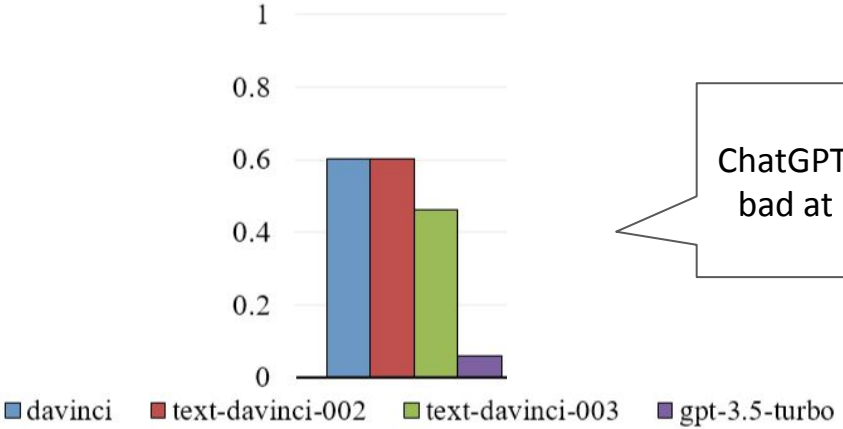
(ate, eight, mouth)
(flour, flower, oven)
…

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

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:







ChatGPT is surprisingly bad at homophones

Today

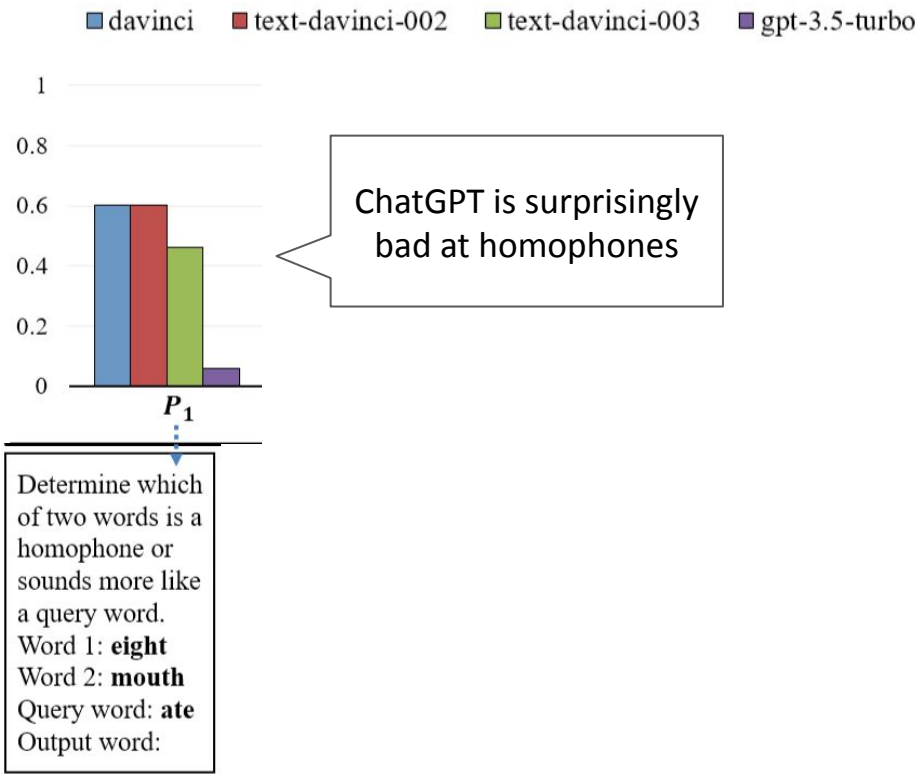
- Common practice in LLM evaluation: Single-prompt benchmarks

- Single prompt evaluation leads to **very brittle results**
  - We evaluate 20 LLMs on more than 6.5M instances in 39 tasks
  - Both in absolute performance, as well as in relative ranking

- How should we evaluate LLMs?

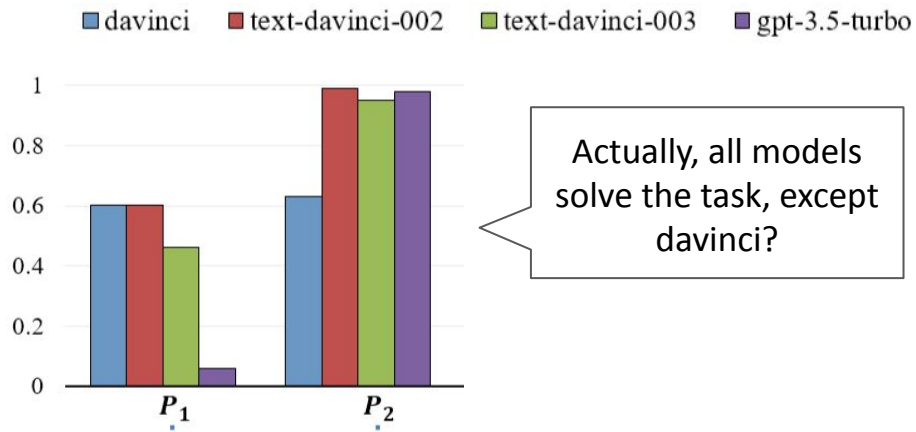
State of what art? Mizrahi et al., TACL 2024 (to be presented at ACL 2024)

There are many ways to phrase instructions



There are many ways to phrase instructions

ChatGPT is surprisingly bad at homophones



Actually, all models solve the task, except davinci?

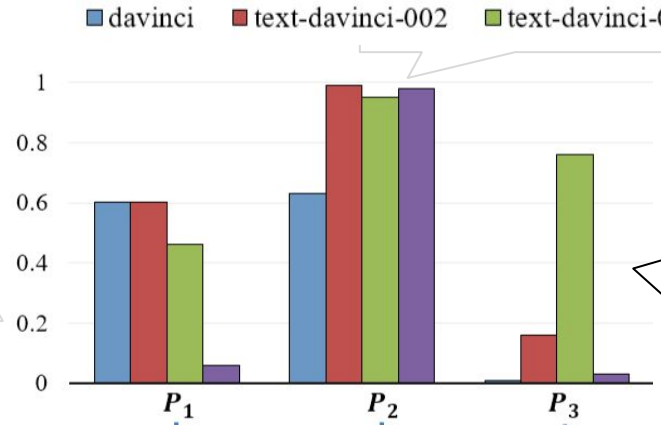
Determine which of two words is a homophone or sounds more like a query word.
Word 1: **eight**
Word 2: **mouth**
Query word: **ate**
Output word:

Given a query word and two other words, determine which of the two words is a homophone of the query word.
Query word: **ate**
Word 1: **eight**
Word 2: **mouth**
Output: The homophone of **ate** is

There are many ways to phrase instructions

Actually, all models solve the task, except

ChatGPT is surprisingly bad at homophones



Only Davinci-003 gets homophones??

