GEORGE MASON UNIVERSITY

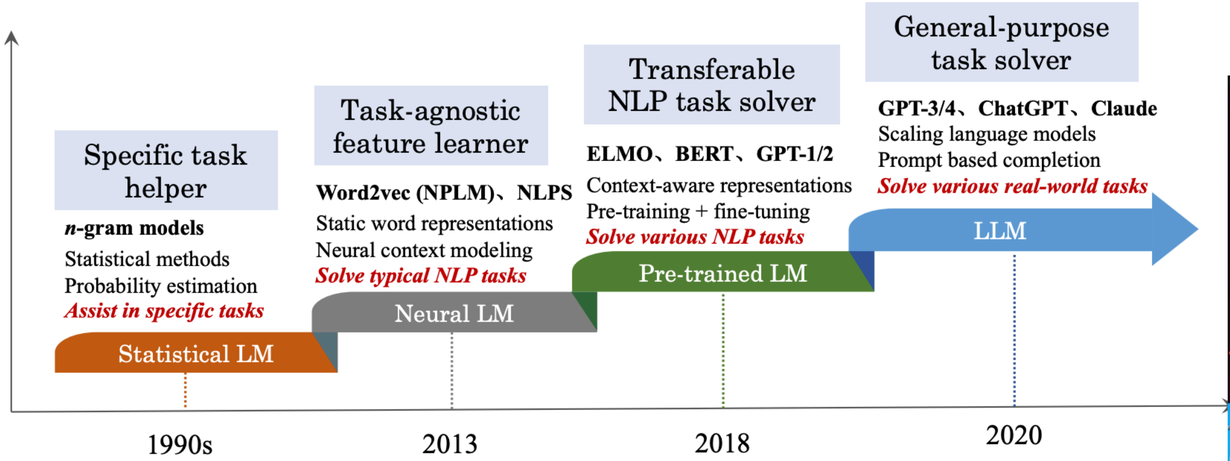
Language Models for Reasoning: The *How* and *Why*

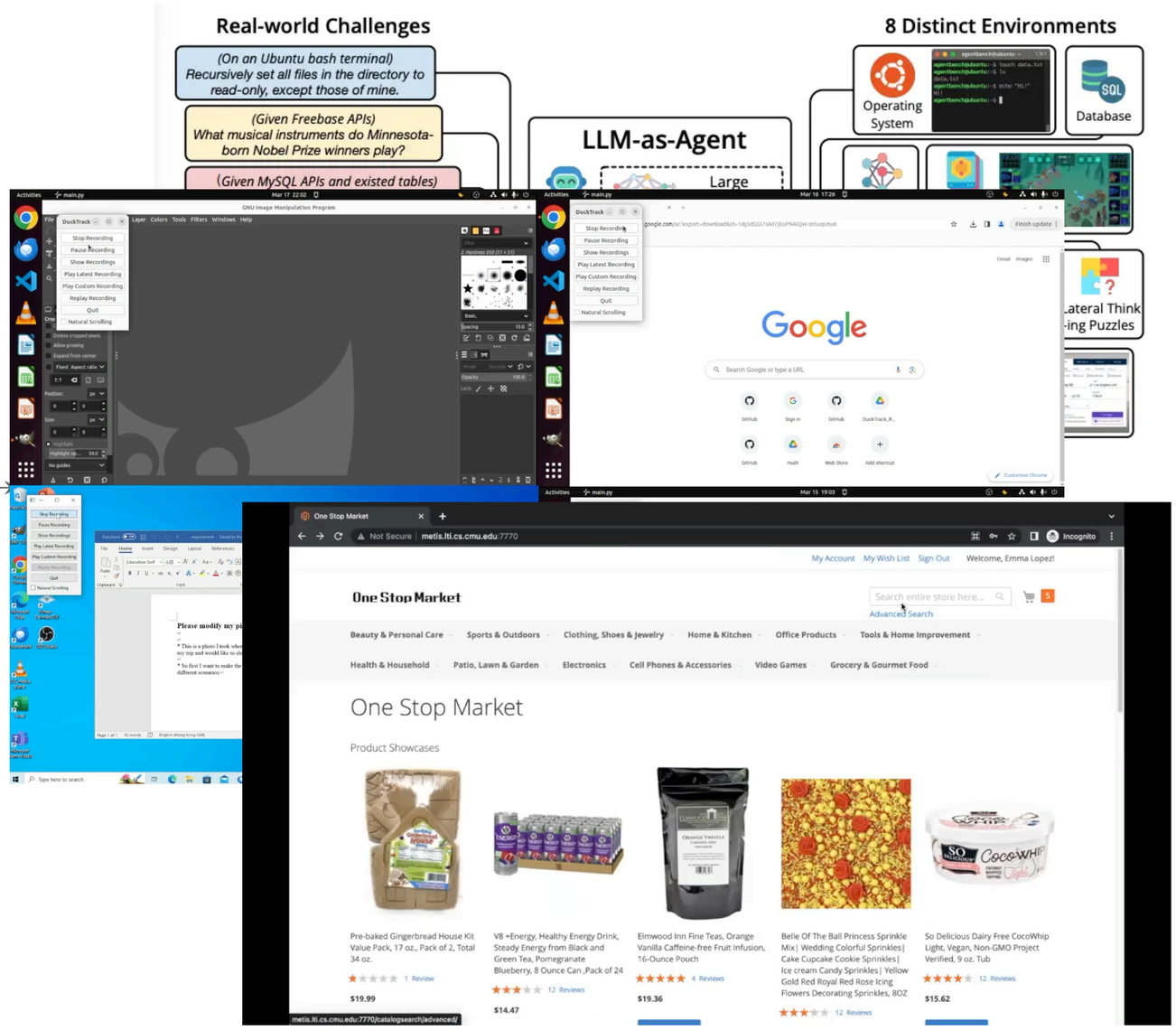
Ziyu Yao
Assistant Professor, Department of Computer Science
George Mason University
https://ziyuyao.org/

Georgetown University
April 11, 2025

Language Models (LMs)

- The paradigm shift





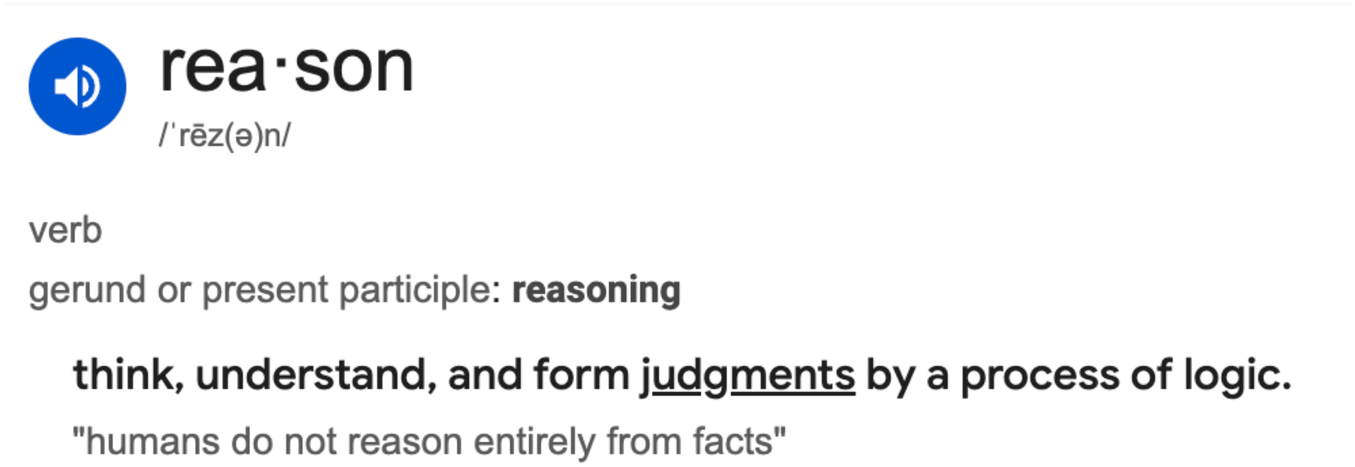
Data Science Agent, Software Agent, Chemistry Agent (AI4Science), etc.

Zhao, Wayne Xin, et al. "A survey of large language models." arXiv preprint arXiv:2303.18223 (2023).

AgentBench (Liu et al., 2023); OSWorld (Xie et al., 2024); WebArena (Zhou et al., 2023)

“Reasoning”

- A cornerstone of machine intelligence
- An umbrella term encompassing various instantiations



LMs for Arithmetic Reasoning

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

Answer: ?

LMs for Arithmetic Reasoning via Chain of Thought (CoT)

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

Answer: 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.

One-shot CoT demonstration

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?

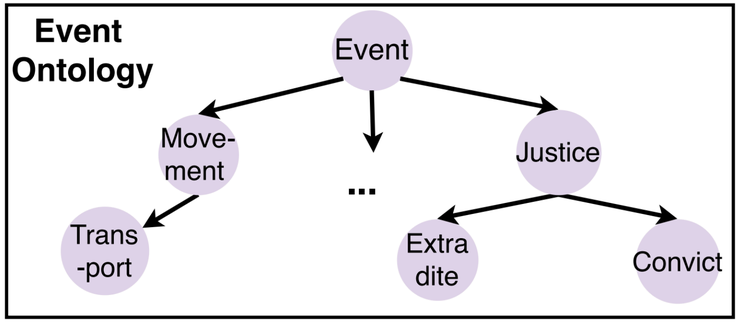
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

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

(Wei et al., 2022)

e.g., zero-shot GPT-4 on GSM8k (Cobbe et al., 2021): 92%+ Accuracy

LMs for Event Extraction (EE)

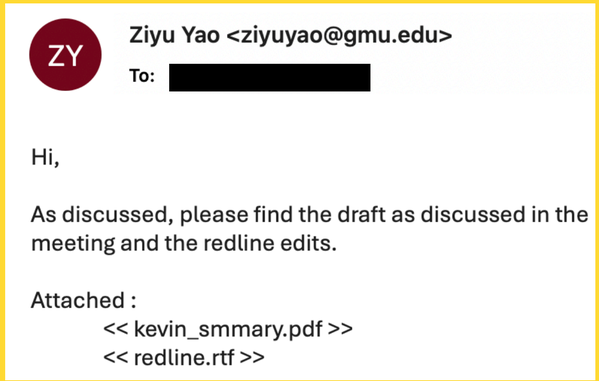


Text: "After getting caught they were transferred to the U.S. for trial."

What event(s) does this sentence include?

GPT-3.5-turbo, few-shot (Huang et al., 2024):
11.8% F1 for Event Detection, 23.8% F1 for Argument Extraction

Text:
(conversational email response)



What event(s) does this sentence include?

GPT-3.5-turbo, few-shot (Srivastava et al., 2023):
4.8% F1 for end-to-end EE, 21.1% F1 for Argument Extraction

vs. fine-tuned BART:
36.3% F1 for end-to-end EE, 56.0% F1 for Argument Extraction

Saurabh Srivastava (4th-yr PhD at GMU CS)

Huang et al. "TextEE: Benchmark, Reevaluation, Reflections, and Future Challenges in Event Extraction." *ACL (Findings)*. 2024.
Srivastava et al. "MailEx: Email Event and Argument Extraction." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023.

