Session #8: Limits of In-Context Learning

Thursday, Sept 22

CSCI 601.771: Self-supervised Statistical Models

JOHNS HOPKINS UNIVERSITY

News: Whisper

OpenAI

- *"The Whisper models are trained for speech recognition and translation tasks, capable of transcribing speech audio into the text in the language it is spoken (ASR) as well as translated into English (speech translation)."*

https://openai.com/blog/whisper

Size	Parameters	English-only model	Multilingual model
tiny	39 M	✓	✓
base	74 M	✓	✓
small	244 M	✓	✓
medium	769 M	✓	✓
large	1550 M		✓

This Week's prompt

This paper is good at ___ but fails to address ___

Impact of Pretraining Term Frequencies on Few-Shot Reasoning

The flow of the presentation

-Background / Motivation

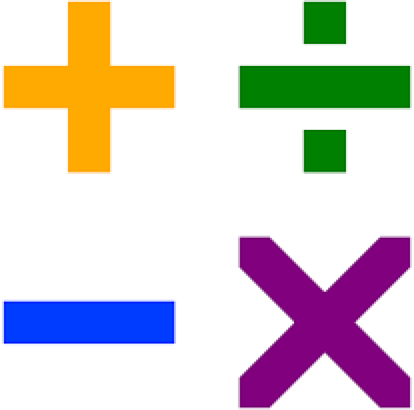
-Method

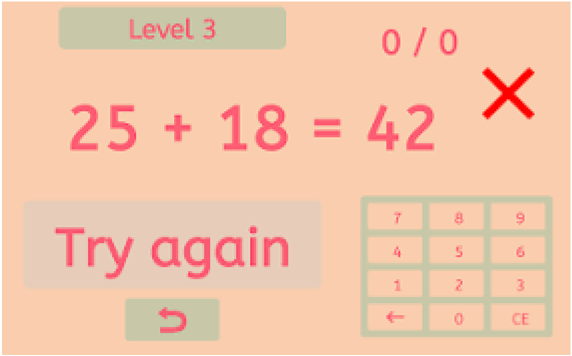
-Experiments

-Conclusions

: Vicky Zeng, Neha Verma, Lingfeng Shen

Numerical reasoning





: Vicky Zeng, Neha Verma, Lingfeng Shen

Background

Background

current evaluation schemes for the reasoning of large language models, often neglect or underestimate the impact of **data leakage**

Background

current evaluation schemes for the reasoning of large language models, often neglect or underestimate the impact of **data leakage**

A model that has learned to reason in the training phase should be able to generalize outside of the narrow context that it was trained in.

Background

current evaluation schemes for the reasoning of large language models, often neglect or underestimate the impact of **data leakage**

A model that has learned to reason in the training phase should be able to generalize outside of the narrow context that it was trained in.

GPT3

Frequent data

Performance

Less frequent data

Performance

: Vicky Zeng, Neha Verma, Lingfeng Shen

Related Work

Sinha et al. (2021) demonstrate that shuffling word order during pretraining has minimal impact on an LMs' accuracy on downstream tasks

Min et al. (2022) similarly find that shuffling labels in in-context learning demonstrations has a minimal impact on few-shot accuracy.

Data privacy researchers have also shown that LMs may memorize sensitive sequences occurring in training data (e.g., social security and credit card numbers), even if they are rare (Carlini et al., 2019; Song & Shmatikov, 2019).

A little question

Q: What is 63 times 24?

Q: What is 24 times 18?

Q: What is 33 times 12?

Q: What is 41 times 19?

A little question

Q: What is 63 times 24?

Q: What is 24 times 18?

Accuracy: >80%

Q: What is 33 times 12?

Q: What is 41 times 19?

A little question

Q: What is 63 times 24?

Q: What is 24 times 18?

Accuracy: >90%

Q: What is 33 times 12?

Q: What is 41 times 19?

Q: What is 61 times 24?

Q: What is 23 times 18?

Q: What is 31 times 17?

Q: What is 47 times 17?

: Vicky Zeng, Neha Verma, Lingfeng Shen

A little question

Q: What is 63 times 24?

Q: What is 24 times 18?

Accuracy: >90%

Q: What is 33 times 12?

Q: What is 41 times 19?

Q: What is 61 times 24?

Q: What is 23 times 18?

Q: What is 31 times 17?

