Cognitive Motivations in Analogical and Physical Reasoning
with Large Language Models

Shane Storks (he/him)

PhD Candidate in Computer Science and Engineering

Cognitive Science Seminar

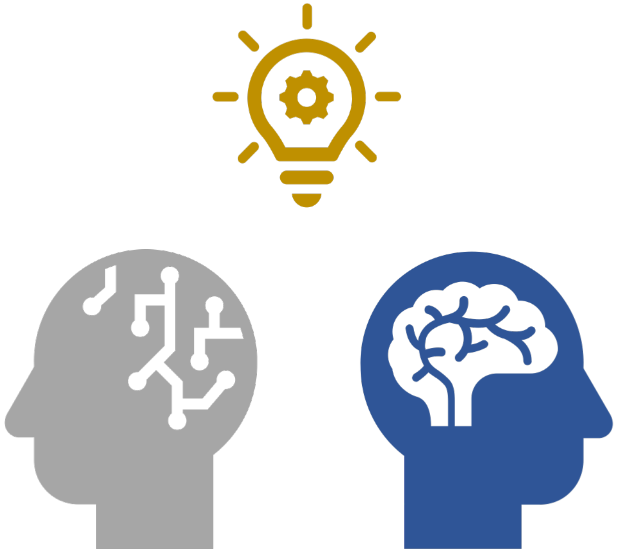
October 10, 2023

AI LAB

Introduction

- Large language models (LLMs) like ChatGPT and GPT-4 have recently attracted attention
- Impressive, seemingly human-like conversation and reasoning capabilities solve many problems for automated language processing
- Enable research on interesting questions:

1. **How can LLMs shed light on the nature of human language and reasoning?**
2. **How can human reasoning strategies empower LLMs to better capture how the world works?**



https://chat.openai.com/
OpenAI. GPT-4 Technical Report. arXiv: 2303.08774.

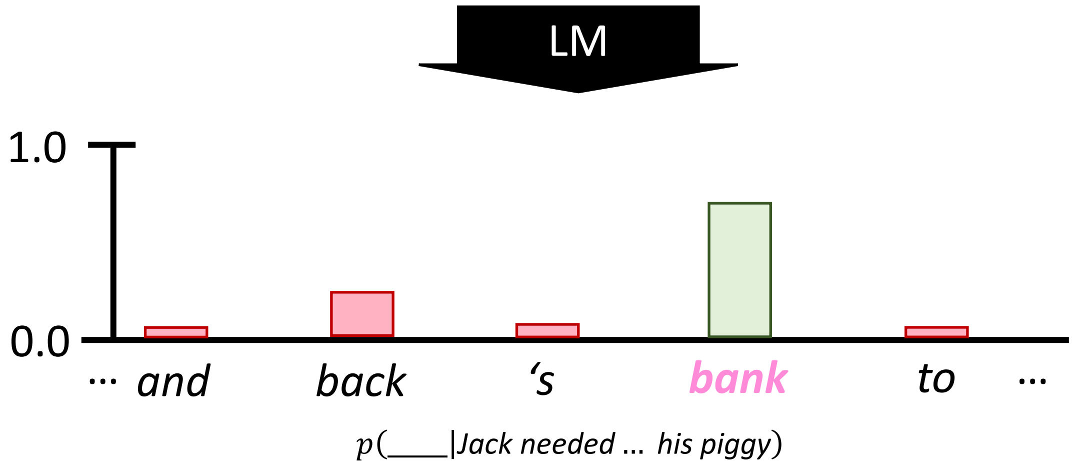
Outline

- **Language Model Basics**
- Application 1: Analogical Reasoning
- Application 2: Physical Commonsense Reasoning

Language Models

$$p(w_n | w_1, w_2, \ldots, w_{n-1})$$

*Jack needed some **money**, so he went and shook his **piggy** ____*

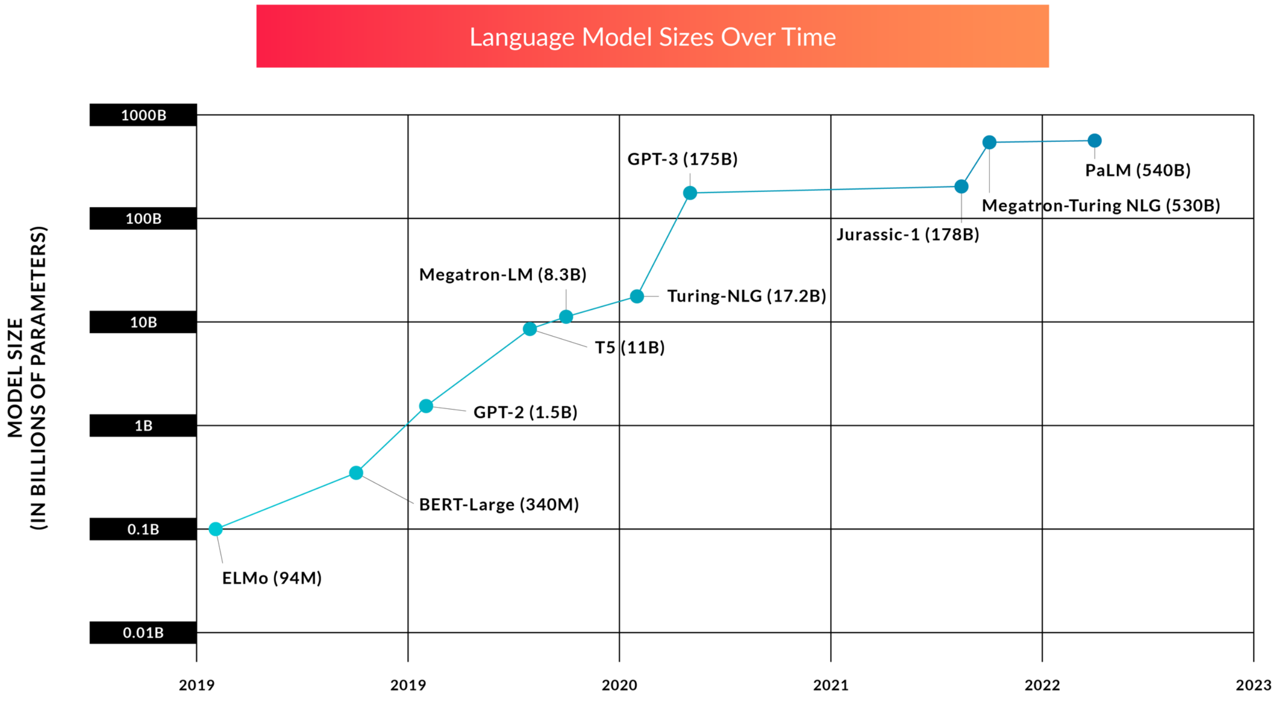


(dreamstime)

Minsky, M. (2000). Commonsense-based interfaces. In *Commun. ACM*, 43(8): p. 66-73.

Large Language Models

- What makes a language model a *large* language model?
- Recent trends:
  - More data
    - Web data
    - Human feedback annotation
  - More learned parameters
- Gives rise to new abilities…



(figure from Vinay Iyengar)