Determine which of two words is a homophone or sounds more like a query word.
Word 1: **eight**
Word 2: **mouth**
Query word: **ate**
Output word:

Given a query word and two other words, determine which of the two words is a homophone of the query word.
Query word: **ate**
Word 1: **eight**
Word 2: **mouth**
Output: The homophone of **ate** is

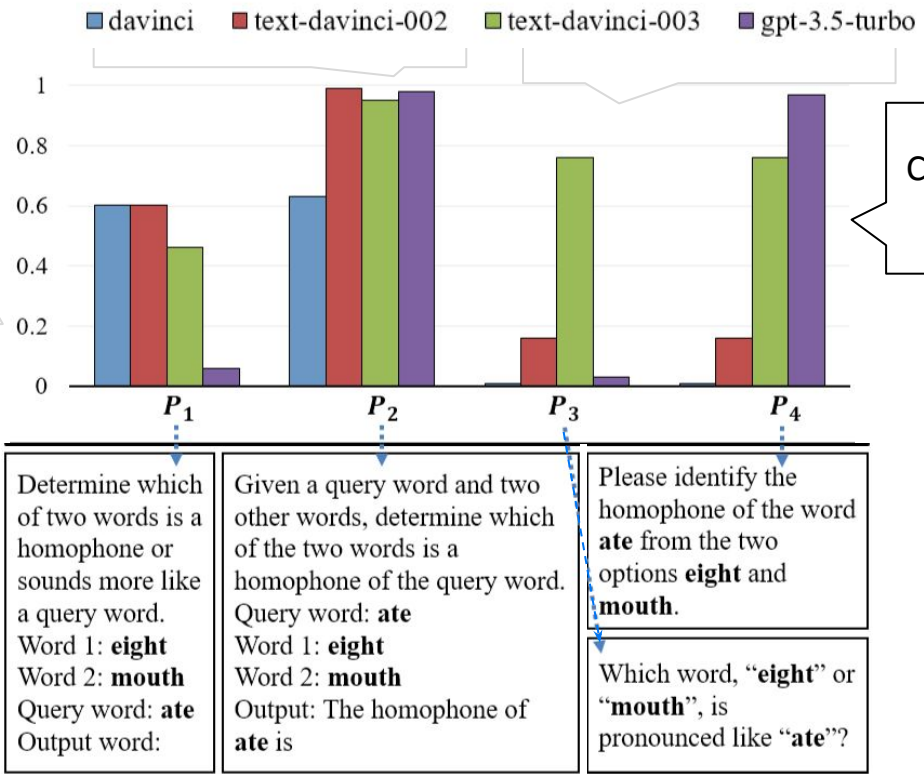
Which word, "**eight**" or "**mouth**", is pronounced like "**ate**"?

There are many ways to phrase instructions

Actually, all models solve the task, except

Only Davinci-003 gets

ChatGPT is surprisingly bad at homophones



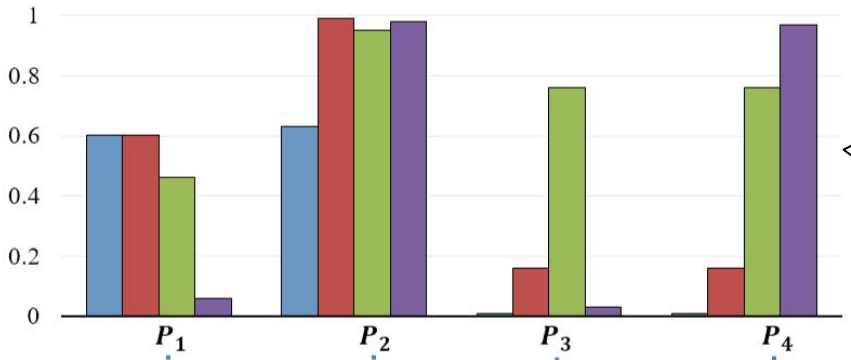
ChatGPT is the best at homophones?!

Actually, all models solve the task, except davinci?

Only Davinci-003 gets homophones?

ChatGPT is surprisingly bad at homophones?

ChatGPT is the best at homophones?



Determine which of two words is a homophone or sounds more like a query word.
Word 1: **eight**
Word 2: **mouth**
Query word: **ate**
Output word:

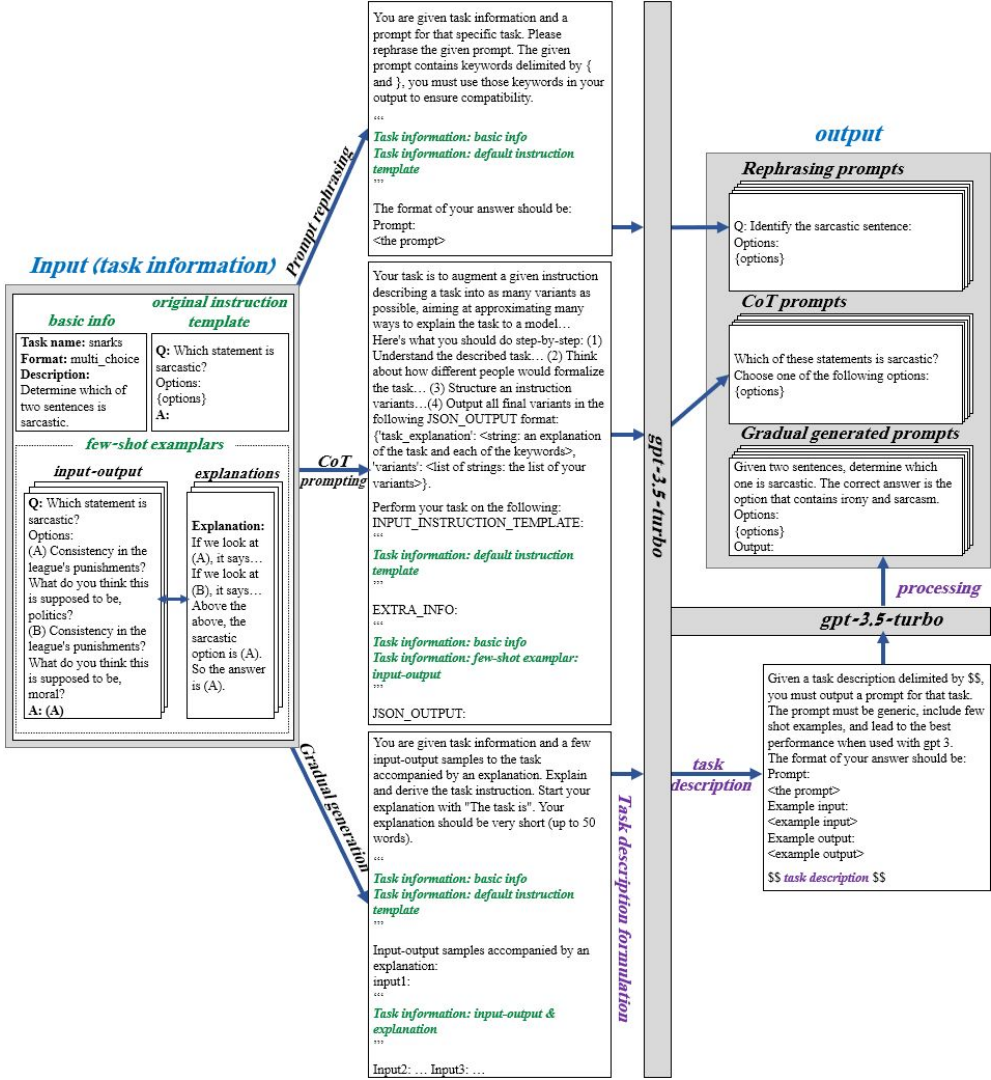
Given a query word and two other words, determine which of the two words is a homophone of the query word.
Query word: **ate**
Word 1: **eight**
Word 2: **mouth**
Output: The homophone of **ate** is

Please identify the homophone of the word **ate** from the two options **eight** and **mouth**.

Which word, "**eight**" or "**mouth**", is pronounced like "**ate**"?

Prompt Paraphrase Corpus

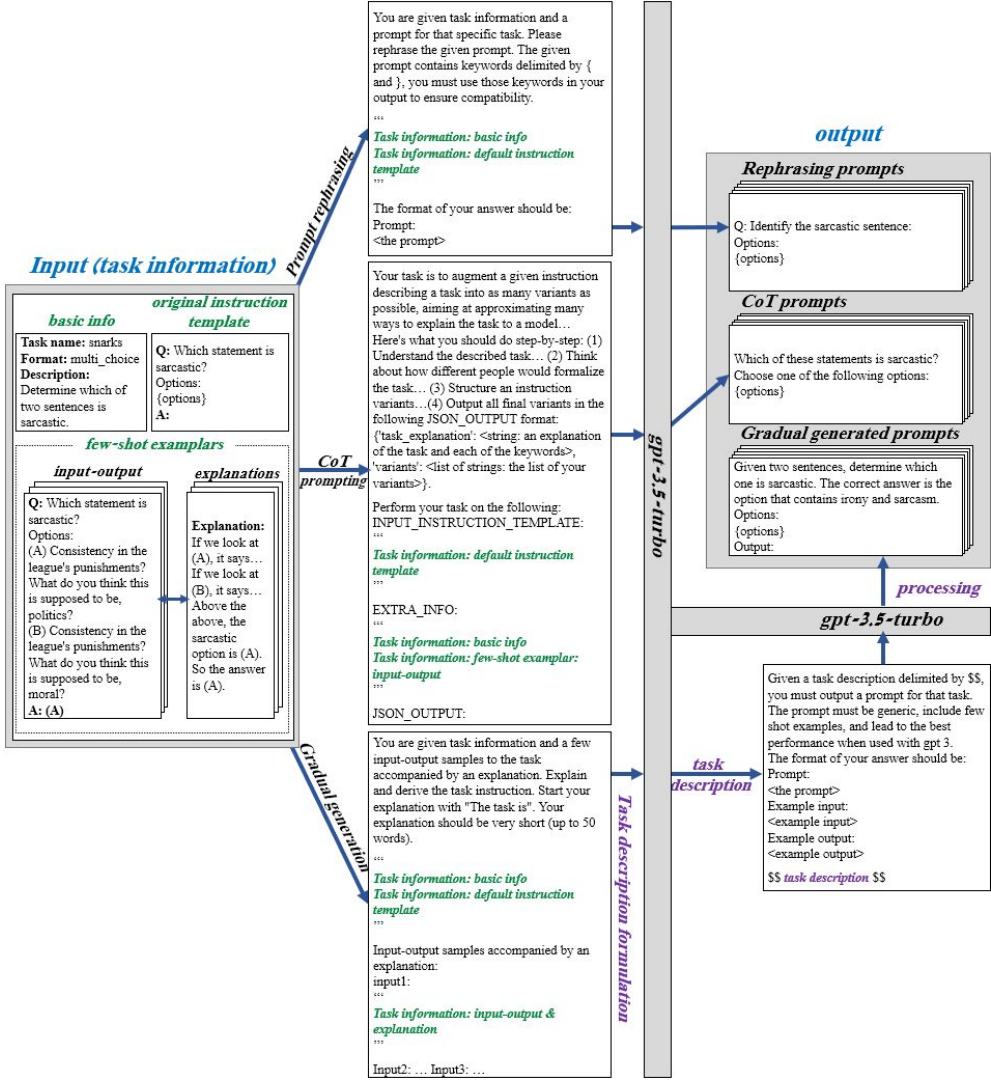
- **Three methods for paraphrasing**
  - Starting with the original prompt in LMEntry & BIG-bench Hard



Prompt Paraphrase Corpus

- **Three methods for paraphrasing**
  - Starting with the original prompt in LMEntry & BIG-bench Hard

- In addition, **manually written** paraphrases for BIG-bench Lite from Sun et al., 2023