Accuracy: <40%

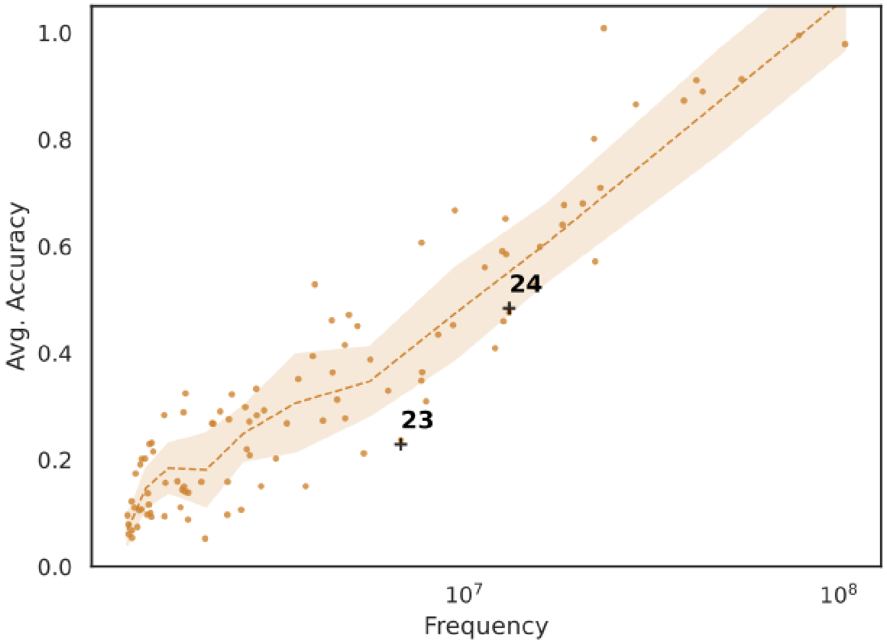
Q: What is 47 times 17?

: Vicky Zeng, Neha Verma, Lingfeng Shen

Problems

Q: What is 24 times 18? A: ____ *Model:* 432 ✓

Q: What is 23 times 18? A: ____ *Model:* 462 ✗



: Vicky Zeng, Neha Verma, Lingfeng Shen

Instances

1. x = positive integers, units of time
2. y = positive integers (optional)
3. ω = frequency, co-occurrences within window=5

$$X^{(1)} = (x_1) \qquad \omega_{X^{(1)}}$$

$$X^{(2)} = (x_1, x_2) \qquad \omega_{X^{(2)}}$$

$$X^{(3)} = (x_1, y) \qquad \omega_{X^{(3)}}$$

$$X^{(4)} = (x_1, x_2, x_3) \qquad \omega_{X^{(4)}}$$

Instances

1. x = positive integers, units of time
2. y = positive integers (optional)
3. ω = frequency, co-occurrences within window=5

$$X^{(1)} = (x_1) \qquad \omega_{X^{(1)}}$$

$$X^{(2)} = (x_1, x_2) \qquad \omega_{X^{(2)}}$$

$$X^{(3)} = (x_1, y) \qquad \omega_{X^{(3)}}$$

$$X^{(4)} = (x_1, x_2, x_3) \qquad \omega_{X^{(4)}}$$

Input numbers:

3 digits, within top 200 frequent numbers

Task prompt templates

Task	Prompt Template	#Test Cases
Arithematic		
Multiplication	*Q:What is* x_1 *times* x_2*? A:* y	5000
Addition	*Q:What is* x_1 *plus* x_2*? A:* y	5000
Operation Inference		
Mult. #	*Q:What is* x_1 *#* x_2*? A:* y	5000
Add. #	*Q:What is* x_1 *#* x_2*? A:* y	5000
Time Unit Inference		
Min→Sec	*Q:What is* x_1 *minutes in seconds? A:* y	79
Hour→Min	*Q:What is* x_1 *hours in minutes? A:* y	100
Day→Hour	*Q:What is* x_1 *days in hours? A:* y	100
Week→Day	*Q:What is* x_1 *weeks in days? A:* y	100
Month→Week	*Q:What is* x_1 *months in weeks? A:* y	100
Year→Month	*Q:What is* x_1 *years in months? A:* y	100
Decade→Year	*Q:What is* x_1 *decades in years? A:* y	100