Prompting and In-Context Learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

task description

prompt

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

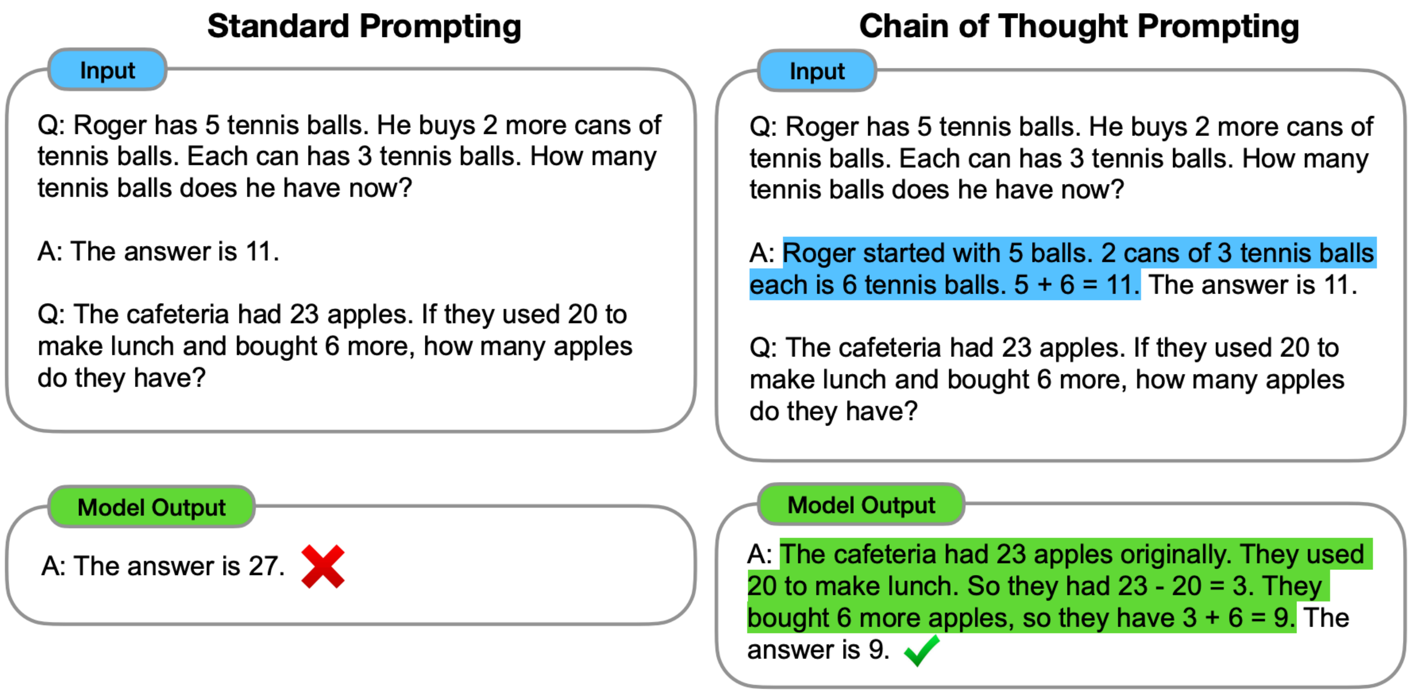
task description

examples

prompt

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. (2020). “Language Models are Few-Shot Learners.” arXiv: 2005.14165.

Chain-of-Thought Prompting



Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems 35.*

Outline

- Language Model Basics
- **Application 1: Analogical Reasoning**
- Application 2: Physical Commonsense Reasoning

In-Context Analogical Reasoning with Pre-Trained Language Models

Xiaoyang Hu[1,2] *, Shane Storks[1] *, Richard L. Lewis[2] †, Joyce Chai[1] †

[1] Computer Science & Engineering Division, University of Michigan
[2] Department of Psychology, University of Michigan
* Equal contribution † Equal advising

ACL 2023 Long Paper

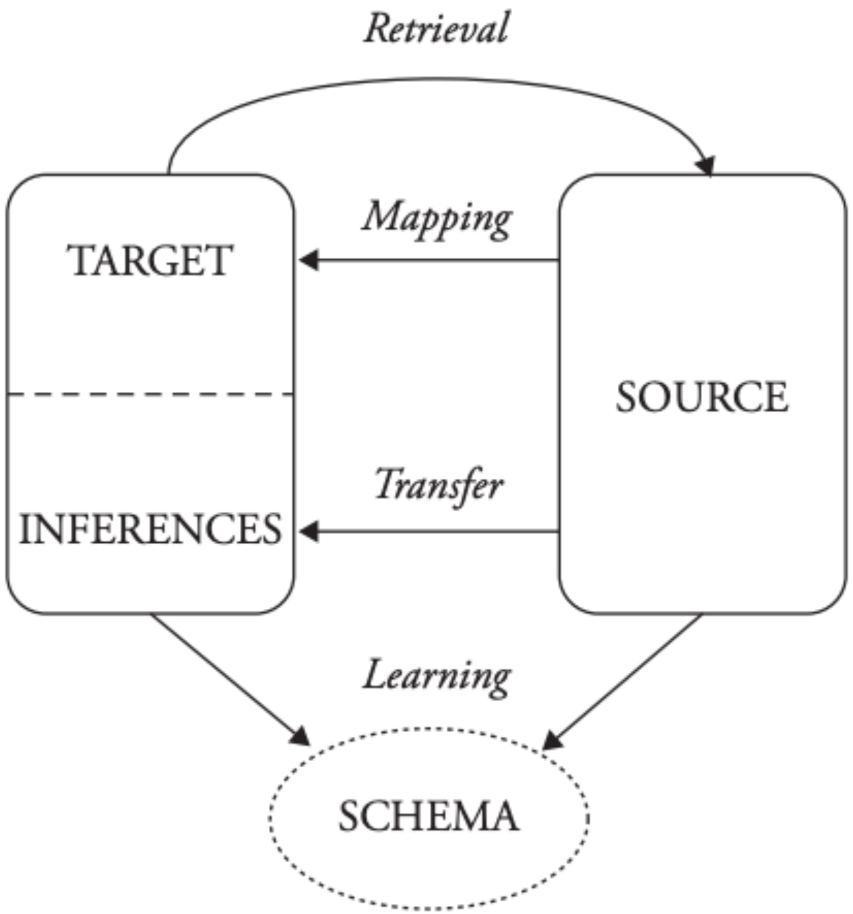
some slides made by Xiaoyang Hu

COMPUTER SCIENCE AND ENGINEERING UNIVERSITY OF MICHIGAN

LSA PSYCHOLOGY UNIVERSITY OF MICHIGAN

Motivation

- Making analogies is a fundamental capability of humans
- Enables us to tackle new situations based on past experience



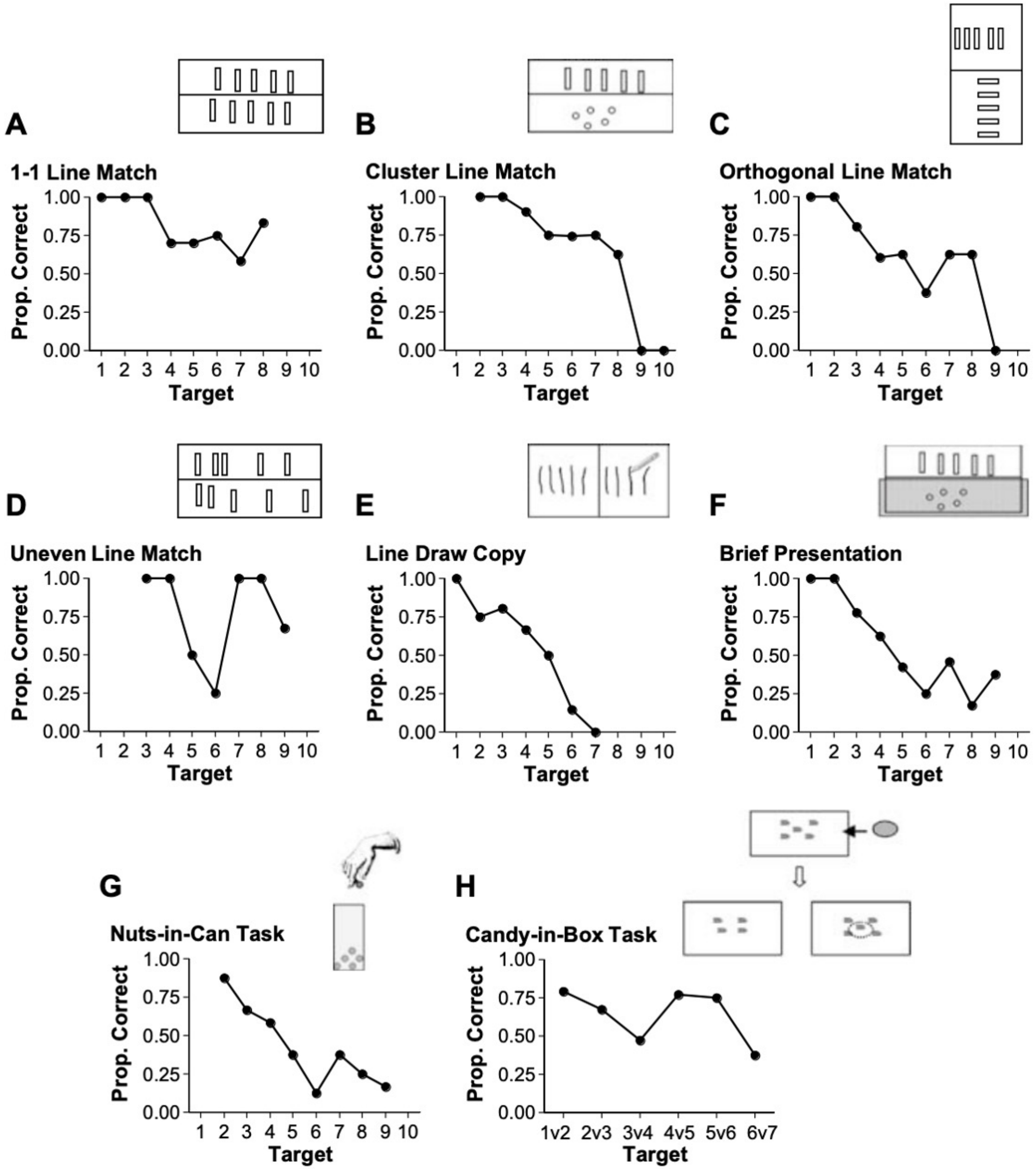
Dedre Gentner. 2010. Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34(5):752–775.
K.J. Holyoak. (2012). Analogy and relational reasoning. *The Oxford Handbook of Thinking and Reasoning* (pp.234-259). New York: Oxford University Press.