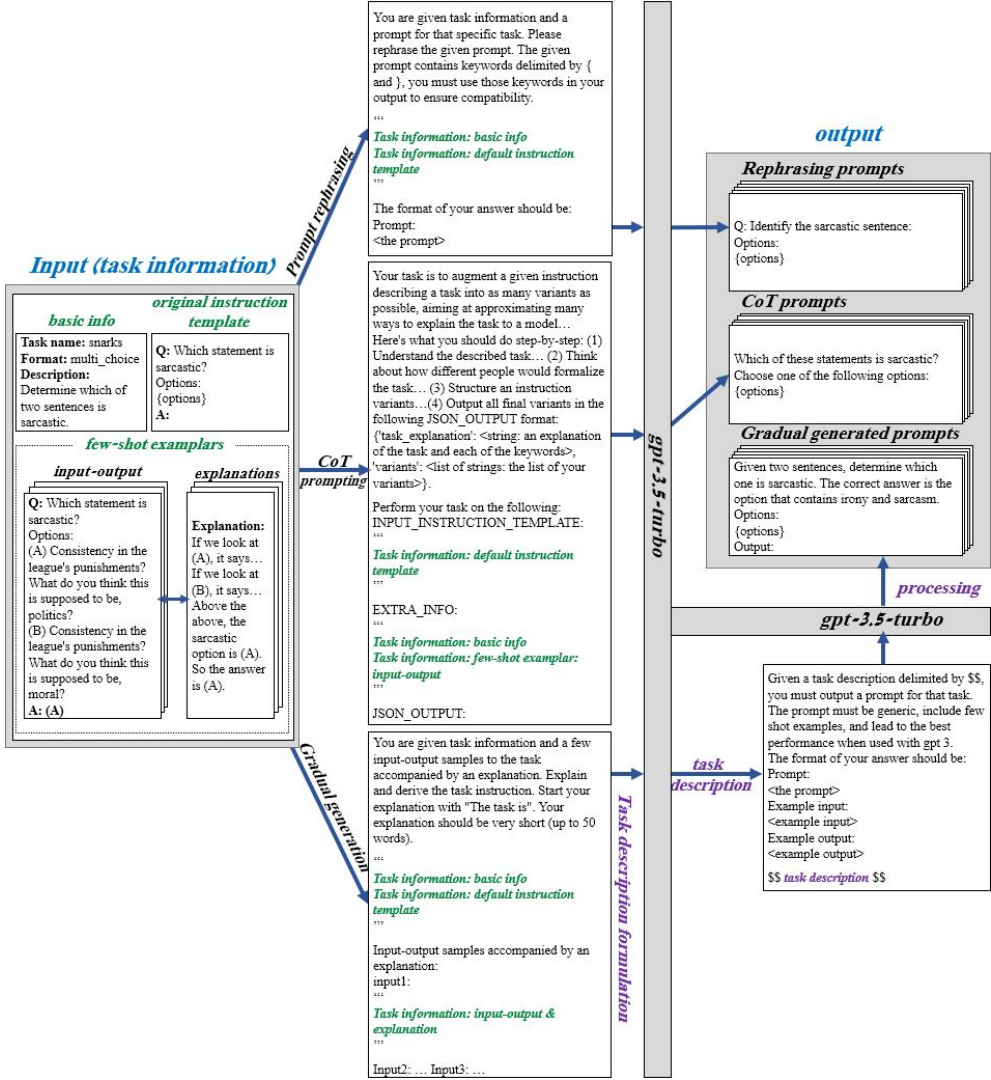


Prompt Paraphrase Corpus

- **Three methods for paraphrasing**
  - Starting with the original prompt in LMEntry & BIG-bench Hard

- In addition, **manually written** paraphrases for BIG-bench Lite from Sun et al., 2023

- All paraphrases were **manually verified and filtered**



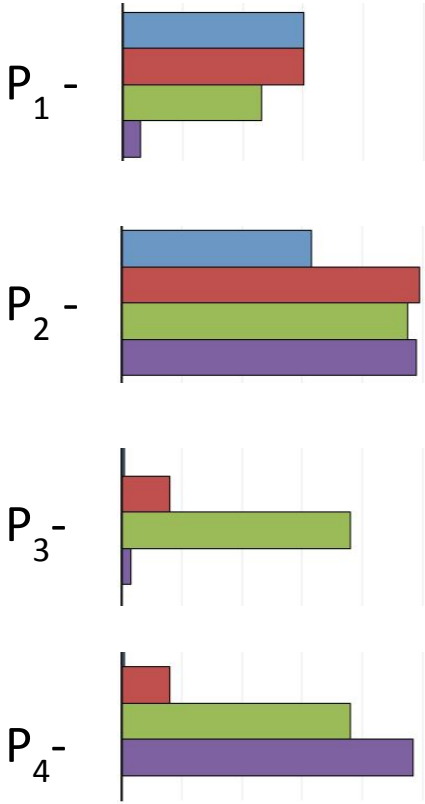
Prompt Paraphrase Corpus

Benchmark	Method	#Automatic Paraphrases	#Correct Paraphrases	Correct Ratio
LMENTRY	All	2429	2186	90.00%
	Rephrase	461	408	88.50%
	CoT	1286	1234	95.96%
	Gradual	652	514	78.83%
BBH	All	2615	2209	84.47%
	Rephrase	734	627	85.42%
	CoT	775	630	81.29%
	Gradual	1091	937	85.88%

- Resulting on average in **175 prompt paraphrases per task**
  - 4.3K paraphrases in total

Quantifying variations due to paraphrasing

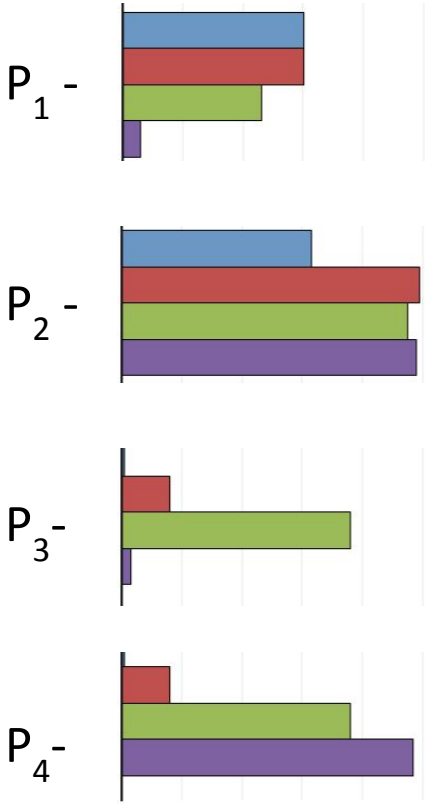
- Each **prompt induces a ranking** over *m* models