Language Models for Reasoning

- **Topic 1: *How*** do we unlock LMs' capabilities in performing complicated reasoning tasks?
  - Using event extraction as an exemplar task
  - How to instruct or prompt an LM for more effective event extraction?

- **Topic 2: *Why*** an LM can or cannot perform reasoning?
  - Using arithmetic reasoning as an exemplar task
  - Mechanistic understanding of how Chain-of-Thought (CoT) prompting elicits an LM's arithmetic reasoning capability

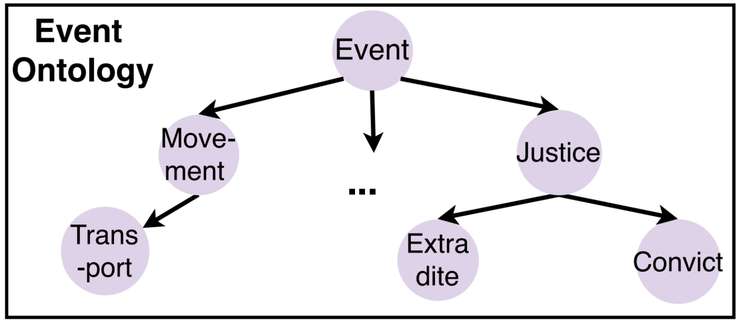
Language Models for Reasoning

- **Topic 1: *How*** do we unlock LMs' capabilities in performing complicated reasoning tasks?
  - Using event extraction as an exemplar task
  - How to instruct or prompt an LM for more effective event extraction?

- **Topic 2: *Why*** an LM can or cannot perform reasoning?
  - Using arithmetic reasoning as an exemplar task
  - Mechanistic understanding of how Chain-of-Thought (CoT) prompting elicits an LM's arithmetic reasoning capability

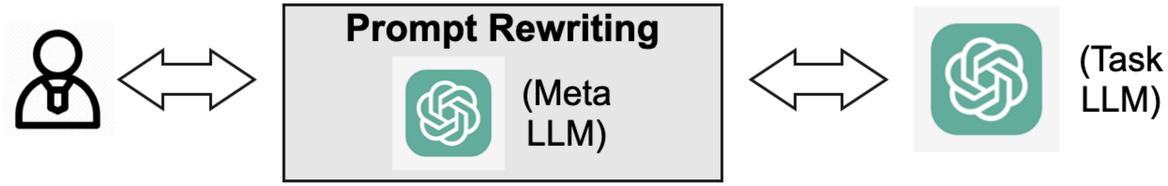
What Makes Event Extraction (EE) Challenging?

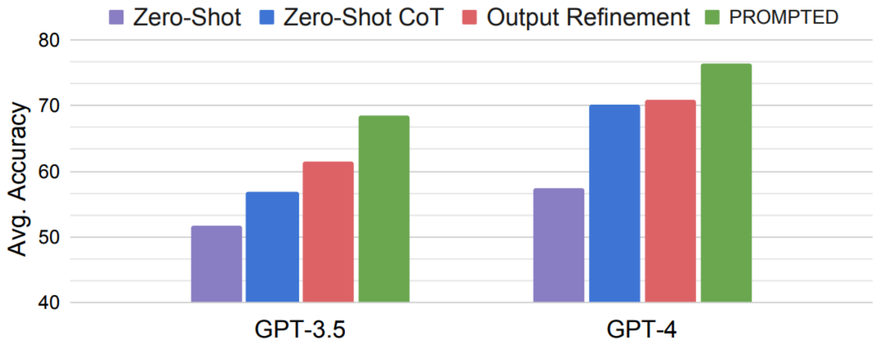
- The task itself is not trivial!
  - Hierarchical event schema
    - Nuanced semantic difference & fine-grained argument types
  - Dataset artifacts



Schema example of ACE05 (Doddington et al., 2004)

- LMs are sensitive to how they are prompted
  - You have to know how to properly instruct or prompt them, e.g.,





Srivastava, Saurabh, et al. "Instances Need More Care: Rewriting Prompts for Instances with LLMs in the Loop Yields Better Zero-Shot Performance." *Findings of ACL* 2024.

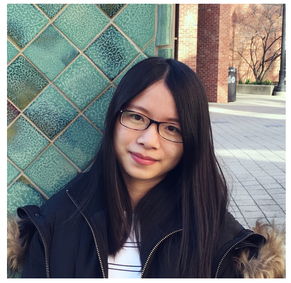
Instruction-Tuning LLMs for Event Extraction with Annotation Guidelines

Saurabh Srivastava*, Sweta Pati*, Ziyu Yao
George Mason University

Revisiting Prompt Optimization with Large Reasoning Models---A Case Study on Event Extraction

Saurabh Srivastava, Ziyu Yao
George Mason University





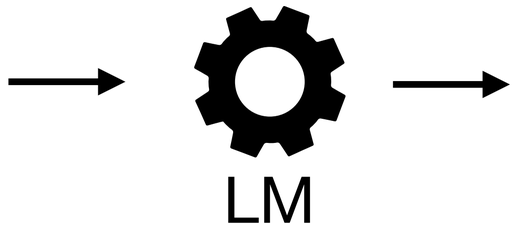
Code Format Facilities Event Extraction

- Wang, Li, and Ji (2023) found that LMs perform better when they read/write the structured event information in code format

```
# This is an event extraction task ...
# The following lines describe the task definition
@dataclass
class Extradite(JusticeEvent):
    mention
    agent
    person
    destination

# This is the text to analyze
text = After getting caught they were transferred to
       the U.S. for trial.
```

Event Schema




```
Extradite(
  mention = “transferred”,
  person = [“they”],
  destination = [“U.S.”]
)
```

(Img source: Srivastava et al., 2025)

LMs were pre-trained to read and write code;

Structured knowledge representation: easier to represent hierarchical schema, type-argument association, type constraints, etc.

Wang, Xingyao, Sha Li, and Heng Ji. "Code4Struct: Code Generation for Few-Shot Event Structure Prediction." *ACL* 2023.

Annotation Guidelines Serve as Helpful Hints

- Saniz et al. (2024) further discovered that the expert-written annotation guidelines, formatted as Python docstrings and comments, are helpful hints
  - Mainly verified in Named Entity Recognition

```
# The following lines describe the task definition
@dataclass
class ProgrammingLanguage(Entity):
    """Refers to a programming language used in the development of AI
    applications and research. Annotate the name of the programming
    language, such as Java and Python."""

    span: str  # Such as: "Java", "R", "CLIPS", "Python", "C + +"

@dataclass
class Metric(Entity):
    """Refers to evaluation metrics used to assess the performance of AI
    models and algorithms. Annotate specific metrics like F1-score."""

    span: str  # Such as: "mean squared error", "DCG", ...
```

Our work: Does including guidelines help the more challenging Event Extraction task?

Sainz, Oscar, et al. "GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction." *ICLR 2024*.

Event Extraction w/ Guidelines in Code Format

```
# This is an event extraction task ...
# The following lines describe the task definition
@dataclass
class Extradite(JusticeEvent):
"""The event is triggered by the formal request and subsequent transfer of an individual from one state or country to another for legal reasor
indicative of this event type, not 'Transport' which involves general movement without legal context."""
    mention  # The text span that triggers the event.
    agent  # The agent plays a crucial role in the extradition process, often being a legal or governmental body.
    person # Examples are 'she', 'him', 'her'. The person is the individual being extradited.
    destination  # Examples are 'jurisdiction', 'Hague', 'state'. The destination is the place to which the person is being extradited.
# This is the text to analyze
text = After getting caught they were transferred to the U.S. for trial.
```

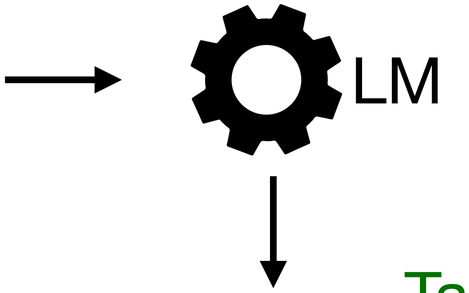
- Problem setup: instruction tuning (FLAN/Google Research 2022)

Wei, Jason, et al. "Finetuned Language Models are Zero-Shot Learners." *ICLR 2022*.

Event Extraction w/ Guidelines in Code Format

Instruction

```
# This is an event extraction task ...
# The following lines describe the task definition
@dataclass
class Extradite(JusticeEvent):
"""The event is triggered by the formal request and subsequent transfer of an in
indicative of this event type, not 'Transport' which involves general movement w
    mention    # The text span that triggers the event.
    agent  # The agent plays a crucial role in the extradition process, often l
    person # Examples are 'she', 'him', 'her'. The person is the individual bei
    destination  # Examples are 'jurisdiction', 'Hague', 'state'. The destina
# This is the text to analyze
text = After getting caught they were transferred to the U.S. for trial.
```



Target to maximize

```
Extradite(
  mention = “transferred”,
  person = [“they”],
  destination = [“U.S.”]
)
```

- Problem setup: instruction tuning (FLAN/Google Research 2022)
- Additionally,
  - Can machines generate more effective guidelines?
  - How do guidelines help in low-data settings (i.e., small amounts of training examples)?

Wei, Jason, et al. "Finetuned Language Models are Zero-Shot Learners." *ICLR 2022*.

Automatic Generation of Annotation Guidelines

- Expert-written guidelines are not always available and may not be the most helpful to LMs
- Can SOTA LMs (e.g., GPT-4o) *reverse-engineer* guidelines from data?

Guideline Generation Prompt (Guideline-PN)

You are an expert in annotating NLP datasets for event extraction. Your task is to generate annotation guidelines for the event type Extradite which is a child event type of super class JusticeEvent.

The event schema is as follows:

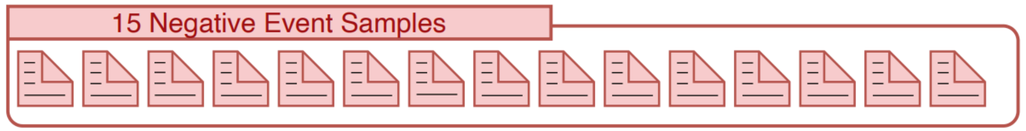
```
@dataclass
class Extradite(JusticeEvent):
    mention
    (...)
```

The below examples are positive examples, as they match the Event Type being annotated

10 Positive Event Samples

The following examples are negative examples, as they illustrate different event types provided for contrast and differentiation:

15 Negative Event Samples



Instructions
1. Identify and List All Unique Arguments.
2. Define the Event Type: Write 5 clear and specific definitions, starting with "The event is triggered by ...":
3. Define Each Argument:** For each argument, provide 5 definitions.

More Task Instructions (...)