Motivation

- Work in cognitive science has found that language and analogy are connected in humans:

Motivation

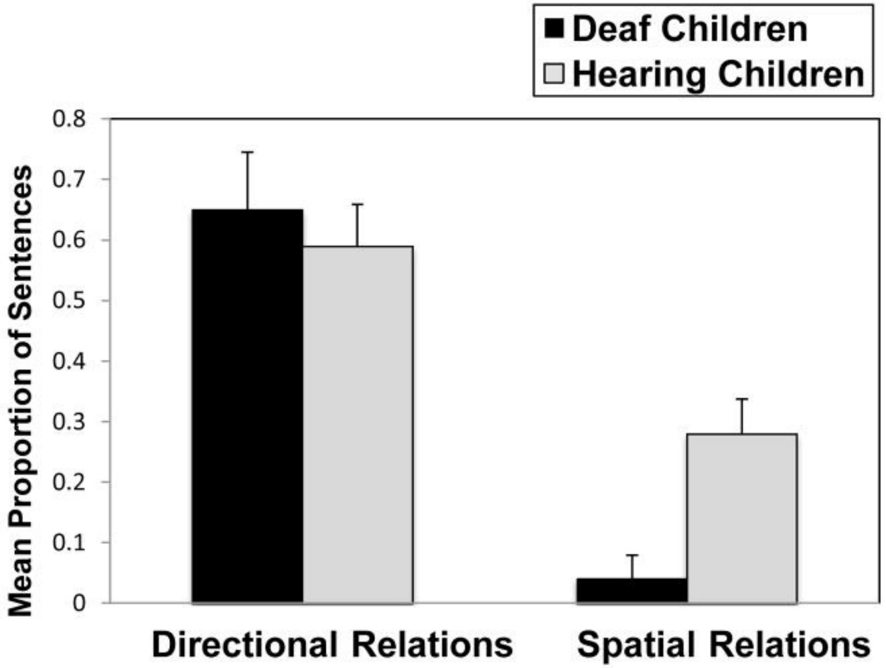
- Work in cognitive science has found that language and analogy are connected in humans:
    - Numerical language facilitates numerical analogies



Peter Gordon. 2004. Numerical cognition without words: Evidence from Amazonia. *Science*, 306(5695):496–499.

Motivation

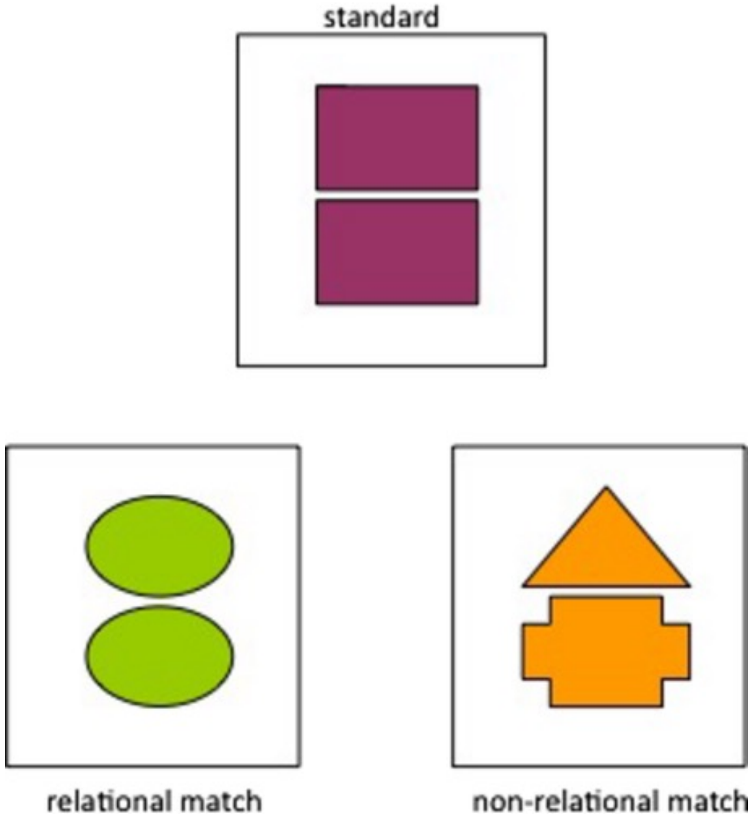
- Work in cognitive science has found that language and analogy are connected in humans:
  - Numerical language facilitates numerical analogies
  - Spatial language facilitates spatial analogies



Dedre Gentner, Asli Özyürek, Özge Gürcanli, and Susan Goldin-Meadow. 2013. Spatial language facilitates spatial cognition: Evidence from children who lack language input. *Cognition*, 127(3):318–330.

Motivation

- Work in cognitive science has found that language and analogy are connected in humans:
  - Numerical language facilitates numerical analogies
  - Spatial language facilitates spatial analogies
  - Names support analogy-making (even nonsense names)



Stella Christie and Dedre Gentner. 2014. Language helps children succeed on a classic analogy task. *Cognitive Science*, 38(2):383–397.

Motivation

- Analogy-making may be key to robust reasoning in AI systems
- Contemporary AI approaches for analogy-making require thousands of training examples to make any progress
- Meanwhile, LLMs can pick up new tasks through in-context learning with just a few relevant examples (more like humans)
  - Are they capable of analogy-making?

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

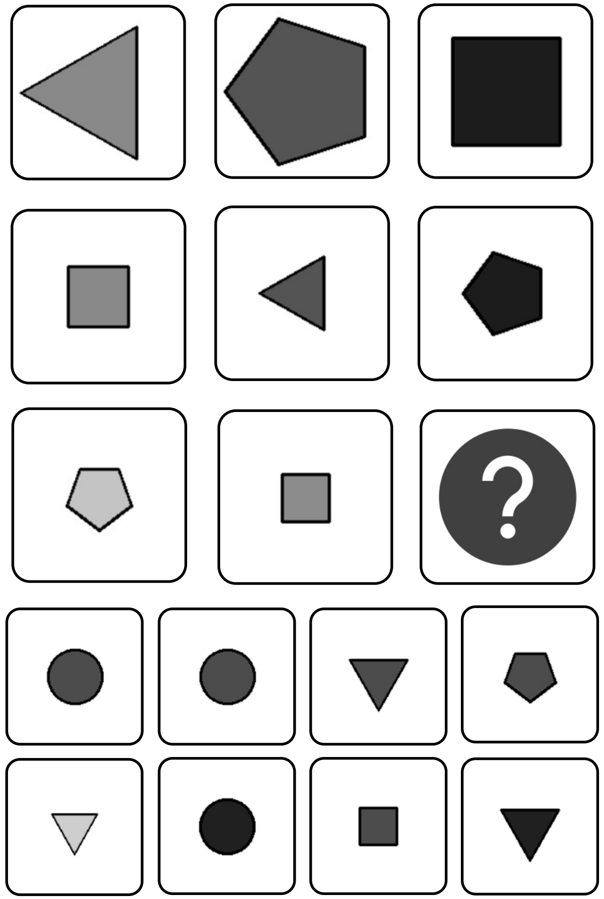
Questions

1. Does training LLMs on *natural language* give rise to the ability to form *abstract* analogies?
2. How do various factors contribute to analogy-making in LLMs?
   - Complexity of situations to make analogies from
   - Language-based abstractions (like names)
   - Complexity (size/# learned parameters) of LLM
   - In-context demonstration of task