Quantifying variations due to paraphrasing

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

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



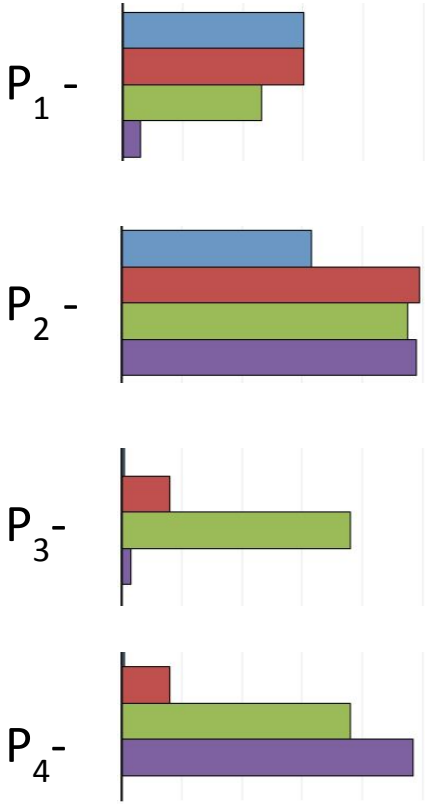
Quantifying variations due to paraphrasing

- 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

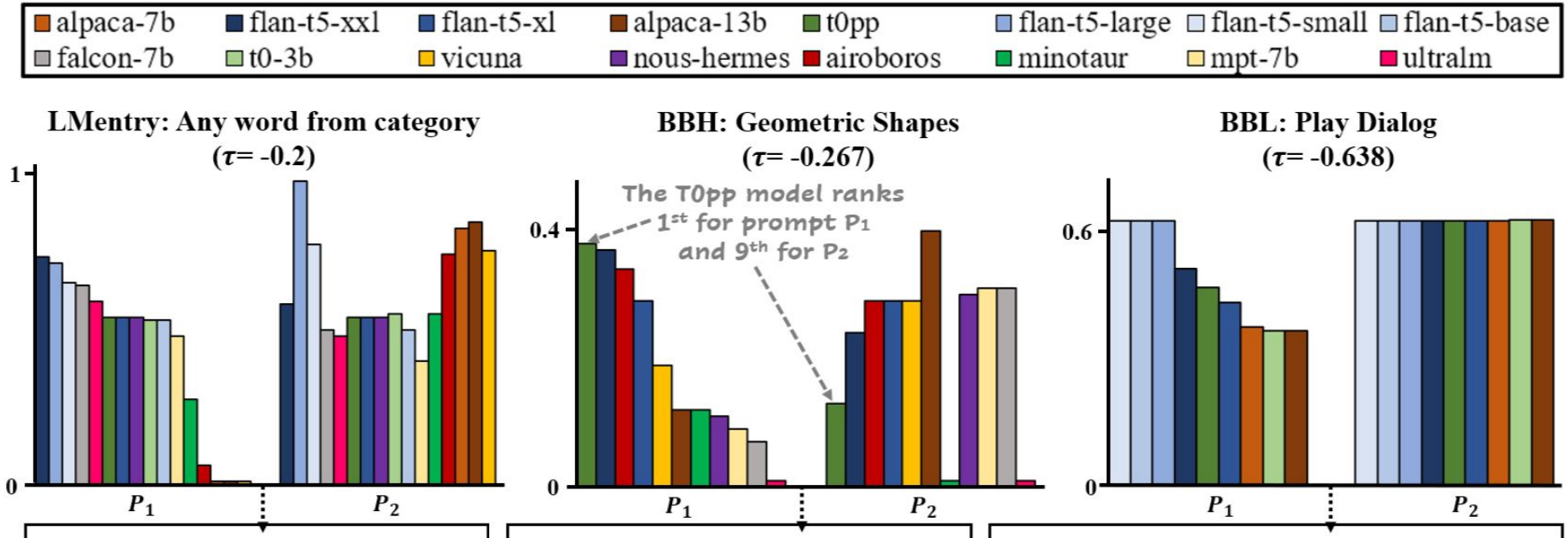
$$W = \frac{12 \sum_{i=1}^{n} (R_i - \bar{R})^2}{m^2 (n^3 - n)}$$



Prompt Phrasing Impacts Results in Most Tasks

- Kendall's *W* is in [0,1]
  - W = 1 is perfect agreement between all prompts
- For most tasks, only weak to moderate correlation
- **Phrasing of prompt greatly affects relative rankings**
- Also affects **absolute performance**
  - See paper for details

Tasks	Kendall's W	Friedman p-val
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**
BIG-bench Lite		
known unknown	.316 (weak)	**4.4E-5**
play dialog	.355 (weak)	**4.3E-5**
winowhy	.520 (weak)	**6.0E-4**
strategic qa	.529 (weak)	.191
hindu knowledge	.560 (weak)	.569
conceptual	.731 (medium)	.132
strange stories	.731 (medium)	.431
code desc	.756 (medium)	**.002**
novel concepts	.787 (medium)	.620
logic grid puzzle	.796 (medium)	**.010**
lang. identification	.811 (medium)	**.002**
vitaminc	.888 (strong)	.772
bbq lite	.890 (strong)	**.023**
logical deduction	.913 (strong)	.895



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:

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}*.