: Vicky Zeng, Neha Verma, Lingfeng Shen

Task prompt templates

Task	Prompt Template	#Test Cases
Arithematic		
Multiplication	*Q:What is* x_1 *times* x_2*? A:* y	5000
Addition	*Q:What is* x_1 *plus* x_2*? A:* y	5000
Operation Inference		
Mult. #	*Q:What is* x_1 *#* x_2*? A:* y	5000
Add. #	*Q:What is* x_1 *#* x_2*? A:* y	5000
Time Unit Inference		
Min→Sec	*Q:What is* x_1 *minutes in seconds? A:* y	79
Hour→Min	*Q:What is* x_1 *hours in minutes? A:* y	100
Day→Hour	*Q:What is* x_1 *days in hours? A:* y	100
Week→Day	*Q:What is* x_1 *weeks in days? A:* y	100
Month→Week	*Q:What is* x_1 *months in weeks? A:* y	100
Year→Month	*Q:What is* x_1 *years in months? A:* y	100
Decade→Year	*Q:What is* x_1 *decades in years? A:* y	100

$(3, 11, 33)$
$(45, 54, 99)$

Task prompt templates

Task	Prompt Template	#Test Cases
Arithematic		
Multiplication	*Q:What is* x_1 *times* x_2*? A:* y	5000
Addition	*Q:What is* x_1 *plus* x_2*? A:* y	5000
Operation Inference		
Mult. #	*Q:What is* x_1 # x_2*? A:* y	5000
Add. #	*Q:What is* x_1 # x_2*? A:* y	5000
Time Unit Inference		
Min→Sec	*Q:What is* x_1 *minutes in seconds? A:* y	79
Hour→Min	*Q:What is* x_1 *hours in minutes? A:* y	100
Day→Hour	*Q:What is* x_1 *days in hours? A:* y	100
Week→Day	*Q:What is* x_1 *weeks in days? A:* y	100
Month→Week	*Q:What is* x_1 *months in weeks? A:* y	100
Year→Month	*Q:What is* x_1 *years in months? A:* y	100
Decade→Year	*Q:What is* x_1 *decades in years? A:* y	100

(3, 11, 33)
(45, 54, 99)

(3, 11, 33)
(45, 54, 99)

Task prompt templates

Task	Prompt Template	#Test Cases
Arithematic		
Multiplication	*Q:What is* x_1 *times* x_2*? A:* y	5000
Addition	*Q:What is* x_1 *plus* x_2*? A:* y	5000
Operation Inference		
Mult. #	*Q:What is* x_1 # x_2*? A:* y	5000
Add. #	*Q:What is* x_1 # x_2*? A:* y	5000
Time Unit Inference		
Min→Sec	*Q:What is* x_1 *minutes in seconds? A:* y	79
Hour→Min	*Q:What is* x_1 *hours in minutes? A:* y	100
Day→Hour	*Q:What is* x_1 *days in hours? A:* y	100
Week→Day	*Q:What is* x_1 *weeks in days? A:* y	100
Month→Week	*Q:What is* x_1 *months in weeks? A:* y	100
Year→Month	*Q:What is* x_1 *years in months? A:* y	100
Decade→Year	*Q:What is* x_1 *decades in years? A:* y	100

(3, 11, 33)
(45, 54, 99)

(3, 11, 33)
(45, 54, 99)

(24, minutes, 60, 1440)

: Vicky Zeng, Neha Verma, Lingfeng Shen

Performance Gap

Difference in accuracy between top 10% of instances and bottom 10% of instances (by frequency)

$$\Omega = \{(\omega_X^{(n)}, a^{(n)})\}$$
$$\Delta(\Omega) = \mathrm{Acc}(\Omega_{>90\%}) - \mathrm{Acc}(\Omega_{<10\%})$$