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Raven’s Progressive Matrices (RPM)

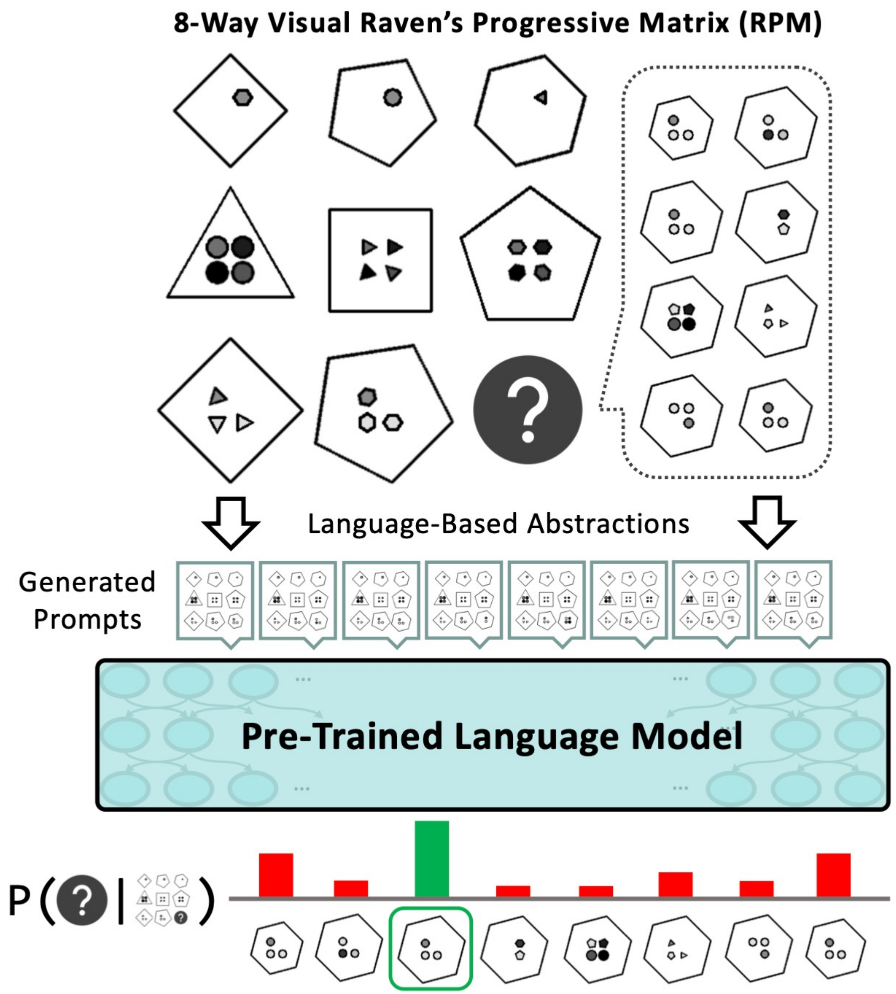
- A canonical test of analogical reasoning often used with human subjects
- Test-taker infers abstract rules from first 2 rows, then apply them to complete the third row
- RAVEN dataset
  - Relations:
    - Constant
    - Progression
    - Arithmetic
    - Distribute-Three



Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019a. RAVEN: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

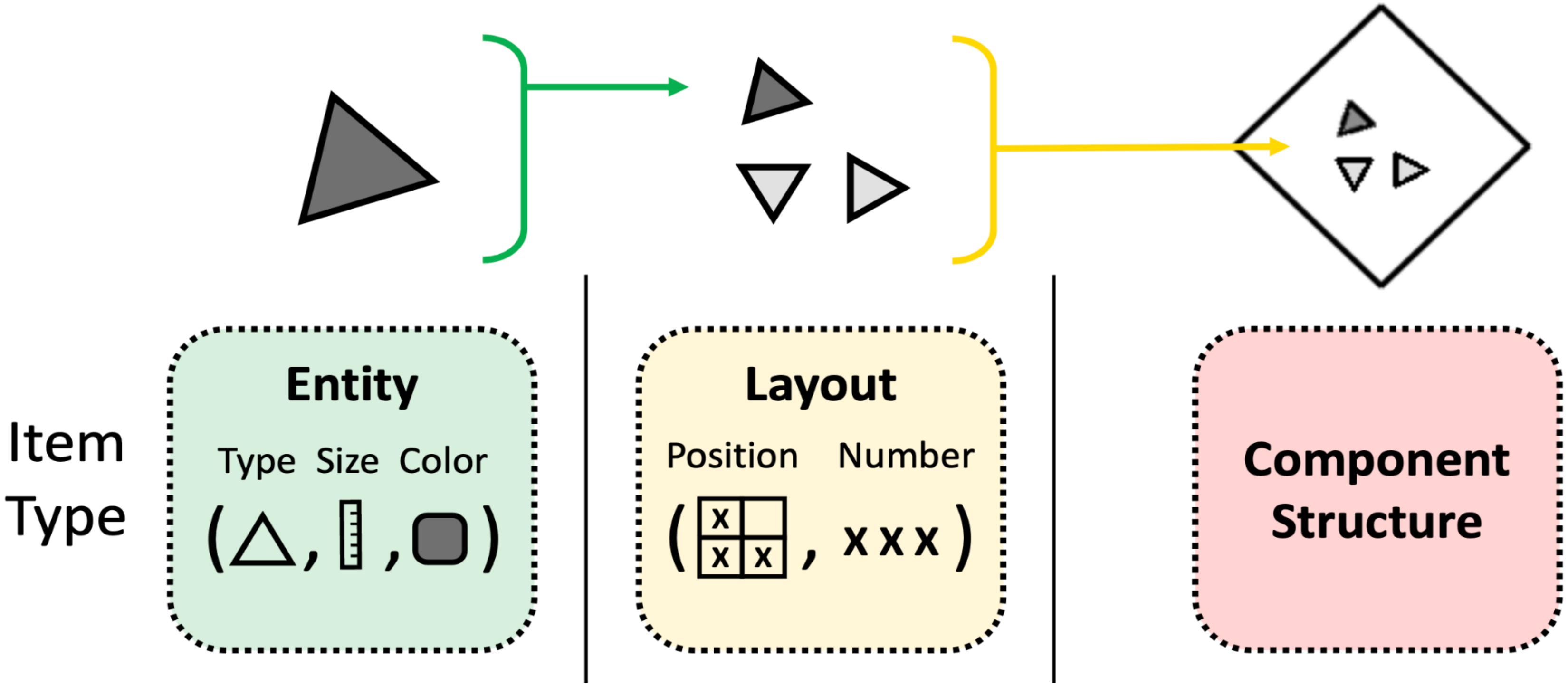
Prompting for Analogical Reasoning

- Created language abstractions for RPMs in RAVEN dataset
- Prompt LLMs to test abstract analogical reasoning capability
  - OPT & InstructGPT at varying model complexity