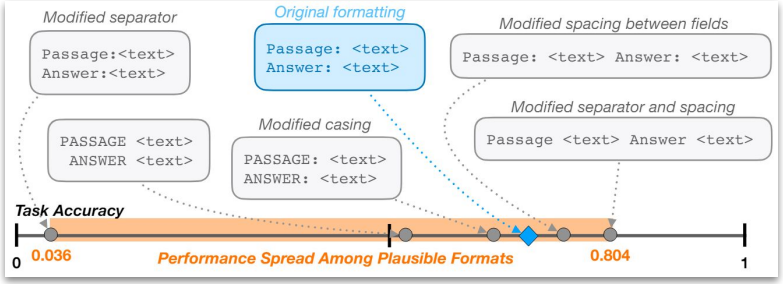
P_1: Now you are a dramatist… Here is the play: *{play}*. Question: In the preceding dialogue, were the lines *{line1}* and *{line2}* spoken by the same person or different people? Please just give a short answer: same or different. Your Answer:

P_2: 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:

Change	Model	P1	Acc.	P2	Acc.	Diff.
{...} –> "{...}"	td002	Which word has a greater number of letters, ***{word1}*** or ***{word2}***?	.50	Which word has a greater number of letters, ***"{word1}"*** or ***"{word2}"***?	.23	-0.27
	td002	Which of the words ***{word1}*** and ***{word2}*** is alphabetically first?	.54	Which of the words ***"{word1}"*** and ***"{word2}"*** is alphabetically first?	.77	+0.23
	td003	Which word has a greater number of letters, ***{word1}*** or ***{word2}***?	.60	Which word has a greater number of letters, ***"{word1}"*** or ***"{word2}"***?	.14	-0.46
	td003	Compare the length of ***{word1}*** and ***{word2}*** and tell me which one is shorter.	.39	Compare the length of ***"{word1}"*** and ***"{word2}"*** and tell me which one is shorter.	.73	+0.34
	cgpt	Which word has a greater number of letters, ***{word1}*** or ***{word2}***?	.55	Which word has a greater number of letters, ***"{word1}"*** or ***"{word2}"***?	.24	-0.31
	cgpt	Compare the length of ***{word1}*** and ***{word2}***. Which one is longer?	.04	Compare the length of ***"{word1}"*** and ***"{word2}"***. Which one is longer?	.70	+0.66
',' –> ':'	td002	Which word is a rhyme for "{query}", "{word1}" or "{word2}"?	.08	Which word is a rhyme for "{query}**:** "{word1}" or "{word2}"?	.85	+0.77
	td003	Which word is a rhyme for "{query}", "{word1}" or "{word2}"?	.48	Which word is a rhyme for "{query}**:** "{word1}" or "{word2}"?	.90	+0.42
',' –> '-'	td002	Which word rhymes with "{query}", "{word1}" or "{word2}"?	.06	Which word rhymes with "{query}" **-** "{word1}" or "{word2}"?	.73	+0.67
	td003	Which word rhymes with "{query}", "{word1}" or "{word2}"?	.17	Which word rhymes with "{query}" **-** "{word1}" or "{word2}"?	.60	+0.43
the –> a	td002	What is ***the*** word that rhymes with "{query}" - "{word1}" or "{word2}"?	.03	What is ***a*** word that rhymes with "{query}" - "{word1}" or "{word2}"?	.78	+0.75
which –> what	td002	***Which*** word rhymes with "{query}" - "{word1}" or "{word2}"?	.73	***What*** word rhymes with "{query}" - "{word1}" or "{word2}"?	.82	+0.09
	td003	***Which*** word rhymes with "{query}" - "{word1}" or "{word2}"?	.60	***What*** word rhymes with "{query}" - "{word1}" or "{word2}"?	.15	-0.45
word –> term	td002	Create a ***word*** that excludes the letter "{letter}".	.54	Create a ***term*** that excludes the letter "{letter}".	.04	-0.50
	td003	Create a ***word*** that excludes the letter "{letter}".	.96	Create a ***term*** that excludes the letter "{letter}".	.58	-0.38
	cgpt	Create a ***word*** that excludes the letter "{letter}".	.81	Create a ***term*** that excludes the letter "{letter}".	.42	-0.39

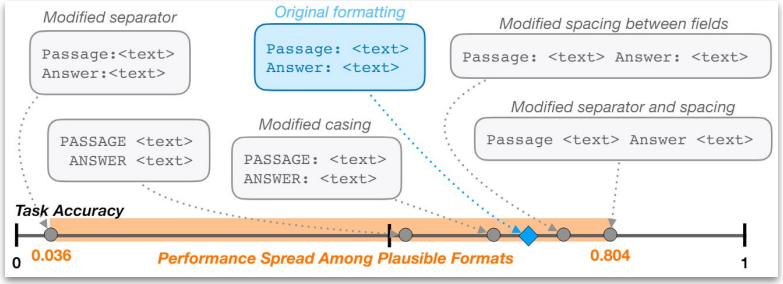
Recent Work – LLMs are sensitive to…

- Sclar et al (2023),(Alzahrani et al., 2024): …minor prompt paraphrases

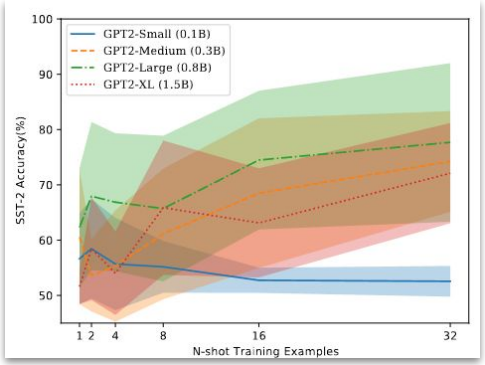


Recent Work – LLMs are also sensitive to…

- Sclar et al (2023),(Alzahrani et al., 2024): …minor prompt paraphrases



- Lu et al, (2022): …order of in-context examples



Today

- **Common practice in LLM evaluation: Single-prompt benchmarks**

- Single prompt evaluation leads to **very brittle results**
  - We evaluate 20 LLMs on more than 6.5M instances in 39 tasks
  - Both in absolute performance, as well as in relative ranking

- **How should we evaluate LLMs?**
  - We propose a set of metrics for different downstream use-cases
  - A proposal for stochastic benchmarking of LLMs

How should we evaluate LLMs?