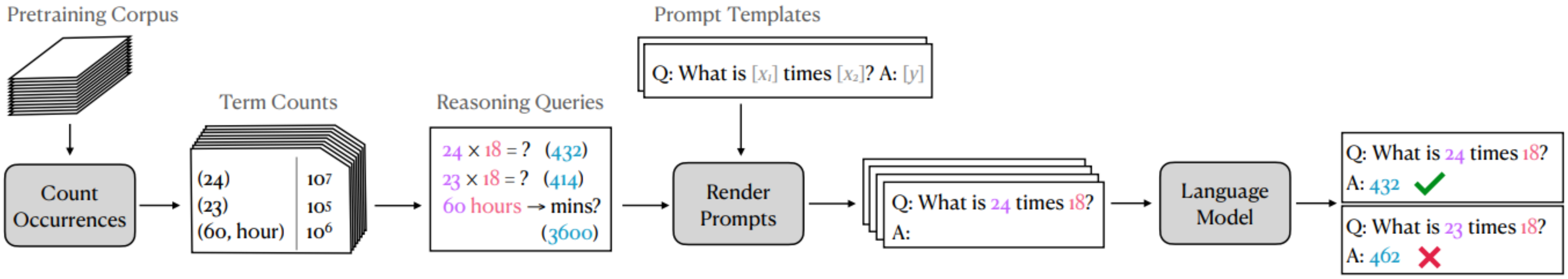
: Vicky Zeng, Neha Verma, Lingfeng Shen

Experimental setup

1. Models
   1. GPT-J-6B
   2. GPT-Neo-1.3B
   3. GPT-Neo-2.7B
2. Corpus
   1. Pile dataset
3. Prompt counts
   1. k = 0,2,4,8,16

: Vicky Zeng, Neha Verma, Lingfeng Shen

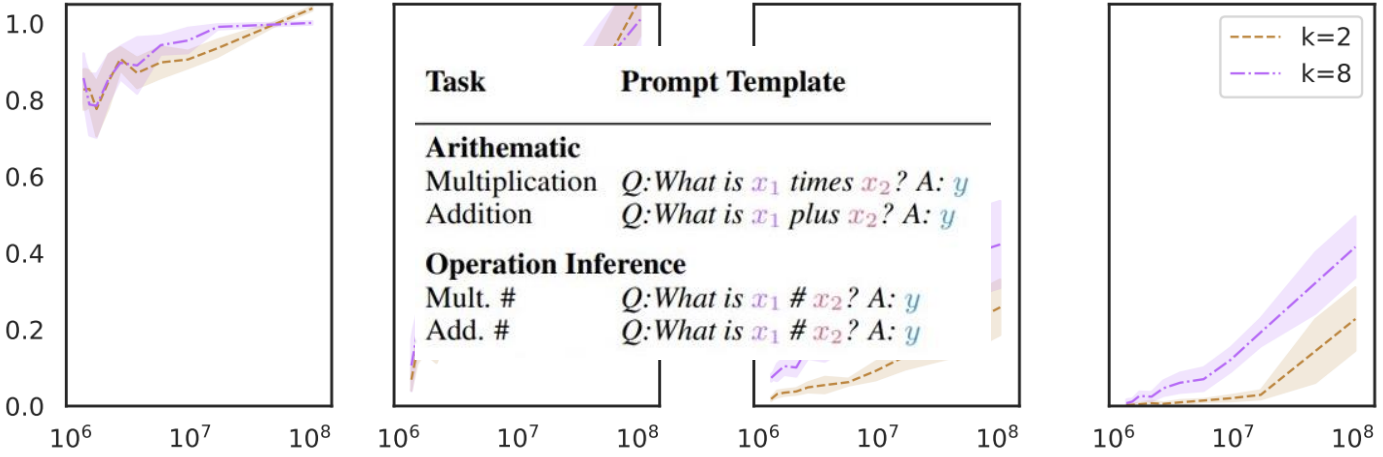
Pipeline for Data Construction



Main Finding: Heavy dependence on pretraining frequency

Main Finding: Heavy dependence on pretraining frequency

Strong positive correlation between test performance and pretraining term frequency