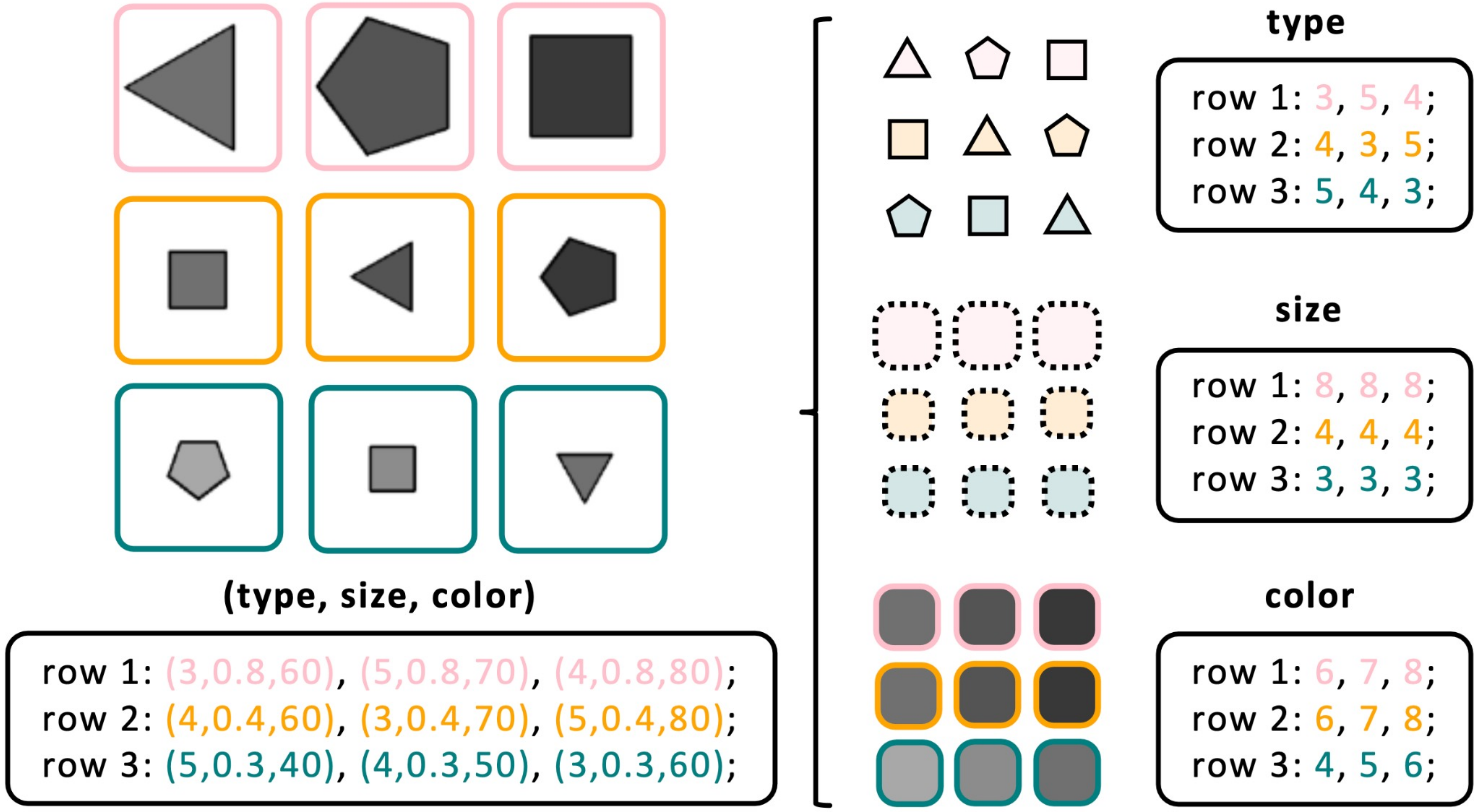


Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019a. RAVEN: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
Long Ouyang, Jeff Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. arXiv: 2203.02155.
Susan Zhang, Stephen Roller, Naman Goyal, et al. 2022. OPT: Open Pre-trained Transformer Language Models. arXiv: 2205.01068.

Components of RAVEN Matrix Items



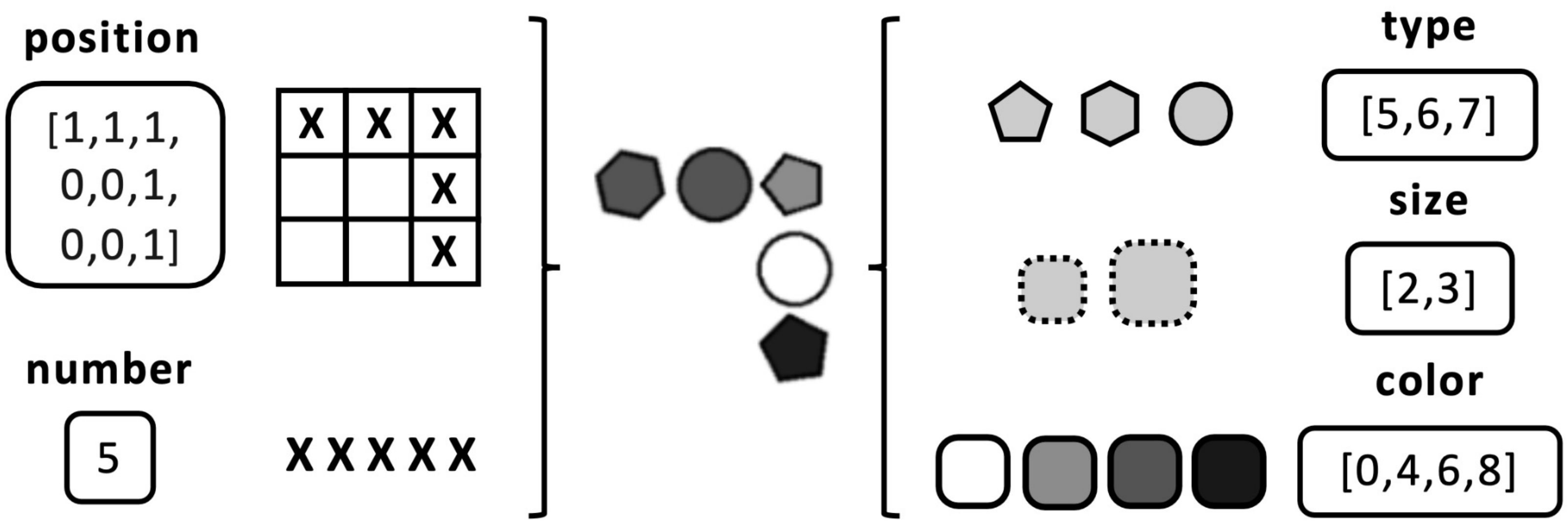
Entity-Level Abstractions



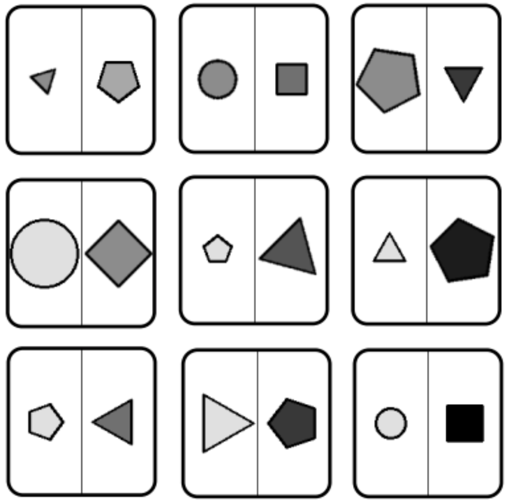
Decomposition

Naming

Layout-Level Abstractions



Component-Level Abstraction



L-R

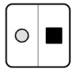
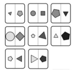
Complete

row 1: A (3,0.1,40) / B (5,0.3,30), A (7,0.2,40) / B (4,0.3,50), A (5,0.6,40) / B (3,0.3,70);
row 2: A (7,0.6,10) / B (4,0.6,40), A (5,0.1,10) / B (3,0.6,60), A (3,0.2,10) / B (5,0.6,80);
row 3: A (5,0.2,10) / B (3,0.4,50), A (3,0.6,10) / B (5,0.4,70), A (7,0.1,10) / B (4,0.4,90);

Left Comp.	**Right Comp.**
row 1: (3,0.1,40), (7,0.2,40), (5,0.6,40);	row 1: (5,0.3,30), (4,0.3,50), (3,0.3,70);
row 2: (7,0.6,10), (5,0.1,10), (3,0.2,10);	row 2: (4,0.6,40), (3,0.6,60), (5,0.6,80);
row 3: (5,0.2,10), (3,0.6,10), (7,0.1,10);	row 3: (3,0.4,50), (5,0.4,70), (4,0.4,90);