- LLMs are evaluated for **diverse purposes**
  - LLM developers may care about mean performance and std
  - Specialized downstream applications may care about maximum performance

- **Why should there be a single way to evaluate LLMs?**

- Instead, we propose **three multi-prompt metrics**
  - Geared for specific use-cases

Maximum Metric – For Downstream Applications

LLM

prompt paraphrases for *T*

$$MaxP(M, T, I_T) = \max_{i \in I_T} \varepsilon(M, T, i)$$

task

standard evaluation for *T*
(e.g., accuracy, F1, EM)

- Optimizes for the **model performance on the task**
- Potentially, **different prompts per model**

Average Metric – for LLM Developers

LLM

prompt paraphrases for *T*

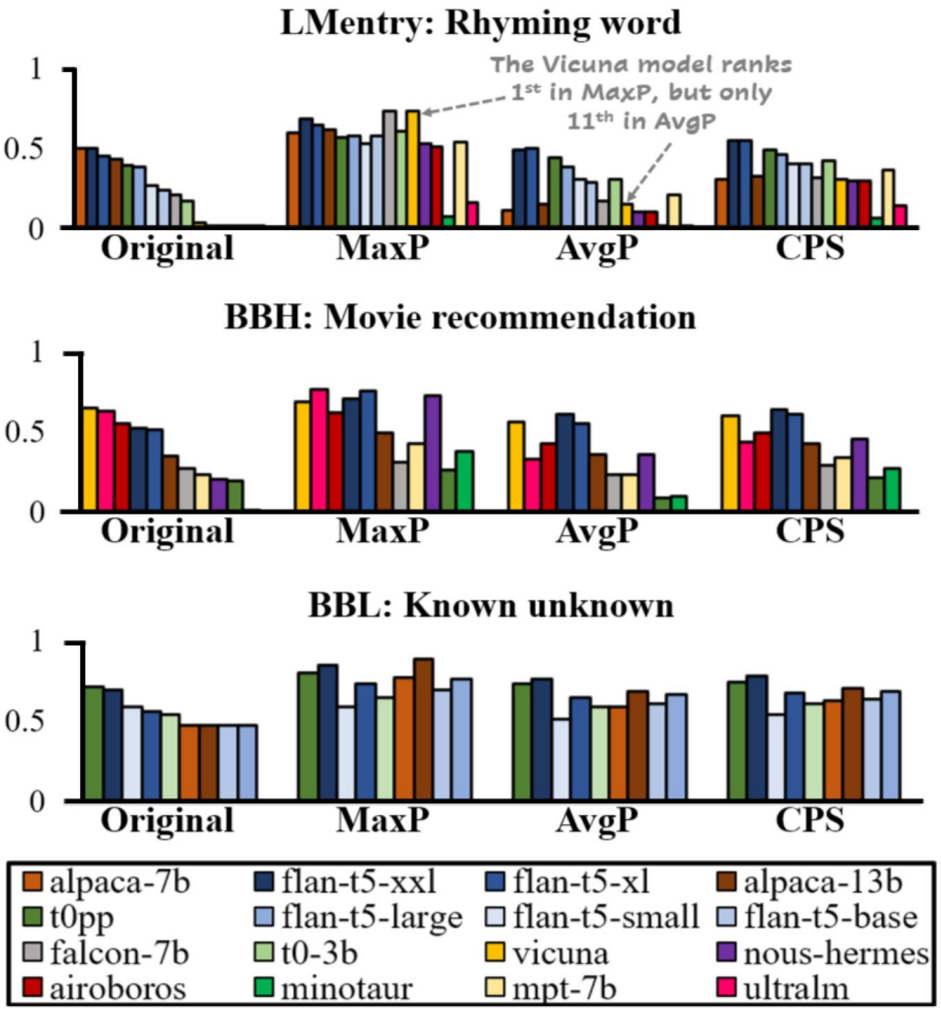
$$AvgP(M, T, I_T) = \frac{1}{|I_T|} \cdot \sum_{i \in I_T} \varepsilon(M, T, i)$$

task

standard evaluation for *T*
(e.g., accuracy, F1, EM)

- Measures LLM **robustness across prompt paraphrases**

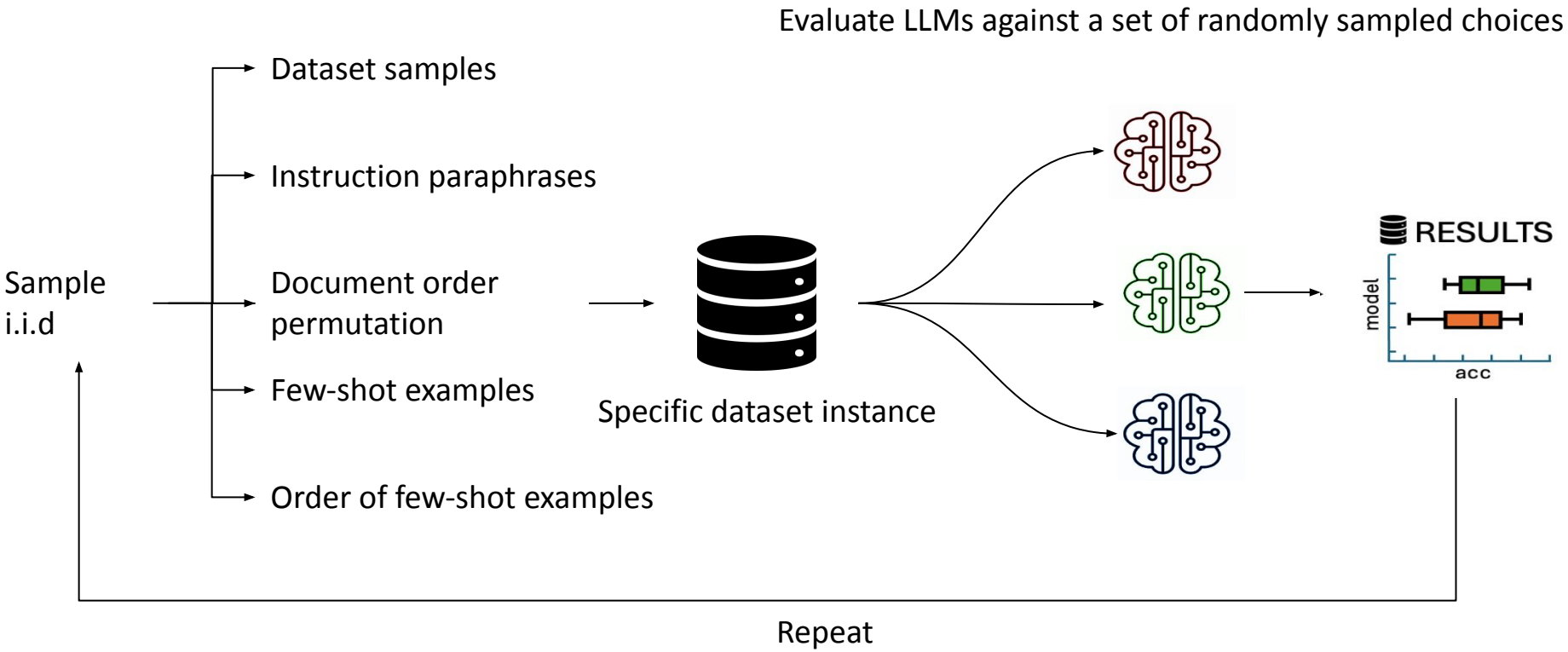
- Different models excel in different metrics
- All metrics differ from ranking on the original prompt
- LLaMA-based models competitive on *MaxP* but fall behind on *AvgP*