Variants:

- Guideline-Positive (P)
- Guideline-Positive+Negative (PN)
- Guideline-Positive+Siblings (PS)

(More variants in paper)

Automatic Generation of Annotation Guidelines

Examples of Annotation Guidelines for Event Type: Extradite (ACE05)	
GUIDELINE-H Avg. Length - 107.67 tokens	**Event Type:** An EXTRADITE Event occurs whenever a PERSON is sent by a state actor from one PLACE to another place for the purposes of legal proceedings there. **Arguments:** - AGENT: The extraditing agent. - PERSON: The person being extradited.
GUIDELINE-P Avg. Length - 163.87 tokens	**Event Type:** The Extradition event refers to the formal process where one jurisdiction delivers a person accused **(...)** The event can be triggered by terms such as 'extradition' **(...)** Edge cases include situations where the term 'extradition' is used metaphorically or in a non-legal context. **Arguments:** - AGENT:**(...)** the agent is the organization or authority **(...)**. Examples include 'court', 'government', **(...)** - PERSON: **(...)** individual who is being transferred to another jurisdiction. Examples are 'she', **(...)**
GUIDELINE-PN Avg. Length - 285.24 tokens	**Event Type:** The event is triggered by the formal request **(...)** for legal reasons. Triggers such as 'extradition' are indicative of this event type, not 'Transport' which involves general movement without legal context. **Arguments:** - AGENT: The agent is responsible for the legal and procedural aspects of the extradition,**(...)**. An example is 'the original court' **(...)** - PERSON: **(...)** one who is being moved from one place to another under legal authority. For example, 'he' **(...)**

(Notation: distinctions from other event types, example mentions, and edge cases)

Does including guidelines help EE?

- Datasets: ACE05 (Doddington et al., 2004) and RichERE (Song et al., 2015)
- Model: Llama-3.1-8B (Meta GenAI, 2024)
- Two settings: w/o or w/ Negative Sampling (NS)
  - NS: 15 negative samples per training example (15x more training examples)

Full training set: 16k and 9k

Experiments	ACE w/o NS				ACE w/ NS				RichERE w/o NS				RichERE w/ NS			
	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC
NoGuideline	39.57	39.57	31.05	29.73					35.11	35.11	27.16	25.32				
Guideline-H	40.71	40.71	30.76	28.64					–	–	–	–				
Guideline-P	**51.46**	**51.46**	**37.82**	**35.20**					34.38	34.38	28.04	26.35				
Guideline-PN	49.60	49.60	35.80	32.81					**40.89**	**40.89**	**30.04**	**27.18**				
Guideline-PS	47.93	47.93	37.19	34.88					32.41	32.41	24.63	22.78				

Guidelines help; Machine Guidelines > Human Guidelines;

Does including guidelines help EE?

- Datasets: ACE05 (Doddington et al., 2004) and RichERE (Song et al., 2015)
- Model: Llama-3.1-8B (Meta GenAI, 2024)
- Two settings: w/o or w/ Negative Sampling (NS)
  - NS: 15 negative samples per training example (15x more training examples)

Full training set: 16k and 9k

Experiments	ACE w/o NS				ACE w/ NS				RichERE w/o NS				RichERE w/ NS			
	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC
NoGuideline	39.57	39.57	31.05	29.73	**84.15**	**84.15**	**64.99**	**61.96**	35.11	35.11	27.16	25.32	42.27	42.27	32.38	31.56
Guideline-H	40.71	40.71	30.76	28.64	56.30	56.30	44.82	43.13	–	–	–	–	–	–	–	–
Guideline-P	**51.46**	**51.46**	**37.82**	**35.20**	72.86	72.86	55.01	53.73	34.38	34.38	28.04	26.35	67.92	67.92	52.29	44.93
Guideline-PN	49.60	49.60	35.80	32.81	80.77	80.77	63.20	60.34	**40.89**	**40.89**	**30.04**	**27.18**	75.35	75.35	**60.85**	**57.10**
Guideline-PS	47.93	47.93	37.19	34.88	79.23	79.23	59.00	56.88	32.41	32.41	24.63	22.78	**76.45**	**76.45**	60.42	56.26

Guidelines help; Machine Guidelines > Human Guidelines;

Guidelines may or may not complement Negative Sampling

How do guidelines help in low-data settings?

- Hypothesis: guidelines gain more in low-data settings with the extracted data heuristics

Training size: 2k

Experiments	ACE w/o NS				ACE w/ NS				RichERE w/o NS				RichERE w/ NS			
	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC
NoGuideline	10.60	10.60	5.19	3.68	31.64	31.64	25.91	24.22	19.87	19.87	13.34	11.69	36.29	36.29	28.15	25.58
Guideline-H	29.01	29.01	16.37	14.78	32.62	32.62	25.35	22.87	–	–	–	–	–	–	–	–
Guideline-P	36.91	36.91	24.17	21.24	56.99	56.99	**43.44**	**40.51**	**40.28**	**40.28**	**21.97**	**18.33**	62.04	62.04	46.33	42.03
Guideline-PN	30.94	30.94	19.27	17.64	**60.29**	**60.29**	42.88	39.95	31.23	31.23	19.48	17.51	**67.16**	**67.16**	**47.85**	**43.39**
Guideline-PS	**40.53**	**40.53**	**28.03**	**26.12**	55.1	55.1	41.57	38.91	26.16	26.16	16.64	15.19	58.95	58.95	42.79	38.1

2k training w/ Guidelines + NS outperforms full-size training

Full training set: 16k

Experiments	ACE w/o NS			
	TI	TC	AI	AC
NoGuideline	39.57	39.57	31.05	29.73

Full training set: 9k

RichERE w/o NS				RichERE w/ NS			
TI	TC	AI	AC	TI	TC	AI	AC
35.11	35.11	27.16	25.32	42.27	42.27	32.38	31.56

How do guidelines help in low-data settings?

- Hypothesis: guidelines gain more in low-data settings with the extracted data heuristics

Training size: 100

	ACE w/ NS				RichERE w/ NS			
	TI	**TC**	**AI**	**AC**	**TI**	**TC**	**AI**	**AC**
NoGuide	**37.08**	**37.08**	**21.53**	**19.18**	24.98	24.98	15.05	13.15
H	29.00	29.00	17.93	16.34	–	–	–	–
P	27.95	27.95	15.94	14.21	23.93	23.93	13.56	12.71
PN	29.60	29.60	17.87	15.92	27.43	27.43	**17.10**	**15.28**
PS	29.85	29.85	19.49	17.04	19.61	19.61	11.77	10.48

But when the training size is too small, LMs cannot be effectively instruction-tuned to utilize the guidelines

More details in our paper!

- After an LM is tuned to utilize the task instruction, will it also gain more **generalization** ability?
  - Yes! See our cross-schema generalization results
- Do the guidelines help **smaller** LMs? **Non-Llama** LMs?
  - Yes! We observed similar patterns w/ Llama-3.2-1B and Qwen2.5-Coder-1.5B
- What event types benefit the most from guidelines?
  - **Frequent and less frequent event types both benefit from guidelines**, but for extremely infrequent event types, LMs cannot use guidelines well

Do We Still Need Better Prompts w/ Large Reasoning Models?

- Early 2025: DeepSeek-R1 blowed everyone's mind
- The emerging model type: **Large Reasoning Models (LRMs)**
  - OpenAI o1, DeepSeek-R1

Prompt Engineering Techniques to Maximize Performance

Keep Prompts Clear and Minimal

Be concise and direct with your ask. Because O1 and O3 perform intensive internal reasoning, they respond best to *focused questions or instructions without extraneous text*. OpenAI and recent research suggest avoiding overly complex or leading prompts for these models. In practice, this means you should **state the problem or task plainly and provide only necessary details**. There is no need to add "fluff" or multiple rephrasing of the query. For example, instead

Can LRMs solve Event Extraction? Is it still worth exploring better prompts to them?

https://techcommunity.microsoft.com/blog/azure-ai-services-blog/prompt-engineering-for-openai%E2%80%99s-o1-and-o3-mini-reasoning-models/4374010

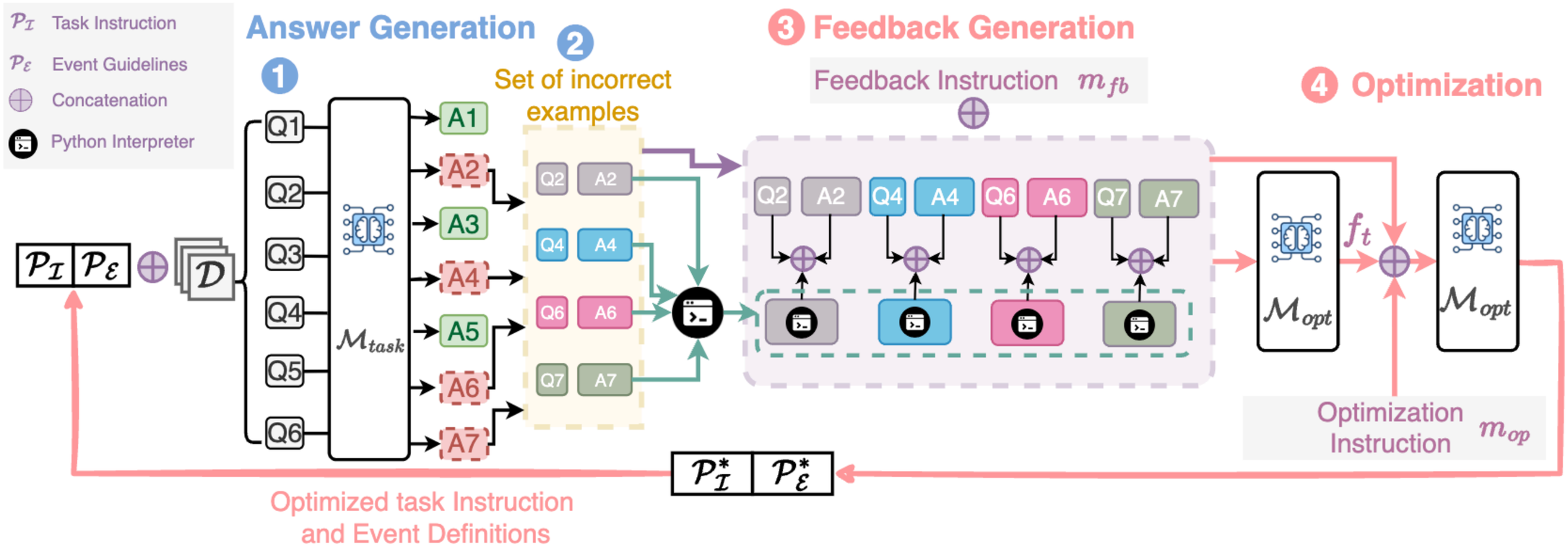
Large Reasoning Models for Event Extraction

- Problem is not solved! Event extraction is proven to still be challenging
- Experimental setup: zero-shot prompt engineering w/ task instruction and event schemas in Python code format
- Argument Classification F1 on a simplified* ACE05 dataset:
  - LLMs: GPT-4o (12.68%), GPT-4.5 (16.47%)
  - LRMs: o1 (13.94%), DeepSeek-R1 (16.45%)

(*a subset w/ 10 event types to avoid long-context inputs)

Prompt Optimization

- Assuming a small training set, can we discover a more effective prompt? (Zhou et al., 2022; Yang et al., 2024; Srivastava et al., 2024)
- Prompt optimization for event extraction:



Srivastava and Yao, Revisiting Prompt Optimization with Large Reasoning Models---A Case Study on Event Extraction, preprint 2025.