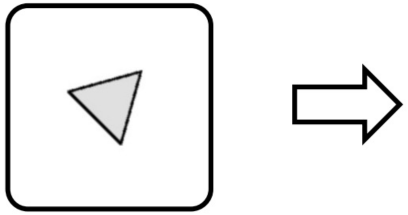
Type: Distr.	**Size: Distr.**	**Color: Const.**	**Type: Distr.**	**Size: Const.**	**Color: Prog.**
row 1: 3, 7, 5;	row 1: 1, 2, 6;	row 1: 4, 4, 4;	row 1: 5, 4, 3;	row 1: 3, 3, 3;	row 1: 3, 5, 7;
row 2: 7, 5, 3;	row 2: 6, 1, 2;	row 2: 1, 1, 1;	row 2: 4, 3, 5;	row 2: 6, 6, 6;	row 2: 4, 6, 8;
row 3: 5, 3, 7;	row 3: 2, 6, 1;	row 3: 1, 1, 1;	row 3: 3, 5, 4;	row 3: 4, 4, 4;	row 3: 5, 7, 9;

LLM LLM LLM LLM LLM LLM

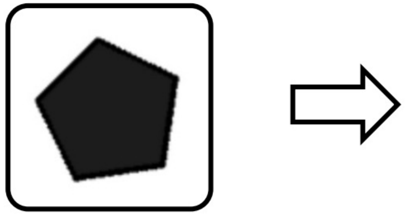
$p($  $|$  $) \propto p(\text{left type}) + p(\text{left size}) + p(\text{left color}) + p(\text{right type}) + p(\text{right size}) + p(\text{right color})$

Baselines

- How helpful are the naming abstractions we chose?
- 2 baselines for comparison:
    1. **Quasi-image**: lower-level "pixel-like" abstraction
    2. **Random naming:** choose random words to represent attributes, removing numerical dependencies between attribute names



```
[[ 2 . . . ],
 [ 2 2 . . ],
 [ 2 2 2 . ],
 [ 2 2 2 2 ],]
```

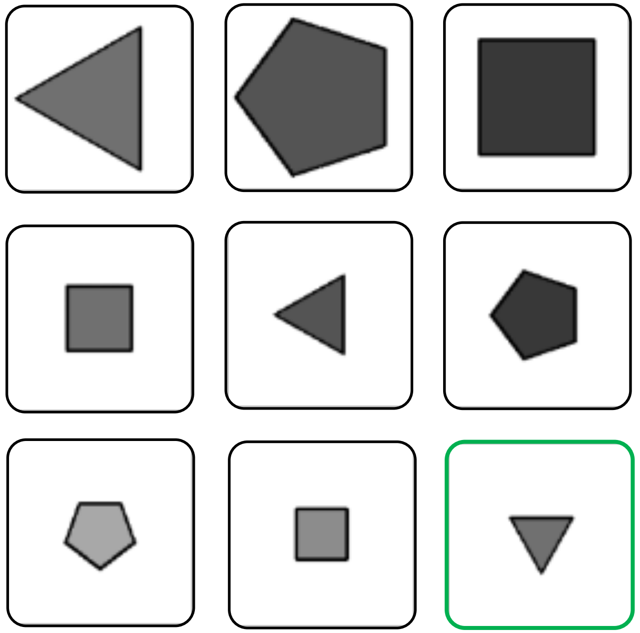


```
[[ . . . 9 . . . ],
 [ . . 9 9 9 . . ],
 [ . 9 9 9 9 9 . ],
 [ 9 9 9 9 9 9 9 ],
 [ 9 9 9 9 9 9 9 ],
 [ 9 9 9 9 9 9 9 ],
 [ 9 9 9 9 9 9 9 ],]
```


Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations (ICLR).*

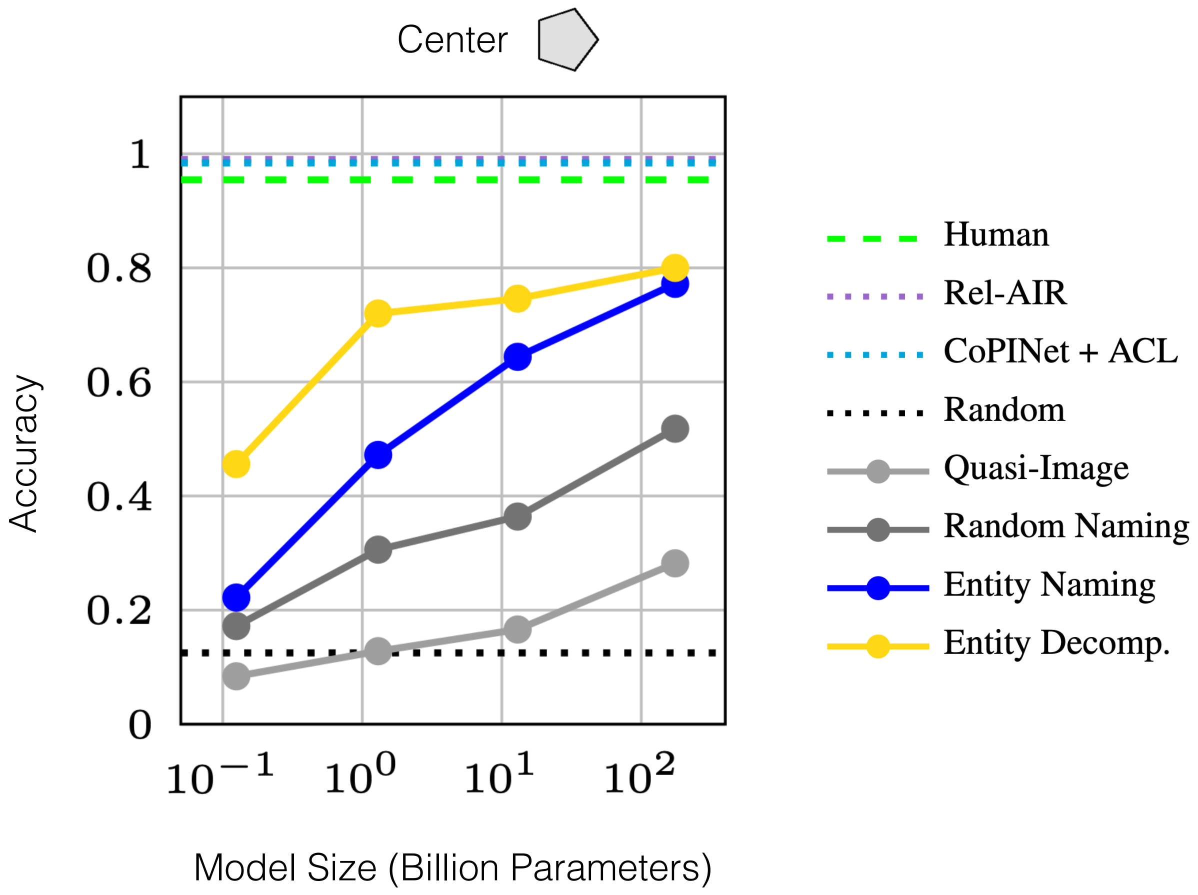
Single Entity Results

Center



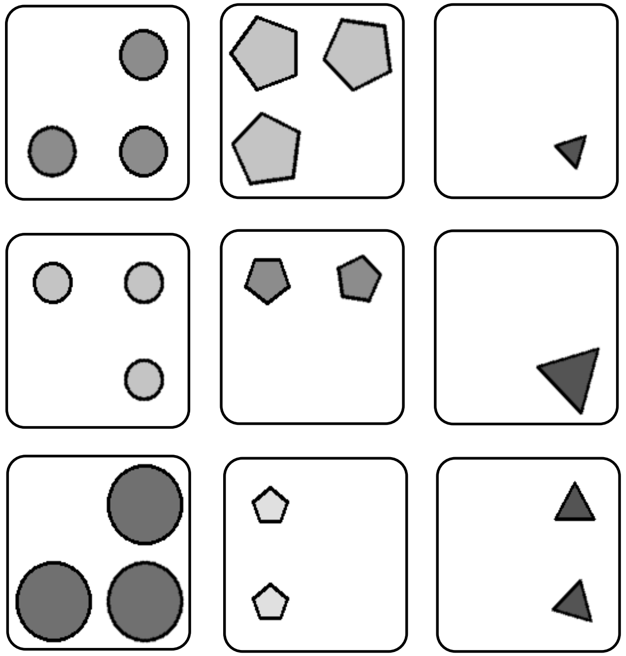
Single Entity Results

- <u>Analogies *do* arise from natural language training!</u>
- Bigger LLMs are better analogy-makers
- Numerical naming enables better analogy-making
- Decomposition abstractions especially help smaller LLMs
    - Model complexity ≈ working memory?