A Stochastic Approach to Benchmarking

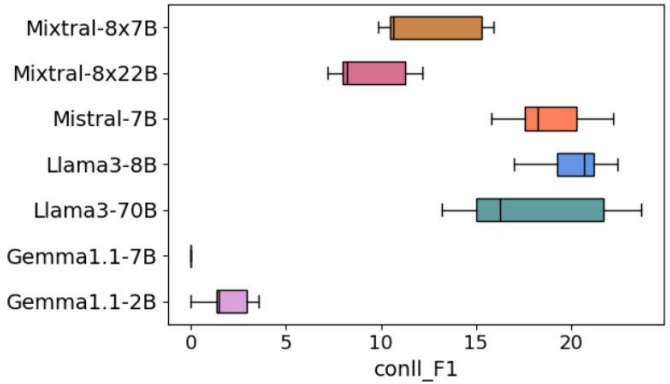
- Evaluation should accurately reflect performance of **current models**
  - E.g., BLEU was a good metric for the models of its time, but it no longer is
- Current models are very sensitive to **seemingly arbitrary choices**
  - Current evaluation should take this into account
- We propose to repeatedly **sample at random** each of 5 arbitrary choices
  - Instruction phrasing, samples, order of docs, few-shot examples & their order

A Stochastic Approach to Benchmarking



SEAM: A Stochastic Benchmark for Multi-Document Tasks

- Models plotted as a **distribution of scores**
- Some datasets and models are **more sensitive** than others
- Model size is **not correlated** with either
  - Higher mean performance
  - Nor more robust performance



(f) ECB+

Conclusion

Conclusion

- **LLMs present a unique novel challenge to evaluation**
  - Task definition is a part of the input, which can be paraphrased in many ways
- **Current evaluations are very brittle**
  - Seemingly arbitrary choices lead to very different observations
- **Proposal for new evaluation paradigms**
  - Different metrics for different use-cases, stochastic approach to benchmarking

Thanks for listening!

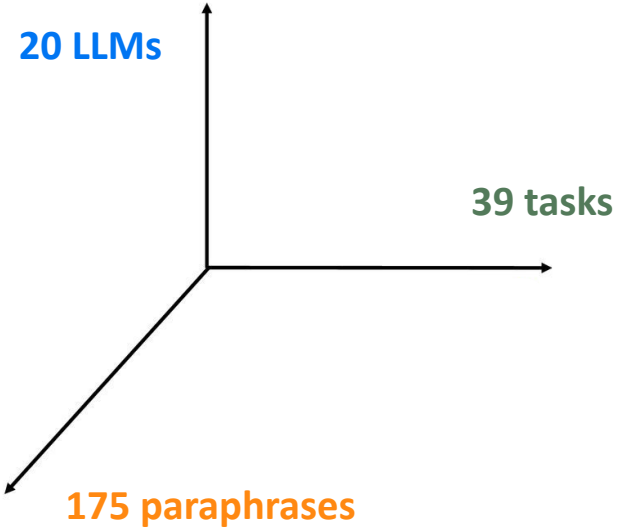
Large-Scale Evaluation: Models

- **No API-based models** in main evaluation
  - Interfere with prompt

- Similar observations on OpenAI models in small-scale evaluation

Model	Size	Base Model	# Params
Flan-T5	Small	T5	80M
	Base		250M
	Large		780M
	XL		3B
	XXL		11B
T0	Small	T5	3B
	T0pp		11B
Alpaca	Small	LLaMA	7B
	Big		13B
Vicuna		LLaMA	13B
Airoboros		LLaMA	13B
UltraLM		LLaMA	13B
Nous-Hermes		LLaMA	13B
Falcon-Instruct		Falcon	7B
MPT		MPT	7B
Minotaur		StarCoder Plus	15B

Large-Scale Evaluation



- We sample 100 instances per task, resulting in **over 6.5M instances**

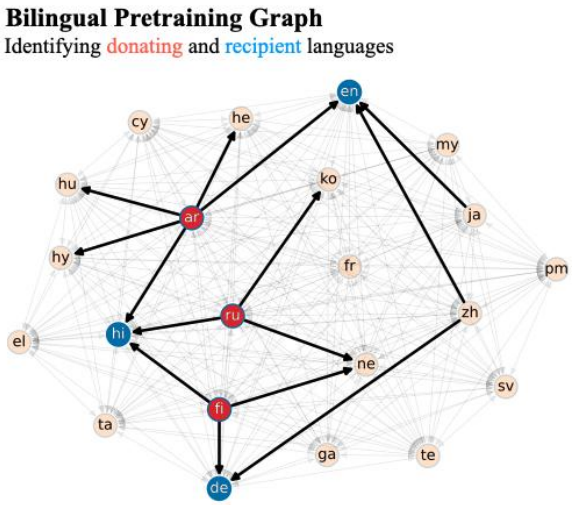
NLP in the Wild

- We develop models models for **real-world settings**
  - Historical texts, legal analysis, **cognitive modelling**: (Itzhak et al., 2024)
    Models exhibit cognitive biases

Bias	Control	Treatment
Decoy	Below you will find three phone brands. Which one would you choose? Brand 1 - price is $130, quality rating is 40. Brand 2 - price is $350, quality rating is 60. Answer: `Brand 1.`	Below you will find three phone brands. Which one would you choose? Brand 1 - price is $130, quality rating is 40. Brand 2 - price is $350, quality rating is 60. Brand 3 - price is $438, quality rating is 60. Answer: `Brand 2.`

NLP in the Wild

- We develop models models for **real-world settings**
- Lead to **broad research questions** beyond the specific task
  - Cross-lingual transfer in controlled setups (Malkin et al., 2022)



Combined Performance Score

$$Sat(M, T, I_T) = 1 - (MaxP - AvgP)$$

Punishes models where max deviates from average

$$CPS(M, T, I_T) = Sat \cdot MaxP$$

… while rewarding for high maximum

- For LLMs applied for a **diverse suite of applications**

There are many ways to phrase instructions

- **How does *prompt wording* affect performance?**