Prompt Optimization

	Optimizer LLMs/LRMs ($\mathcal{M}_{opt}$)				
$\mathcal{M}_{task}$	*No Opt.*	**GPT-4o**	**GPT-4.5**	**o1**	**DS-R1**
			ACE$_{low}$		
GPT-4o					
GPT-4.5					
o1					
DS-R1					
			ACE$_{med}$		
GPT-4o					
GPT-4.5					
o1					
DS-R1					

Training size: 15 (1 example per event type + 5 non-events)

Training size: 120 (10 example per event type + 20 non-events)

Prompt Optimization

$\mathcal{M}_{task}$	Optimizer LLMs/LRMs ($\mathcal{M}_{opt}$)					#Output
	No Opt.	GPT-4o	GPT-4.5	o1	DS-R1	Tokens
			ACE$_{low}$			
GPT-4o	12.68	18.18 +5.50	16.67 +3.99	18.83 +6.15	**20.15** +7.47	15.31
GPT-4.5	16.47	19.33 +2.86	16.47 00.00	19.32 +2.85	**22.31** +5.84	24.57
o1	13.94	18.96 +5.02	18.57 +4.63	20.29 +6.35	**21.92** +7.98	489.67
DS-R1	16.45	18.67 +2.22	18.57 +2.12	21.83 +5.38	**24.66** +8.21	217.71
			ACE$_{med}$			
GPT-4o	12.68	22.32 +9.64	**27.54** +14.86	26.30 +13.62	25.10 +12.42	17.31
GPT-4.5	16.47	29.63 +13.16	35.94 +19.47	**36.51** +20.04	35.42 +18.95	28.75
o1	13.94	30.19 +16.25	36.67 +22.73	**36.98** +23.04	36.96 +23.02	543.45
DS-R1	16.45	32.20 +15.75	37.14 +20.69	38.77 +22.32	**40.00** +23.55	277.11

Training size: 15 (1 example per event type + 5 non-events)

Training size: 120 (10 example per event type + 20 non-events)

Both LLMs and LRMs benefit from prompt optimization, and LRMs gain further; LRMs are better prompt optimizers than LLMs.

Color notation:
task instruction output format
extraction rules exception handling

Prompt Examples: Instruction

Examples of Task Instructions Optimized by Different Models	
NO OPTIMIZATION	# This is an event extraction task where the goal is to extract structured events from the text following structured event definitions in Python. **(...)** For each different event type, please output the extracted information from the text into a python list format **(...)** you should always output in a valid pydantic format: result = [EventName(”mention” = ”trigger”, ”arg1_key” = ”arg1_span”, ...), EventName(”mention” = ”trigger”, ”arg1_key” = ”arg1_span”, ...)]. **(...)**
DEEPSEEK-R1	# Event Extraction Task: Extract structured events from text using Python class definitions.**(...)**: 1. Span Extraction:- Triggers: Minimal contiguous spans (verbs/nouns) directly expressing the event. Include both verbal and nominal forms **(”death” = Die, ”killings” = Die)**.**(...)** - Arguments: - Remove articles **(”a/an/the”)** and possessive pronouns EXCEPT when part of official names or temporal phrases **(”The Hague”, ”the past year”)** - Resolve pronouns AND POSSESSIVE NOUNS to named entities immediately using same-sentence antecedents **(”airline’s plan” → [”airline”])** - Strip role/location/age descriptors from arguments **(”Philadelphia lawyers” → ”lawyers”) (...)** - Keep FULL spans for crimes/money including sources/amounts **(”stereo worth $1,750 from family”)** unless legal terms **(...)** 2. Special Handling:- Bankruptcy Triggers: **”went bust” → EndOrg(...)** - Crime Spans: Retain full contextual clauses **(”If convicted of killings...”)** without truncation - Temporal Phrases: Keep original spans with articles when part of phrase **(”the early 90’s”)** 3. Output Rules: Always output in Python-format as **(...)** 4. Critical Exceptions:**-(...)**

(Optimized by DeepSeek-R1)

Prompt Examples: Guidelines

```
class Convict(JusticeEvent):
    """Extract convictions where entity is found guilty of crime.
    Key Updates:
    - crime: Retain FULL spans including amounts/sources ("received
        stereo worth $1,750 from family")
    Example: "convicte
        bribes worth $1
    Counterexample: Tr
    """
    mention: str  # Tr
    defendant: List[st
        descriptors)
    adjudicator: List[
    crime: List[str]
    time: List[str]  #
    place: List[str]
        )
```

```
class DeclareBankruptcy(BusinessEvent):
    """Formal bankruptcy declarations.
    Key Rules:
    - entity: Resolve org pronouns AND possessive nouns ("airline's
        bankruptcy" → ["airline"])
    - Triggers: "bankruptcy", "Chapter 11" (exclude "collapse"/"went bust
        " without explicit bankruptcy context)

    Example: "airline's bankruptcy filing" → mention="bankruptcy", org=["
        airline"]
    Counterexample: "near-collapse" → EndOrg
    """
    mention: str  # Triggers indicating financial collapse: "bankruptcy",
         "Chapter 11"
    entity: List[str]  # ["Enron Corp"] (resolved orgs from pronouns/
        possessives in same sentence)
    time: List[str]  # ["2003"] (declaration time phrase)
    place: List[str]  # ["Texas"] (jurisdiction noun if specified)
```

Take-aways

- Better prompts are still needed for challenging reasoning tasks such as event extraction
- LLMs/LRMs for **insight discovery from data**
  - Discover rules and heuristics from data
  - Discover *generalizable* insights from a *small amount* of data
- How is this helpful?
  - Improving model performance
  - LLMs/LRMs as assistants in data annotation: defining guidelines/standards, identifying pitfalls in existing guidelines, discovering new data types, etc.

Language Models for Reasoning

- **Topic 1: *How*** do we unlock LMs' capabilities in performing complicated reasoning tasks?
  - Using event extraction as an exemplar task
  - How to instruct or prompt an LM for more effective event extraction?

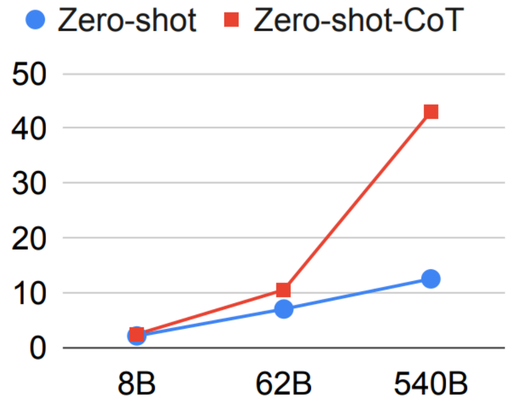
- **Topic 2: *Why*** an LM can or cannot perform reasoning?
  - Using arithmetic reasoning as an exemplar task
  - Mechanistic understanding of how Chain-of-Thought (CoT) prompting elicits an LM's arithmetic reasoning capability

Optimizing Zero-Shot Prompts for Arithmetic Reasoning

- Chain-of-Thought (CoT) w/ the magic spell of "Let's think step by step" (Kojima et al., 2022)

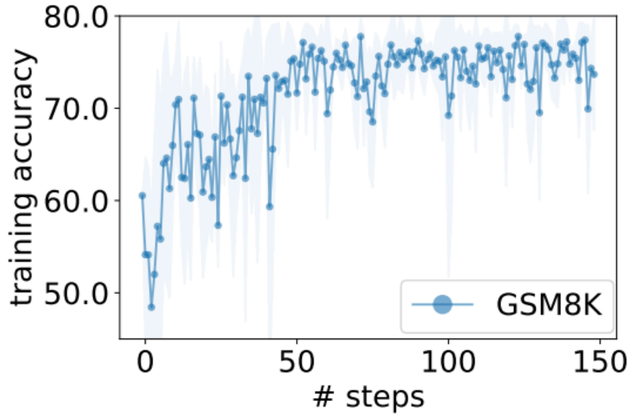
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓



PaLM on GSM8k (Cobbe et al., 2021)

- LLMs as prompt optimizers (Yang et al., 2023)



Source	Instruction	Acc
Baselines		
(Kojima et al., 2022)	Let's think step by step.	71.8
(Zhou et al., 2022b)	Let's work this out in a step by step way to be sure we have the right answer.	58.8
	(empty string)	34.0
Ours		
`PaLM 2-L-IT`	Take a deep breath and work on this problem step-by-step.	**80.2**
`PaLM 2-L`	Break this down.	79.9
`gpt-3.5-turbo`	A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem.	78.5
`gpt-4`	Let's combine our numerical command and clear thinking to quickly and accurately decipher the answer.	74.5

Optimizing Few-Shot CoT Prompts for Arithmetic Reasoning

- Empirical studies found important designs in few-shot CoT prompts

Research Questions	Examples in CoT Prompts (few-shot demonstrations)	Prior Work	Findings
Does equation matter? (RQ3)	**w Equation:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been **21 - 15 = 6** trees. The answer is 6. **w/o Equation:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been **6** trees. The answer is 6.	Wang et al. (2023); Ye et al. (2023); Madaan and Yazdanbakhsh (2022)	Yes
Does textual explanation matter? (RQ4)	**w Textual Explanation:** Let's think step by step. First Leah had 32 chocolates and her sister had 42 chocolates. So in total they had 32 + 42 = 74 chocolates. Then they ate 35 chocolates. So there must be 74 - 35 = 39 chocolates. The answer is 39. **w/o Textual Explanation:** 32 + 42 = 74. 74 - 35 = 39. The answer is 39.	Wang et al. (2023); Ye et al. (2023); Madaan and Yazdanbakhsh (2022)	Yes

More, e.g., *Does the diversity of arithmetic operators matter? (Yes) Does incorrect reasoning label matter? (No* if IID *Yes* if OOD*)*

But WHY?

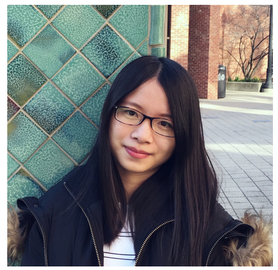
An Investigation of Neuron Activation as a Unified Lens to Explain Chain-of-Thought Eliciting Arithmetic Reasoning of LLMs

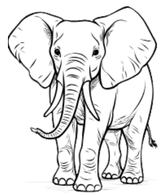
Daking Rai, Ziyu Yao
George Mason University

Press Cover:

MIT 科技评论 Technology Review

Daking Rai
(4th-yr PhD
at GMU CS)





ACL 2024
Bangkok, Thailand

Hypothesis

- Maybe there's a "switch" deciding an LM's reasoning performance, and these CoT variants end up tuning the switch differently?