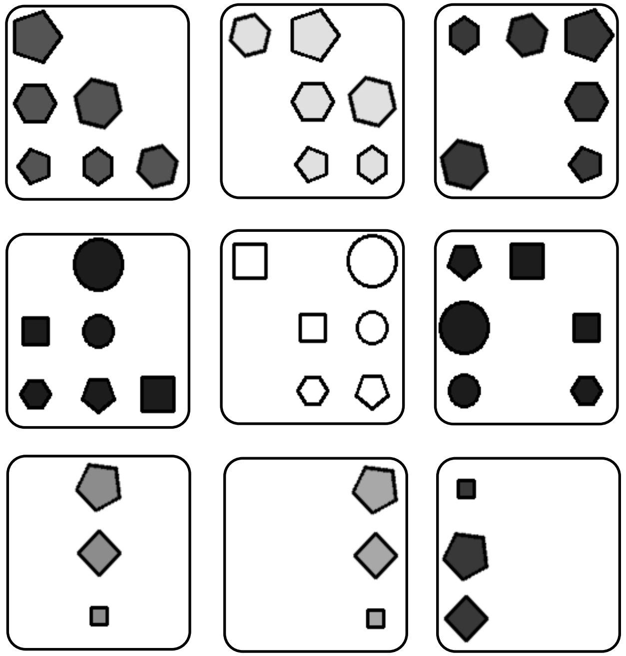


Multiple Entity Results

2x2Grid

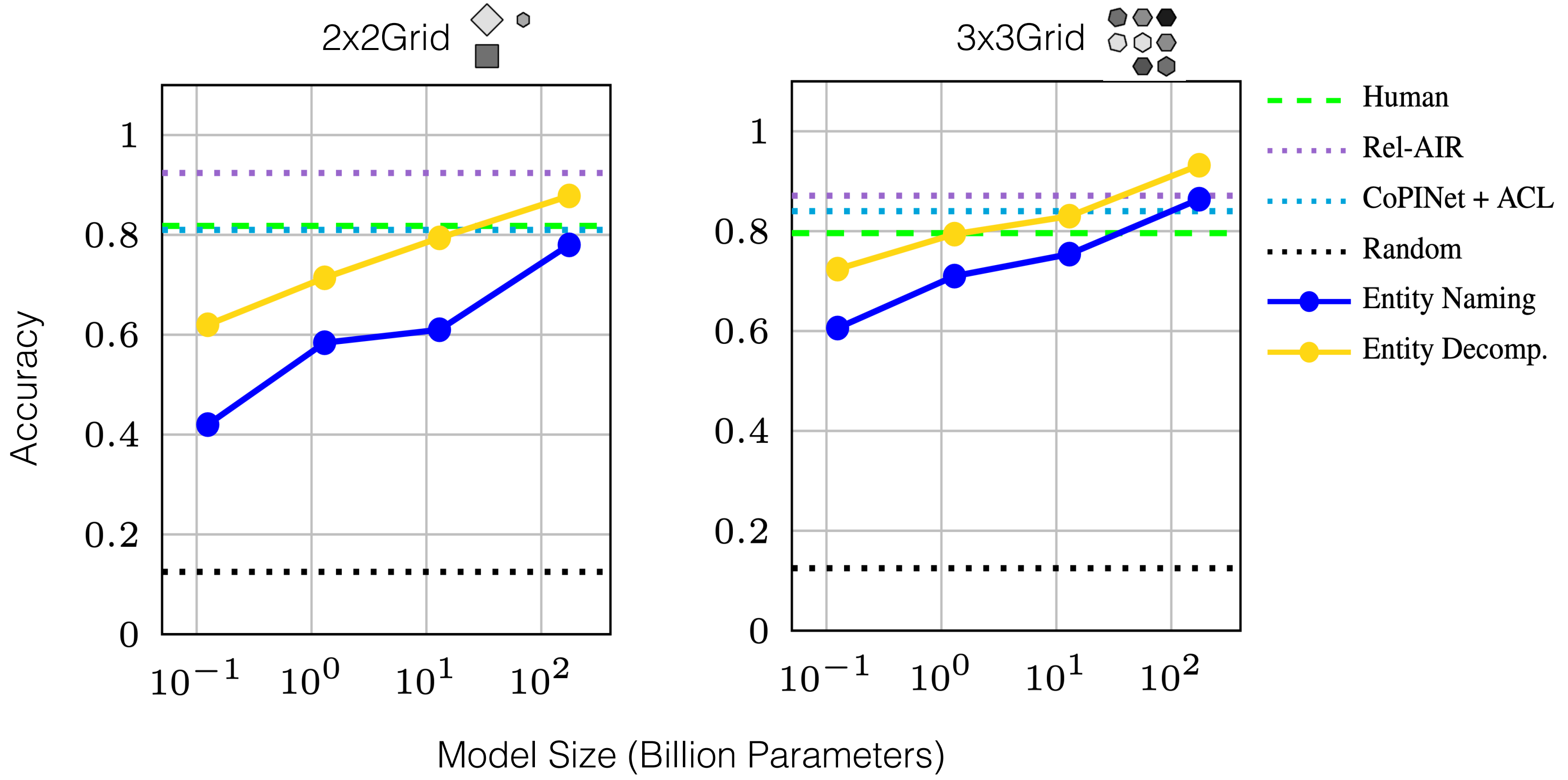


3x3Grid



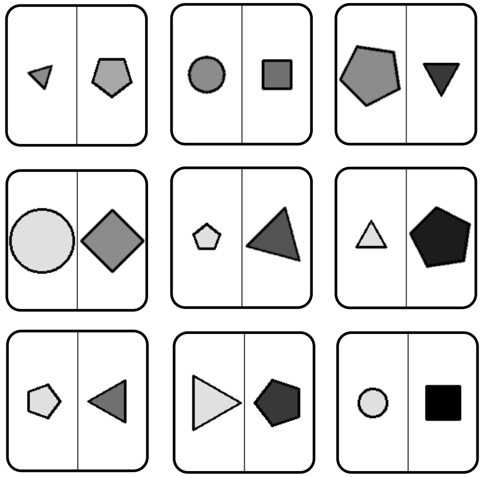
Multiple Entity Results

- Humans struggle more with task complexity than LLMs
  - Model complexity ≈ working memory?
- Can outperform humans and supervised approaches