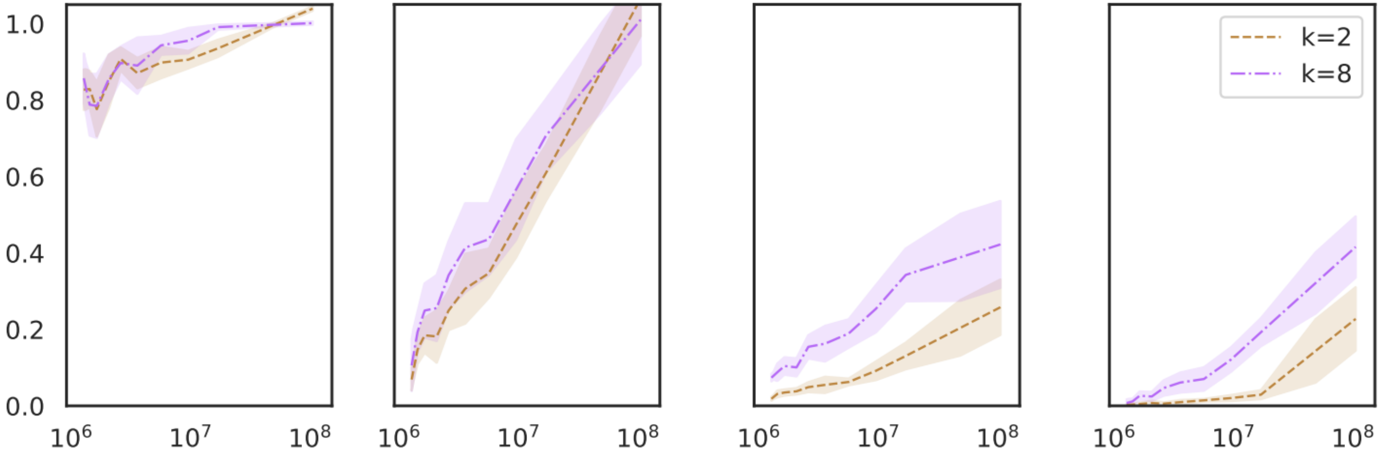
Task	Prompt Template
Arithematic	
Multiplication	*Q:What is x_1 times x_2? A: y*
Addition	*Q:What is x_1 plus x_2? A: y*
Operation Inference	
Mult. #	*Q:What is x_1 # x_2? A: y*
Add. #	*Q:What is x_1 # x_2? A: y*

(a) Arithmetic-Addition (b) Arithmetic-Multiplication (c) Op.Inference-Addition (d) Op. Inference-Multiplication

: Vicky Zeng, Neha Verma, Lingfeng Shen

Main Finding: Heavy dependence on pretraining frequency

Strong positive correlation between <u>test performance</u> and <u>pretraining term frequency</u>



(a) Arithmetic-Addition (b) Arithmetic-Multiplication (c) Op.Inference-Addition (d) Op. Inference-Multiplication

: Vicky Zeng, Neha Verma, Lingfeng Shen

Additional Support: Performance Gap, Inference vs Arithmetic Gap

- $\omega_{\{x_1\}}$: the number of times that x_1 (e.g., 23) appears in the pretraining data.
- $\omega_{\{x_1,x_2\}}$: the number of times that the input terms x_1 (e.g., 23) and x_2 (e.g., 18) appear in the pretraining data within a specific window size.
- $\omega_{\{x_1,y\}}$: the number of times that the first input term x_1 (e.g., 23) and the output term y (e.g., 414) appear in the pretraining data within a specific window size.

- Performance gap **increases** as number of *k* shots increases
- Inference task performance is much **lower** than arithmetic task performance

k	Addition				Multiplication				Addition (#)				Multiplication (#)			
	Acc.	Δ_1	$\Delta_{1,2}$	$\Delta_{1,y}$	Acc.	Δ_1	$\Delta_{1,2}$	$\Delta_{1,y}$	Acc.	Δ_1	$\Delta_{1,2}$	$\Delta_{1,y}$	Acc.	Δ_1	$\Delta_{1,2}$	$\Delta_{1,y}$
0	1.6	8.4	6.9	8.0	5.4	18.0	20.6	30.8	-	-	-	-	-	-	-	-
2	88.2	16.8	21.7	21.9	35.9	77.6	79.3	89.9	7.8	18.1	25.3	28.3	3.1	14.1	13.7	14.2
4	91.4	15.0	24.8	26.4	39.2	70.8	76.4	83.5	9.8	24.8	30.1	30.4	5.7	20.9	21.3	23.4
8	89.6	16.3	26.5	29.6	42.9	74.6	80.8	86.0	19.8	31.0	44.8	45.2	9.4	31.3	33.2	34.7
16	88.6	16.4	27.3	31.0	40.9	73.3	77.7	82.6	26.2	38.5	47.2	49.9	11.0	39.6	38.7	42.6

Generalization (Def.): A form of abstraction whereby common properties of specific instances are formulated as general concep ts or claims.