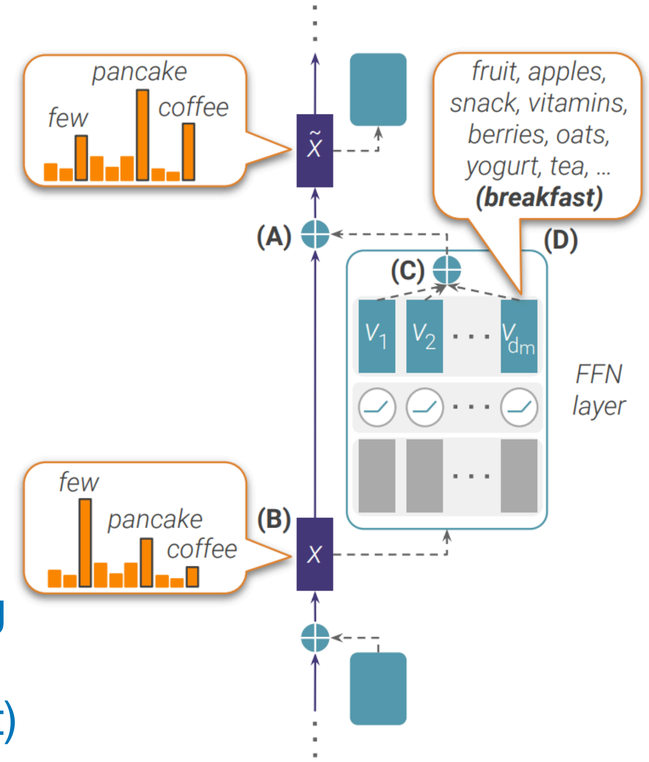
- Switch = critical **neurons** contributing to the model's arithmetic reasoning
- Geva et al. (2022): concept neurons in Feed-Forward (FF) layers of the transformer

$$FF^l(x_i^l) = f(K^l x_i^l)V^l$$

$$= \sum_{j=1}^{d_m} f(x_i^l \cdot k_j^l)v_j^l = \sum_{j=1}^{d_m} m_{ij}^l v_j^l$$

Activation coefficient (input dependent)

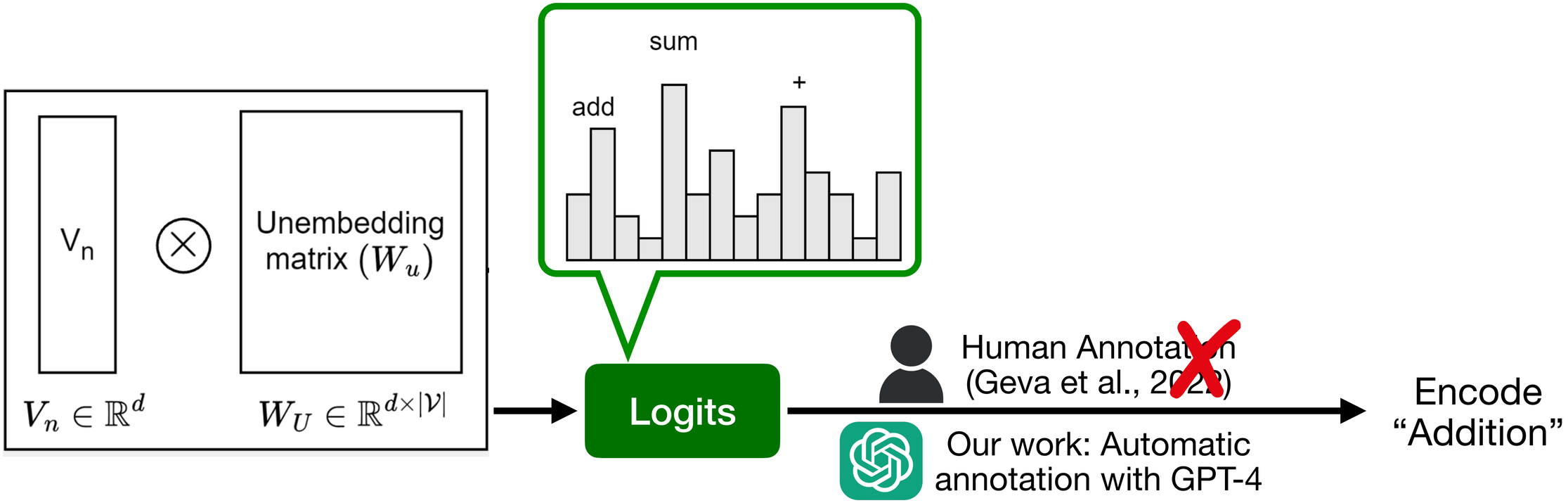
Neurons encoding critical concepts (input independent)



Knowledge neurons (Dai et al., 2021), skill neurons (Wang et al., 2022b), sentiment neurons (Radford et al., 2017), universal neurons (Gurnee et al., 2024) etc.

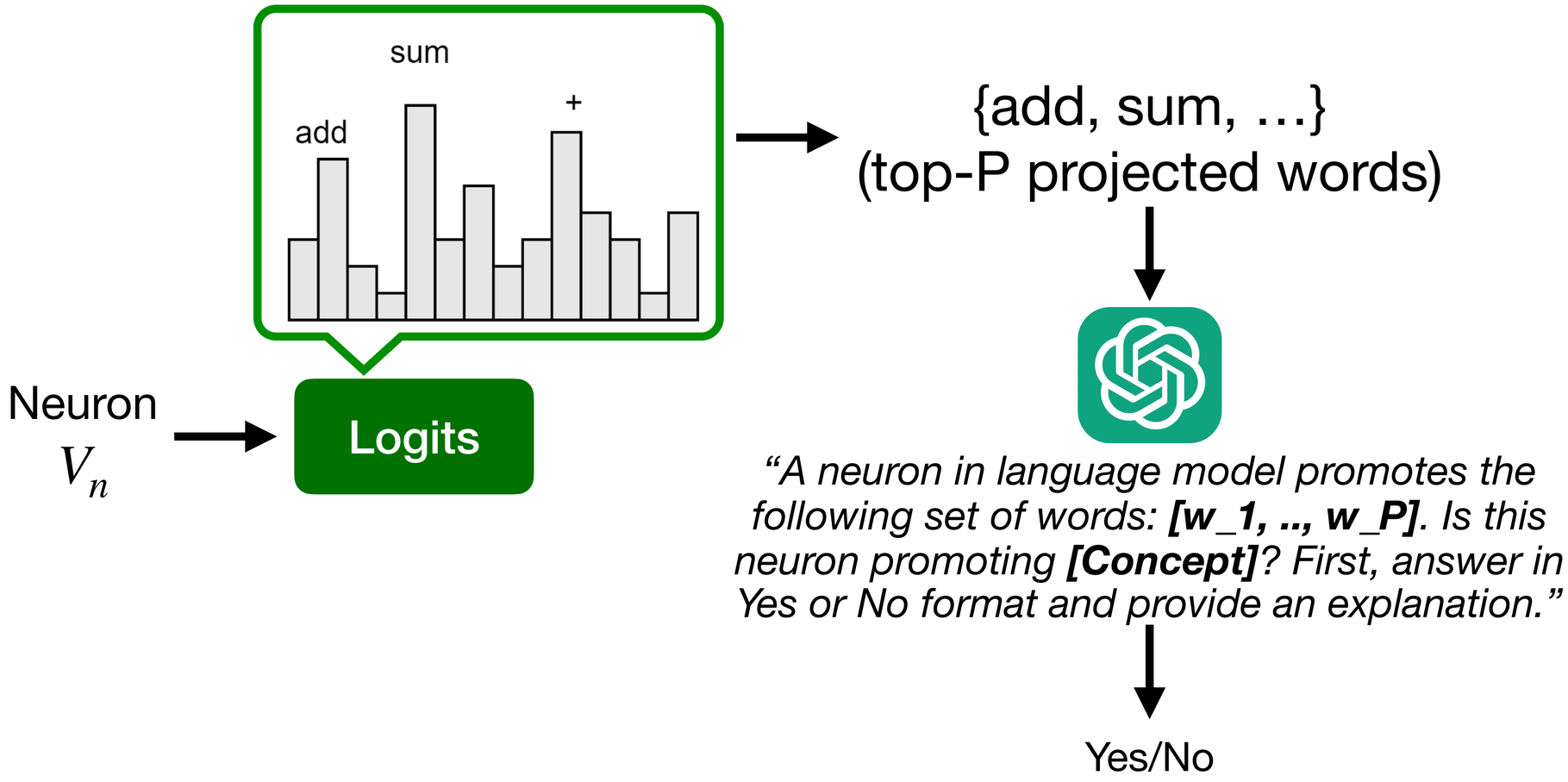
RQ1: Are there neurons related to the concept of “arithmetic reasoning”?

- How do we know what concept(s) each neuron has encoded?
- Approach: Logits lens (nostalgebraist 2020)



RQ1: Are there neurons related to the concept of “arithmetic reasoning”?

Concept
Logical Connectors ($C_{\texttt{logic}}$)
Addition ($C_{\texttt{add}}$)
Subtraction ($C_{\texttt{sub}}$)
Multiplication ($C_{\texttt{mul}}$)
Division ($C_{\texttt{div}}$)
Equals to ($C_{\texttt{eq}}$)
Calculation ($C_{\texttt{cal}}$)



RQ1: Are there neurons related to the concept of "arithmetic reasoning"?

- Llama-2 7B (Touvron et al., 2023)

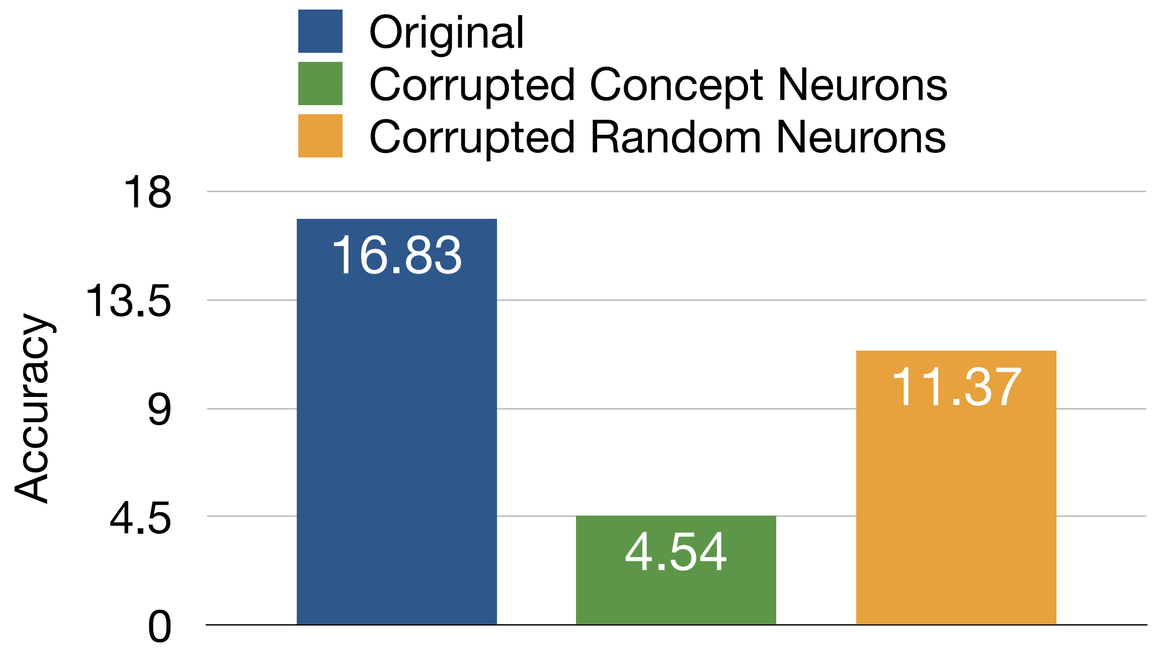
A total of 113 neurons (V_n's)!