Outlier – Possible (limited) generalization

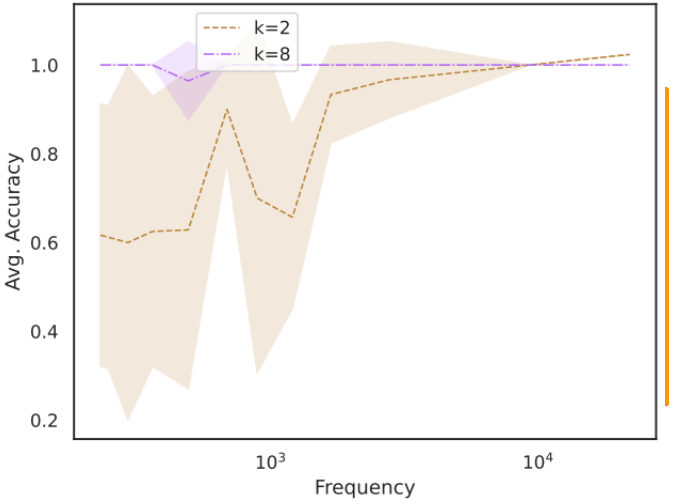
k	Min→Sec				Hour→Min				Day→Hour				Week→Day			
	Acc.	$\Delta_{1,2}$	$\Delta_{1,2,3}$	$\Delta_{1,2,y}$	Acc.	$\Delta_{1,2}$	$\Delta_{1,2,3}$	$\Delta_{1,2,y}$	Acc.	$\Delta_{1,2}$	$\Delta_{1,2,3}$	$\Delta_{1,2,y}$	Acc.	$\Delta_{1,2}$	$\Delta_{1,2,3}$	$\Delta_{1,2,y}$
0	1.3	0.0	0.0	12.5	1.0	0.0	0.0	5.0	1.0	0.0	0.0	10.0	1.0	0.0	0.0	10.0
2	25.5	62.5	67.5	67.5	19.4	58.0	40.5	44.0	12.1	28.9	24.0	28.0	13.1	43.5	50.0	54.0
4	35.5	60.0	71.7	63.1	29.1	76.4	50.5	59.0	22.7	46.4	45.0	47.5	19.2	40.9	43.3	47.0
8	49.9	72.1	79.0	52.7	36.3	74.6	52.5	63.0	31.0	59.1	52.5	54.5	28.6	70.6	62.0	67.0
16	58.4	82.7	74.4	48.5	42.8	80.1	49.0	62.5	43.3	62.8	56.0	54.8	28.0	22.1	31.4	33.2

Shots, k	Month→Week				Year→Month				Decade→Year			
	Acc.	$\Delta_{1,2}$	$\Delta_{1,2,3}$	$\Delta_{1,2,y}$	Acc.	$\Delta_{1,2}$	$\Delta_{1,2,3}$	$\Delta_{1,2,y}$	Acc.	$\Delta_{1,2}$	$\Delta_{1,2,3}$	$\Delta_{1,2,y}$
0	1.0	0.0	0.0	10.0	1.0	0.0	0.0	10.0	3.1	14.3	14.3	28.6
2	30.1	8.5	9.3	21.0	21.8	58.0	64.0	53.0	76.5	38.8	47.1	43.1
4	63.3	22.9	26.2	10.5	31.9	64.8	69.5	66.8	96.7	2.9	0.0	2.9
8	80.9	33.8	30.8	24.0	45.4	55.0	72.0	50.0	99.6	0.0	0.0	0.0
16	84.5	43.4	57.0	30.3	56.7	58.7	65.3	61.3	100.0	0.0	0.0	0.0

Outlier – Possible (limited) generalization

Task: Time-Unit Conversion for Decade -> Year

➢ Performance gap **disappears** as *k* increases

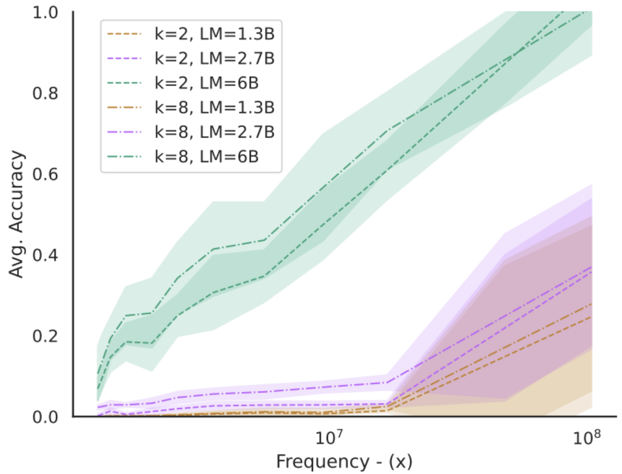


Why?

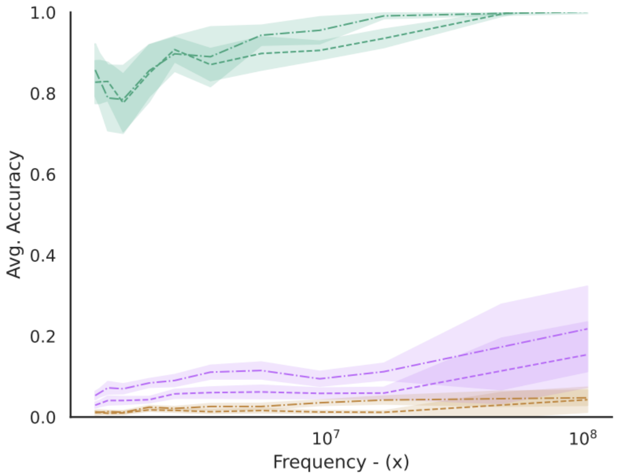
- Task simple enough to generalize?
- Bad frequency range, External factors that were neglected..?

Model Size on Performance

Models perform better on high-frequency terms across all model sizes



(a) Arithmetic-Multiplication



(b) Arithmetic-Addition

: Vicky Zeng, Neha Verma, Lingfeng Shen

Overview of the paper

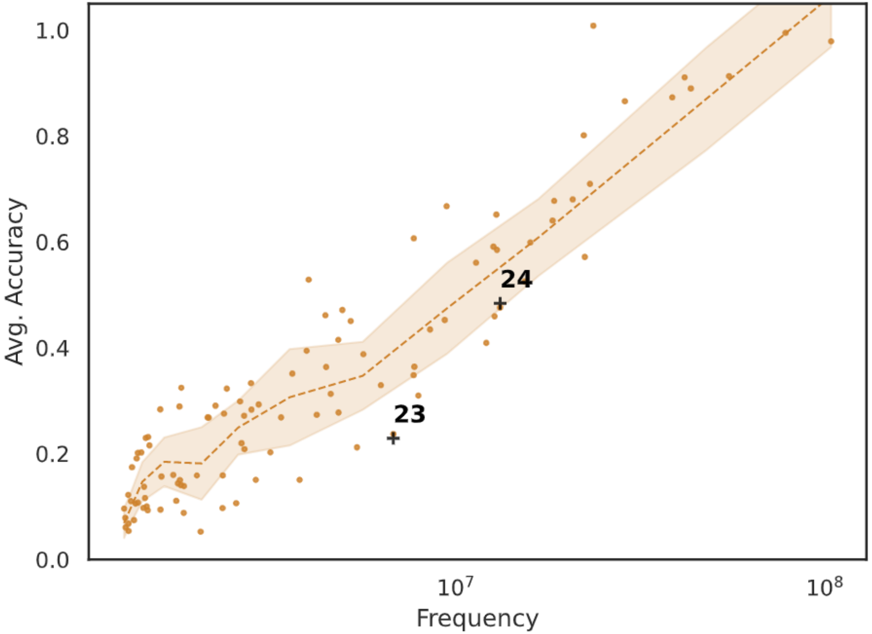
present analysis on these numerical reasoning tasks for three sizes of the EleutherAI/GPT models pretrained

show a consistently large performance gap between highest-frequency terms and lowest-frequency terms in all of our experiments

Call for a revisit evaluation of language models with respect to their pretraining data on numerical reasoning.

Paper Summary

Relation between term frequency and reasoning (over 11 tasks)