Concept	Seed Tokens	Expanded Tokens	#of Neurons	Exemplar Neurons
Logical Connectors ($C_{\texttt{logic}}$)	{first, so, meaning, therefore, then, next, hence }	{logic, implies, thus, however, accordingly, subsequently, later, corresponding, etc. }	65	L10N9818{then, THEN, Then, then, ..}, L11N3000{therefore, Therefore, accordingly, donc,..}, L11N7742, L12N1030
Addition ($C_{\texttt{add}}$)	{add, addition, +, sum, plus }	{added, U+002B, adding, ++, increment, total, etc.}	18	L12N4814{added ,addition ,add,..}, L21N7027{+ ,add ,U+4e0e, U+306,..}, L27N10751{+, plus, -, minus,..}
Subtraction ($C_{\texttt{sub}}$)	{subtract, -, minus, sub }	{ -=, negative, U+2212, etc. }	2	L19N7900{-= ,- ,minus,2̌212, ..}, L25N5227
Multiplication ($C_{\texttt{mul}}$)	{multiply, product, times, mult, *, x }	{ multip, multi, U+00D7, double, twice, triple, fold, larger, etc. }	5	L16N10193{multip, double, multip, multiply, ..}, L18N4462, L20N6554, L22N1345, L22N1236
Division ($C_{\texttt{div}}$)	{divide, division, div, /, % }	{ div, divided, divisions, U+00F7, partition, partitions, etc. }	2	L20N10457{div ,divided ,division ,U+00F7,.. }, L26N1378{div, Div, div, Div, division,.. }
Equals to ($C_{\texttt{eq}}$)	{ =, total, equals, equal, equivalent }	{ equality, identical, same, exactly, contain, exact, etc. }	6	L14N7597{identical, difference, differences, equal,..}, L18N7531, L18N1850, L20N3177, L20N5535, L24N154
Calculation ($C_{\texttt{cal}}$)	{formula, equation, calculation, algorithm, expression, computation }	{rewrite, sum, application, ratio, percentage, eqn, rate, etc. }	14	L11N815{equation, formula, Formula, diagram,..}, L7N7176, L8N3689, L13N2019, L15N3958

RQ2: Are the discovered neurons important for eliciting the reasoning capability of LLMs?

- Evaluating the *faithfulness* of the discovered neurons
- If these neurons are critical, then corrupting them should yield substantial loss in LM's reasoning performance

$$FF^l(x_i^l) = \sum_{j=1}^{d_m} m_{ij}^l (v_j^l + Noise)$$

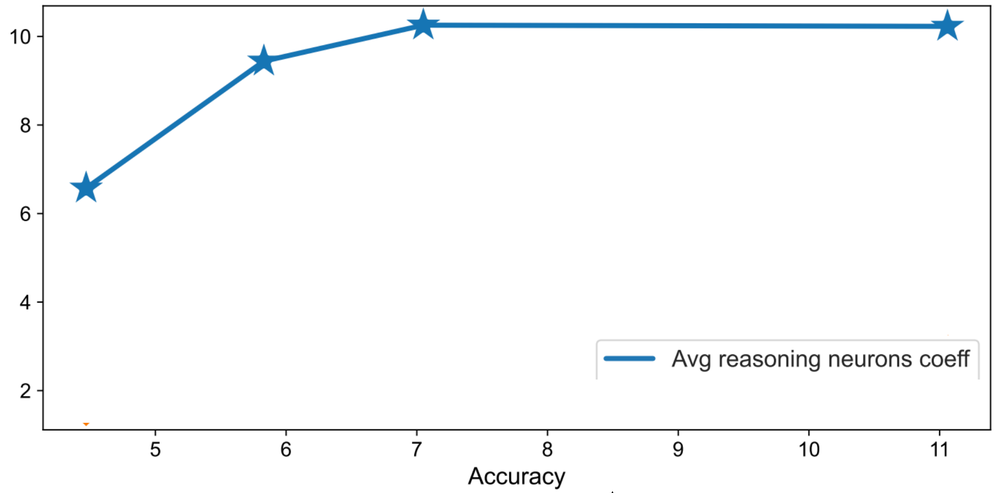


Using The Activation of Reasoning Neurons as Explanations

- Now that we know that these neurons are critical, can they be used to explain an LM's different performances under various zero-shot prompts?

Yes!

Activation coefficient m_{ij}^{l}



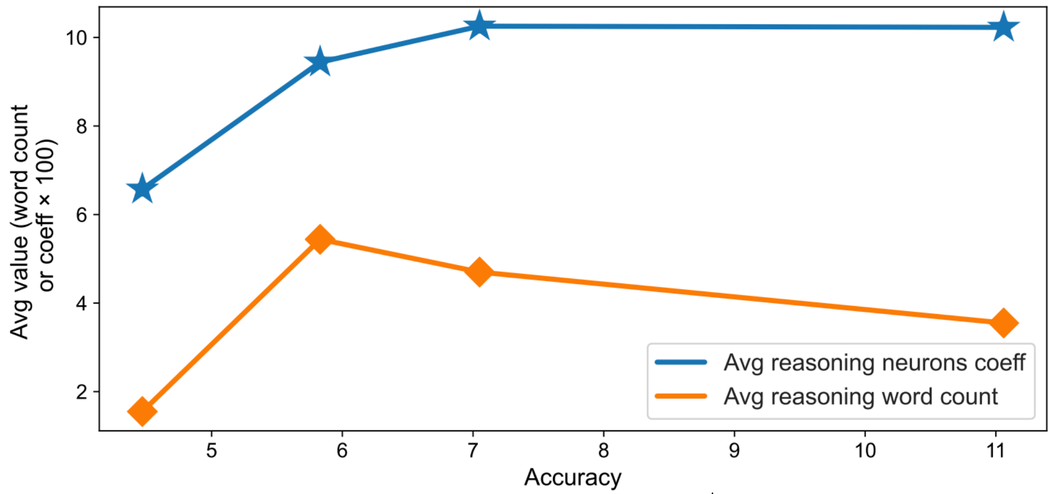
Zero-Shot CoT Prompt	Accuracy (%)
Let's think step by step	7.05
Take a deep breath and work on this problem step-by-step	4.47
Break this down	11.06
A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem	5.83

Performance of Llama2-7b on GSM8k; Prompts from Yang et al. (2023)

Using The Activation of Reasoning Neurons as Explanations

- Now that we know that these neurons are critical, can they be used to explain an LM's different performances under various zero-shot prompts?

Yes! (Cannot be trivially measured at the word level)



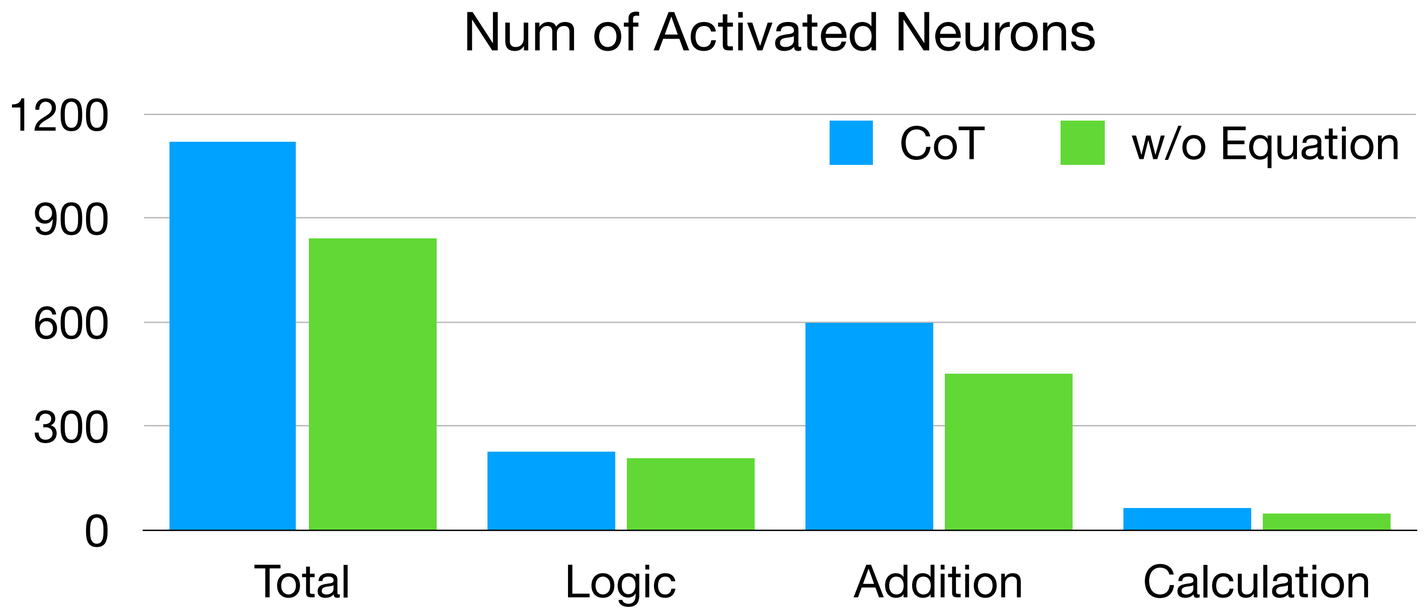
Zero-Shot CoT Prompt	Accuracy (%)
Let's think step by step	7.05
Take a deep breath and work on this problem step-by-step	4.47
Break this down	11.06
A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem	5.83

Performance of Llama2-7b on GSM8k; Prompts from (Yang et al., 2023)

RQ3: Why does equation matter?

Research Questions	Examples in CoT Prompts	Prior Work	Findings
Does equation matter? (RQ3)	**w Equation:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been **21 - 15 = 6** trees. The answer is 6. **w/o Equation:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been **6** trees. The answer is 6.	Wang et al. (2023); Ye et al. (2023); Madaan and Yazdanbakhsh (2022)	Yes

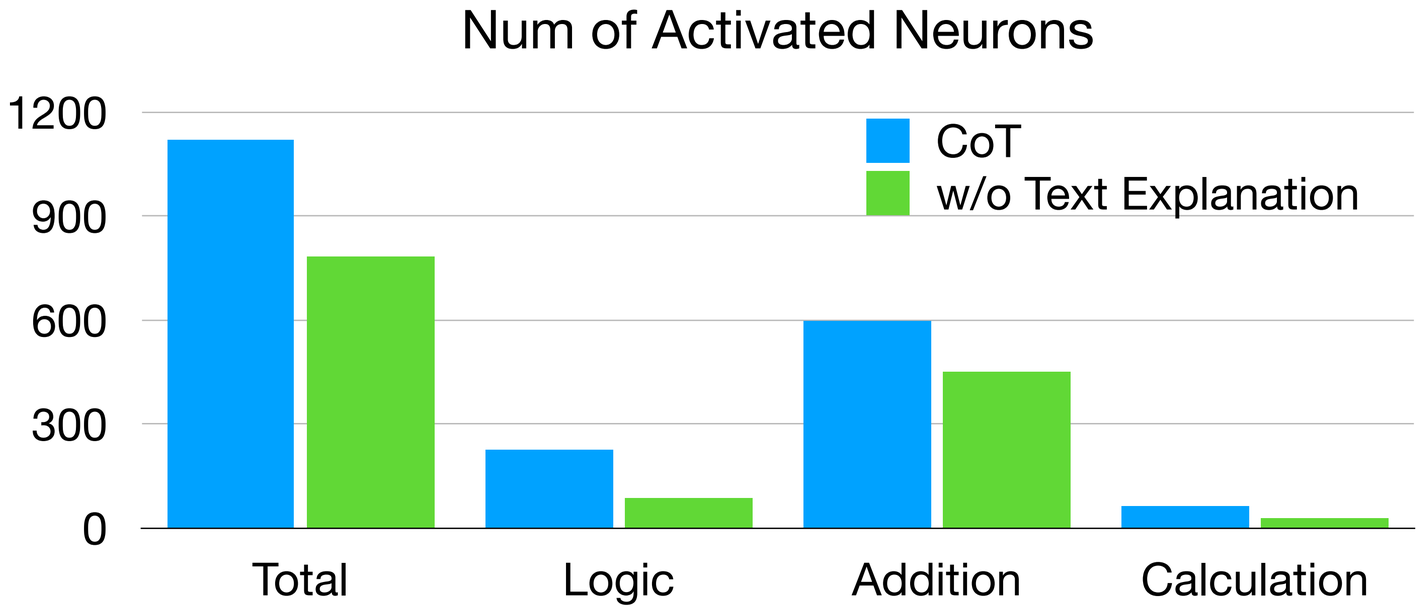
Prompts	Accuracy
CoT (w/ equation)	16.83%
w/o equation	12.58%



RQ4: Why does text explanation matter?

Research Questions	Examples in CoT Prompts	Prior Work	Findings
Does textual explanation matter? (RQ4)	**w Textual Explanation:** Let's think step by step. First Leah had 32 chocolates and her sister had 42 chocolates. So in total they had 32 + 42 = 74 chocolates. Then they ate 35 chocolates. So there must be 74 - 35 = 39 chocolates. The answer is 39. **w/o Textual Explanation:** 32 + 42 = 74. 74 - 35 = 39. The answer is 39.	Wang et al. (2023); Ye et al. (2023); Madaan and Yazdanbakhsh (2022)	Yes

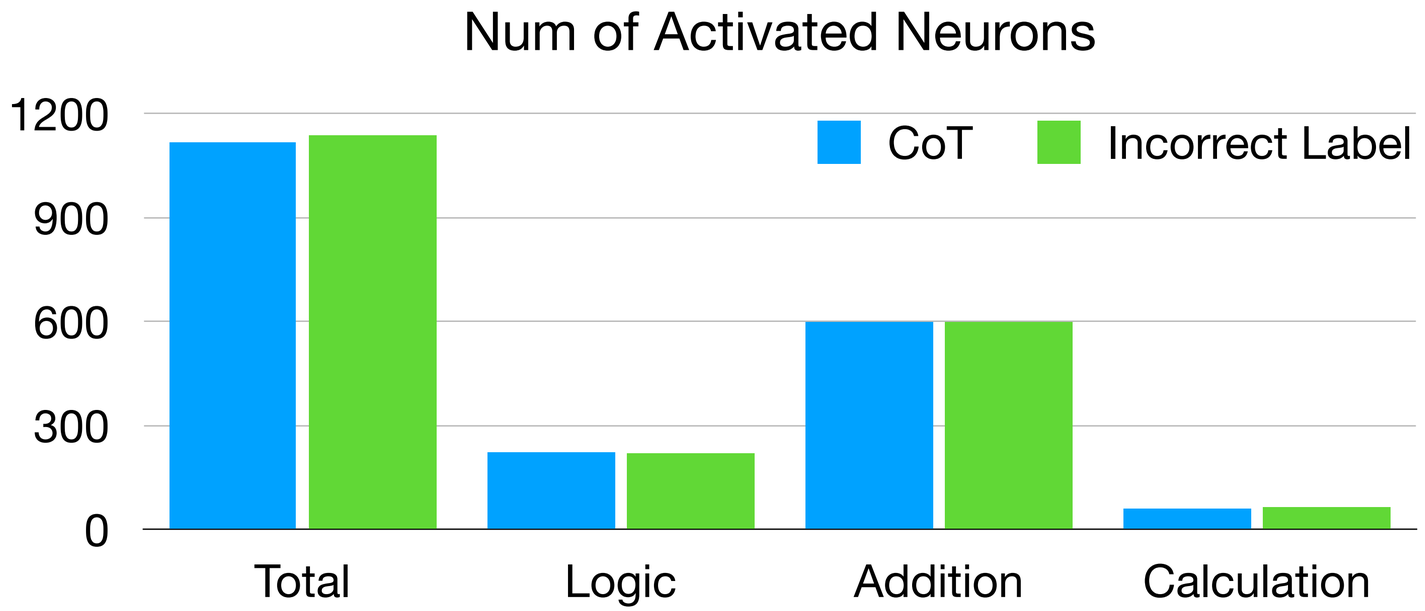
Prompts	Accuracy
CoT (w/ text explanation)	16.83%
w/o text explanation	13.41%



RQ6: Why does incorrect reasoning label *not* matter?

Research Questions	Examples in CoT Prompts	Prior Work	Findings
Does incorrect reasoning or gold label not matter? (RQ6)	**Correct Label:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6 trees. The answer is 6. **Incorrect Label:** Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = **1** trees. The answer is **1**.	Wang et al. (2023); Ye et al. (2023)	No

Prompts	Accuracy
CoT	16.83%
Incorrect label	16.45%



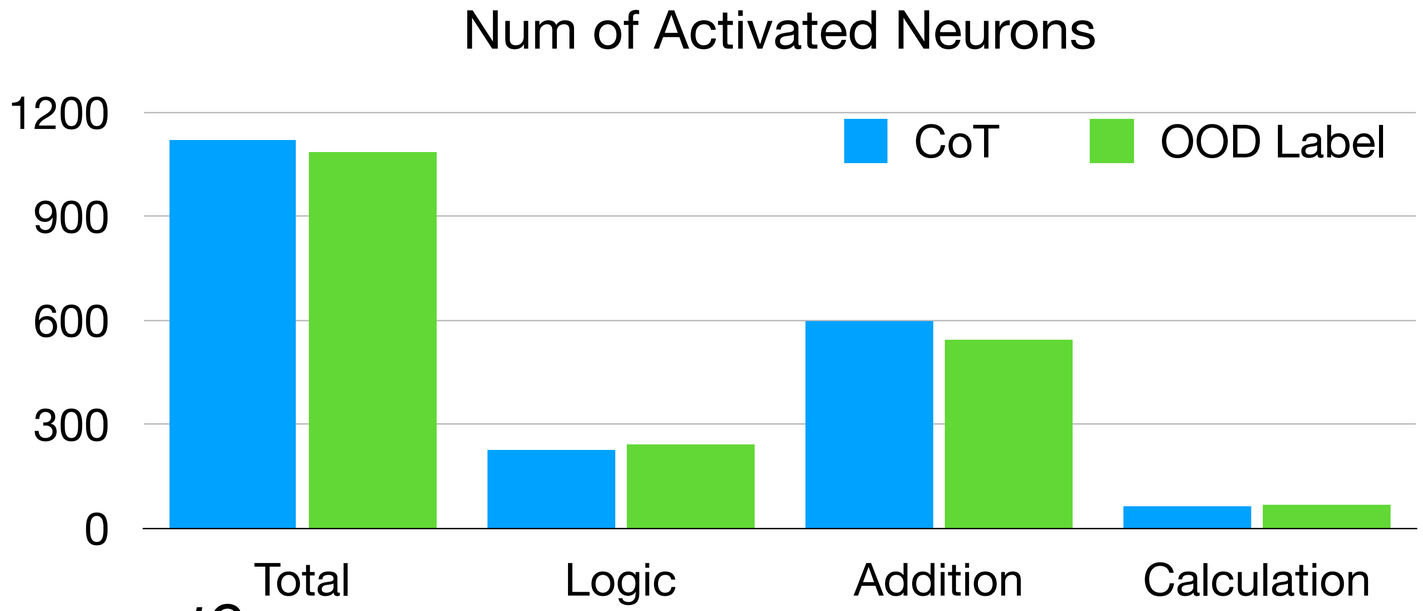
Does Neuron Activation Explain Everything?

- Unfortunately, it doesn't. LMs are way more complicated than just neuron activation.

- e.g., we cannot explain why OOD labels matter

OOD Label: Let's think step by step. First there are 15 trees. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = **Dawson** trees. The answer is **Dawson**.

Prompts	Accuracy
CoT	16.83%
OOD label	7.58%



Contradict to our faithfulness assessment?

Does Neuron Activation Explain Everything?

- Unfortunately, it doesn't. LMs are way more complicated than just neuron activation.
- **Necessity** vs. **Sufficiency**
  - Necessity: certain neurons are necessary to enable an LM's reasoning
  - Sufficiency: but only these neurons being activated is not sufficient
- **Lack of universality**: observations do not generalize to Llama3-8B
  - e.g., Llama3-8B is very sensitive to *any* noise to random neurons
- **Need other ways to discover the full picture of LM reasoning**

Mechanistic Interpretability

ICML International Conference On Machine Learning

Tutorial at ICML'25

- Reverse engineering the inner mechanism of a model
  - Features & Circuits (Olah et al., 2020)

https://arxiv.org/pdf/2407.02646

A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models

Language Models for Reasoning

- **Topic 1: *How*** do we unlock LMs' capabilities in performing complicated reasoning tasks?
  - Using event extraction as an exemplar task
  - How to instruct or prompt an LM for more effective event extraction?