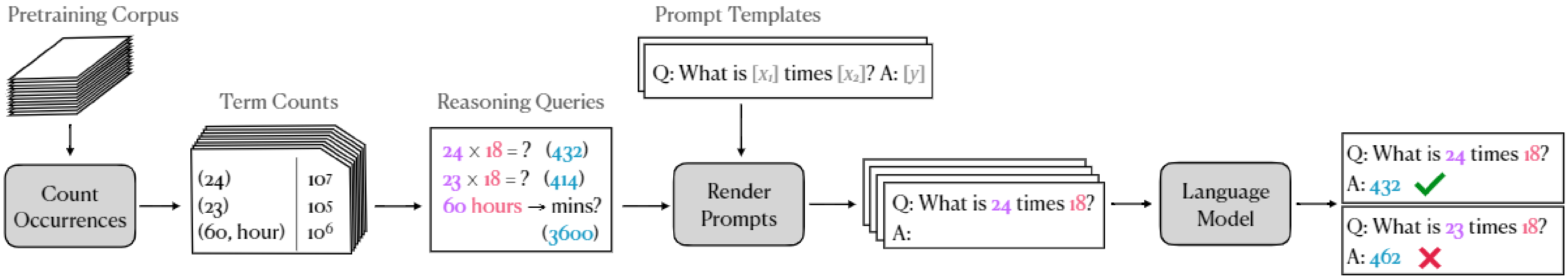


Strengths

- Stress the importance of pre-training dataset

Strengths

- Stress the importance of pre-training dataset
- Experiment setup is intuitive



Strengths

- Stress the importance of pre-training dataset
- Experiment setup is intuitive
- Consistent results over 11 tasks (arithmetic, operational, time unit conversion)

Strengths

- Stress the importance of pre-training dataset
- Experiment setup is intuitive
- Consistent results over 11 tasks (arithmetic, operational, time unit conversion)
- Reproducible methods and code

Strengths

- Stress the importance of pre-training dataset
- Experiment setup is intuitive
- Consistent results over 11 tasks (arithmetic, operational, time unit conversion)
- Reproducible methods and code
- Does well in tying work to related research

Weakness

- Limited to GPT models

Weakness

- Limited to GPT models
- Limited by numerical reasoning tasks
    - Analysis on commonsense reasoning would be interesting

Weakness

- Limited by numerical reasoning tasks
- Hard to explain the performance gap:
  - Does the gap come from memorization?
  - Other confounders?
    - Sentence length
    - Context (numbers occurs in arithmetic context during training?)
    - ...

Weakness

- Limited by numerical reasoning tasks
- Hard to explain the performance gap:
- Is "frequency vs performance" enough?
  - Frequency is a simple heuristic measurement
  - Even if frequency gives no performance gap, can we say model has reasoning ability?

Weakness

- Only talked about the result they found but didn’t explained in detail why this happened. Solution needed.
- Multiplication task.

Arithematic		
Multiplication	*Q:What is x_1 times x_2? A: y*	5000
Addition	*Q:What is x_1 plus x_2? A: y*	5000

  - Why The other operand was chosen from [1,50].
  - What if the number was very Unique .

: Steven, Aowei, Illiana

Models used for experiments

- GPT-Neo- 1.3B
- GPT-2
- GPT-3

gpt-neo-1.3B

https://colab.research.google.com/drive/1Nv4Qjmhe3PKenQy2OHeocfrV-y539jOC#scrollTo=hdGXP51_6NDc

https://huggingface.co/spaces/gradio/gpt-neo

gpt2

https://colab.research.google.com/drive/1rH1EvXmEbSnLjoi7nZgI8noxzlA1CvCI#scrollTo=AHJzdDt6Tagy

Take Away Messages

1.Low-order co-occurrence statistics impact reasoning tasks significantly

2.Pretraining data, unknown black box?

3.Characterzing the impacting factors on reasoning ability is still an issue

: Boyuan Zheng, Zhiqing Zhong

Short-Term Follow-Ups

Better benchmarks for reasoning ability considering the impact of the training data

1.Mathmatically and Logical as a playground

2.A benchmark without impact of the pretraining data

3.More general form of tasks (natural languages)

: Boyuan Zheng, Zhiqing Zhong

What about 5 years impacts?

More pretraining data aware benchmarks

- Quantify the impact through a set of metrics/tools and use that to investigate how much the model is influenced
- Remove data points heavily impacted by pretraining data out of the evaluation dataset
- Consider the impact of pretraining data when building the evaluation dataset

Frequency Effects on Syntactic Rule Learning in Transformers

- using the case study of BERT's performance on English subject–verb agreement.
- train multiple instances of BERT from scratch, allowing us to perform a series of controlled interventions at pre-training time.
- subject–verb pairs that never occurred in training
- performance is heavily influenced by word frequency
- What if we change the syntactic to logistic, semantic, etc.

Shortcomes…

- BERT appears to represent the correct rule but fails to predict agreement features for low frequency verb forms.
- BERT fails to apply the rule when doing so requires overcoming strong item-specific priors.