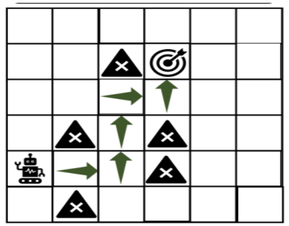
- **Topic 2: *Why*** an LM can or cannot perform reasoning?
  - Using arithmetic reasoning as an exemplar task
  - Mechanistic understanding of how Chain-of-Thought (CoT) prompting elicits an LM's arithmetic reasoning capability

Other Types of Reasoning

- e.g., spatial reasoning

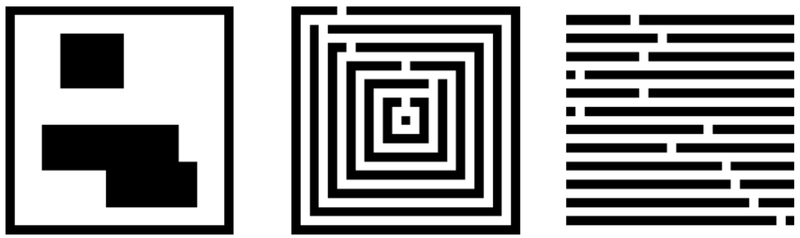
CAN LARGE LANGUAGE MODELS BE GOOD PATH PLANNERS? A BENCHMARK AND INVESTIGATION ON SPATIAL-TEMPORAL REASONING

Mohamed Aghzal, Erion Plaku, Ziyu Yao
Department of Computer Science
George Mason University
Fairfax, VA, 22030, USA
{maghzal, plaku, ziyuyao}@gmu.edu



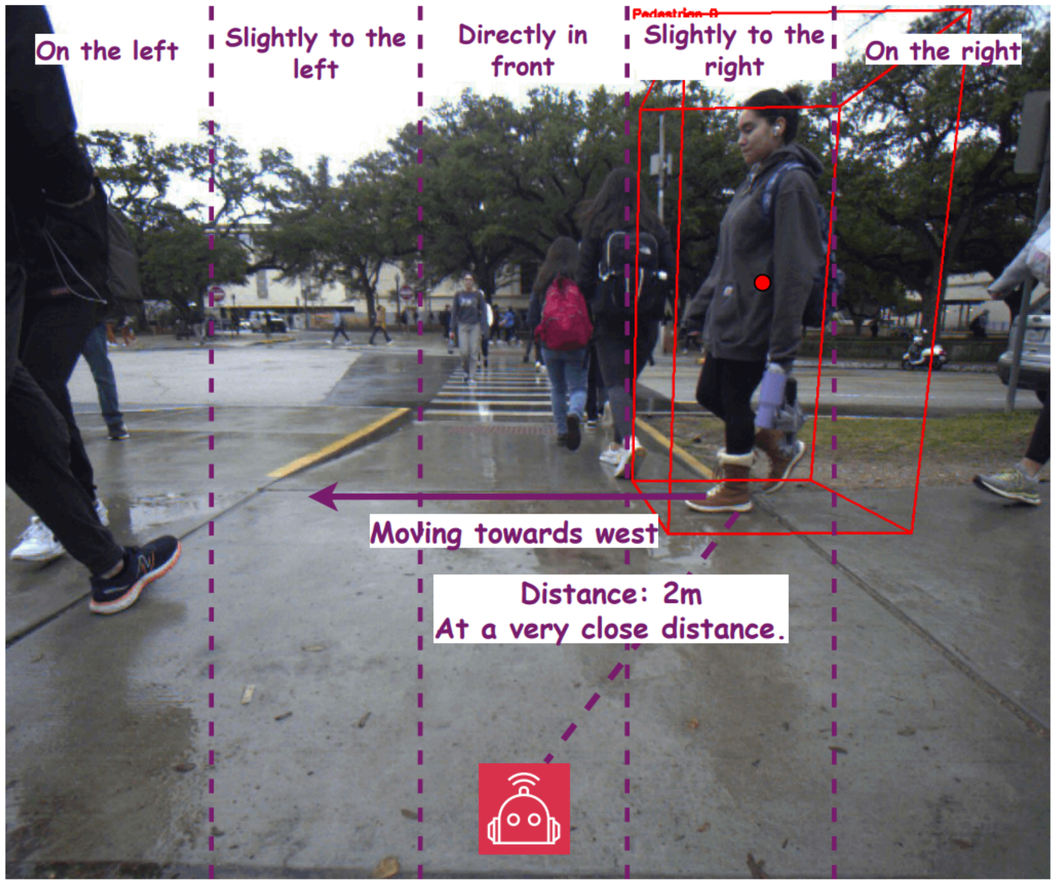
Look Further Ahead: Testing the Limits of GPT-4 in Path Planning

Mohamed Aghzal[1] Erion Plaku[2] Ziyu Yao[1]



AutoSpatial: Visual-Language Reasoning for Social Robot Navigation through Efficient Spatial Reasoning Learning

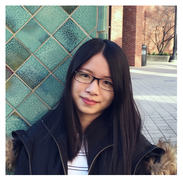
Yangzhe Kong[1], Daeun Song[1], Jing Liang[2], Dinesh Manocha[2], Ziyu Yao[1], and Xuesu Xiao[1]



Reasoning Agents for Education

- LLMs for Math Education
  - LLMs good at Math Reasoning -> role-playing students solving math problems
- MathVC: a platform for collaborative math problem solving and math education

Murong Yue
(3rd Yr CS PhD
at GMU)



Now let's start a group discussion!
Alice: " So, we need to figure out the right amount of each soup flavor and type of bread roll Martha should buy based on the survey results to avoid leftovers and maximize profit."
Bob: " Right, so we need to work out the best quantities of soup and bread rolls to purchase. This means calculating the preferred soup flavors and bread types from the survey, determining the servings and bottles needed for each soup flavor, the packs of bread rolls required, and then figuring out the costs, revenue, and profit."
Charlie: " I can take on calculating the quantities and costs for the soups based on the survey results. Does anyone want to handle the bread rolls, or should we split that task too?"
Alice: " I can handle the calculations for the bread rolls, and we can review each other's findings to ensure accuracy. Does that sound good to everyone?"
Bob: " Let's start the problem solving by calculating how popular each soup flavor is according to the survey responses, which will help us determine how much of each flavor to prepare."
Alice: "Based on the survey, 15% prefer carrot and coriander, 40% like tomato, 25% choose leek and potato, and 20% go for chicken and vegetable, so for 500 customers, we'll need 75 servings of carrot and coriander, 200 of tomato, 125 of leek and potato, and 100 of chicken and vegetable, which means 8 bottles of carrot and coriander, 20 of tomato, 12.5 of leek and potato, and 10 of chicken and vegetable."
Bob: "Alice, we actually need to round up to 13 bottles for the leek and potato soup because we can't buy half a bottle, so we'll have enough for all the servings."
Alice: "Oh, you're right, Bob; we'll need to get 13 bottles for the leek and potato soup, and now I'll calculate how many packs of bread rolls we need based on the survey."
Bob: "Good catch on the soup bottles, Alice; now let's figure out the bread rolls—we'll need 38 packs of white bread and 13 packs of brown bread to match the survey preferences!"
Charlie: "Looks like we've got the right numbers for the bottles and bread packs, so let's add up the costs and see how much money Martha can make!"
Alice: "Okay, so if we add up the costs for the soup and bread rolls, we get a total cost of $357, and since we're selling 500 mugs with bread for $1.25 each, our total revenue will be $625, which means our profit will be $268!"
Charlie: "Great, the numbers look good, so let's start planning where to buy the soup and bread rolls and how we'll set up the stall!"
Alice: "Yep, the profit of $268 sounds awesome, so let's figure out where to buy all the soup and bread rolls and get our stall ready for the market!"

(Highlights of my group)

Research Landscape about Language Models (LMs)

- Pursuit of machine intelligence
  - Revisiting Prompt Optimization with Large Reasoning Models---A Case Study on Event Extraction, Preprint'25
  - Instruction-Tuning LLMs for Event Extraction with Annotation Guidelines, Preprint'25
  - AutoSpatial: Visual-Language Reasoning for Social Robot Navigation through Efficient Spatial Reasoning Learning, Preprint'25
  - DOTS: Learning to Reason Dynamically in LLMs via Optimal Reasoning Trajectories Search, ICLR'25
  - Large Language Model Cascades with Mixture of Thoughts Representations for Cost-efficient Reasoning, ICLR'24
  - Instances Need More Care: Rewriting Prompts for Instances with LLMs in the Loop Yields Better Zero-Shot Performance, ACL Findings'24
  - Evaluating Vision-Language Models as Evaluators in Path Planning, CVPR'25
  - Look Further Ahead: Testing the Limits of GPT-4 in Path Planning, IEEE CASE'24
  - Can Large Language Models be Good Path Planners? A Benchmark and Investigation on Spatial-temporal Reasoning, ICLR'24 Workshop

Reasoning

Planning

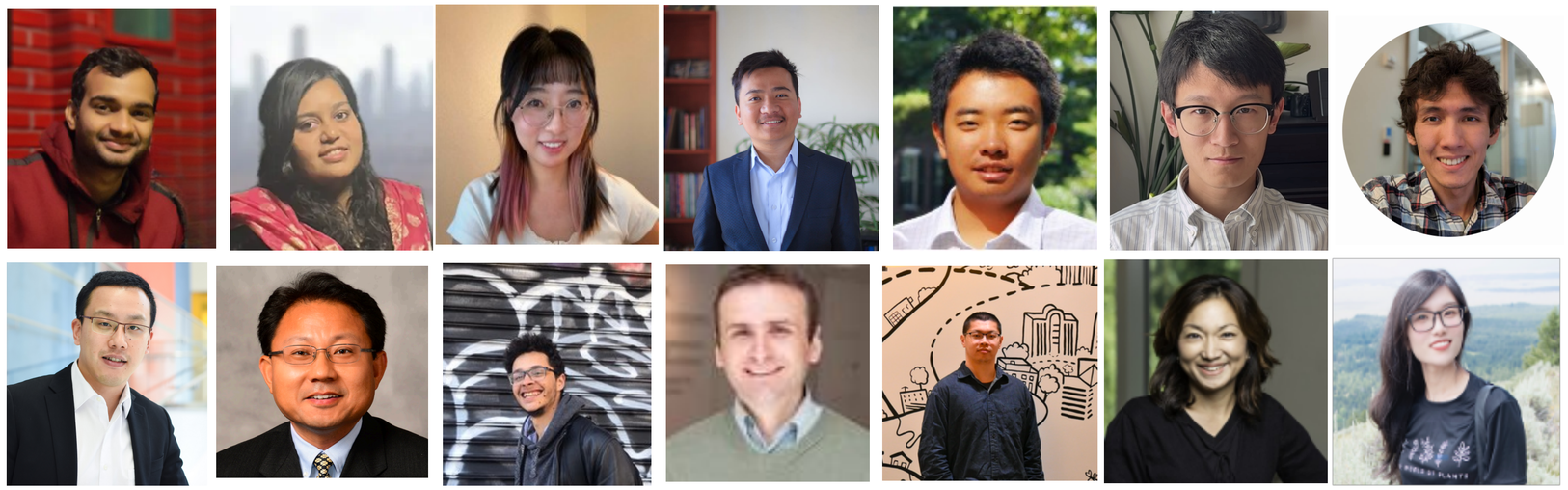
- Trustworthy language interface
  - Efficient but Vulnerable: Benchmarking and Defending LLM Batch Prompting Attack, Preprint'25
  - Beneath the Surface: How Large Language Models Reflect Hidden Bias, Preparing'25
  - An Investigation of Neuron Activation as a Unified Lens to Explain Chain-of-Thought Eliciting Arithmetic Reasoning of LLMs, ACL'24
  - Navigating the Shortcut Maze: A Comprehensive Analysis of Shortcut Learning in Text Classification by Language Models, EMNLP Findings'24

Interpretability, Human-AI Interaction, AI Safety

- Applications of LMs/LM agents to education, software engr, robotics, etc.

Thank You!

Email: ziyuyao@gmu.edu
Webpage: https://ziyuyao.org/
X/Twitter: @ZiyuYao



… and many others!