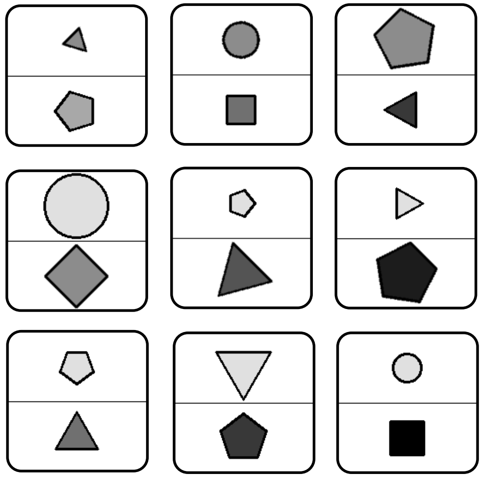


Multiple Component Results

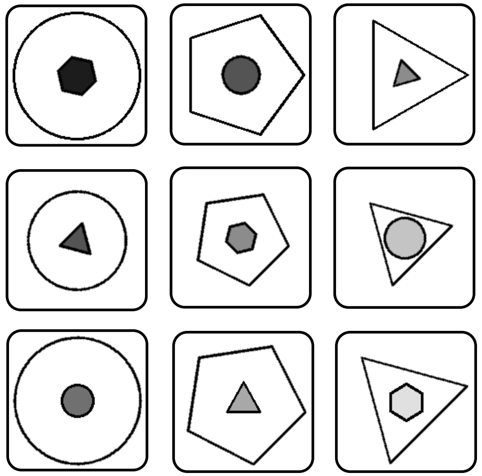
L-R



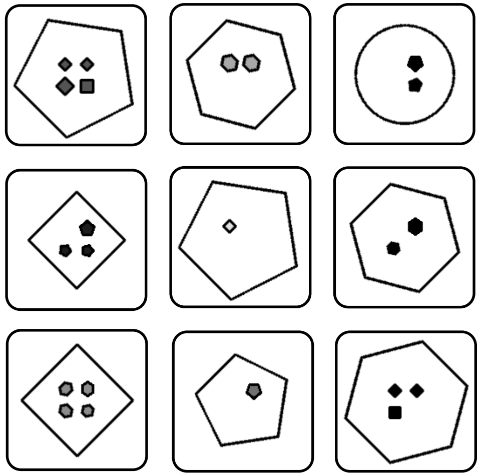
U-D



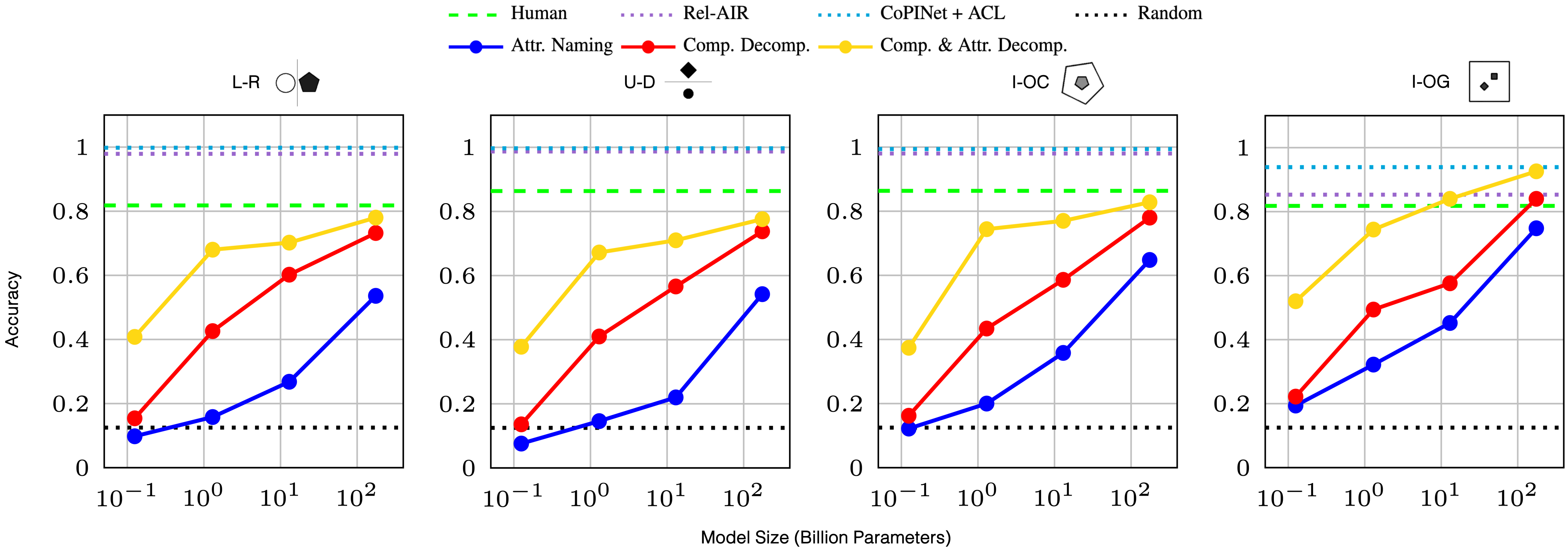
I-OC



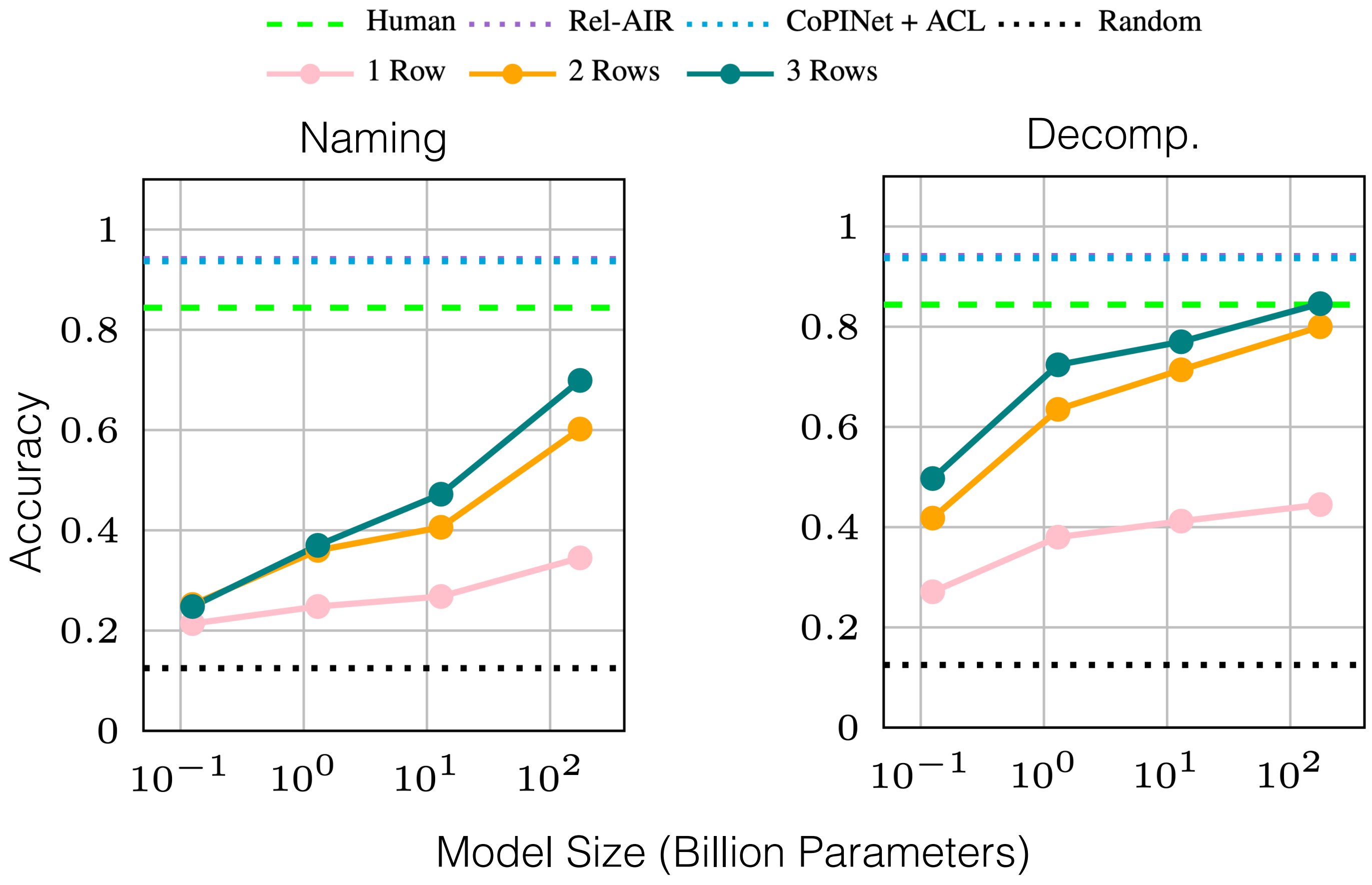
I-OG



Multiple Component Results



Impact of In-Context Learning



Takeaways

1. LLMs gain a fair capacity for abstract analogical reasoning from large-scale natural language training!
2. A number of factors strengthen their capability to make analogies:
   - Stronger language abstractions
   - LLM size
   - In-context demonstration
3. Complexity of context does not seem to impact LLMs as much as humans!

Outline

- Language Model Basics
- Application 1: Analogical Reasoning
- **Application 2: Physical Commonsense Reasoning**

From Heuristic to Analytic:
Cognitively Motivated Reasoning Strategies for Coherent Physical Commonsense in Pre-Trained Language Models

Zheyuan Zhang[1] Shane Storks[1] Fengyuan Hu[1] Sungryull Sohn[2]
Moontae Lee[2] Honglak Lee[1,2] Joyce Chai[1]

[1]University of Michigan, Computer Science and Engineering Division
[2]LG AI Research

EMNLP 2023 Long Paper

Tiered Reasoning for Intuitive Physics (TRIP)

***Story* A**

1. Ann sat in the chair.
2. Ann turned off the telephone.
3. Ann picked up a pencil.
4. Ann opened the book.
5. Ann wrote in the book.

***Story* B**

1. Ann sat in the chair.
2. **2.** Ann turned off the telephone.
3. Ann picked up a pencil.
4. Ann opened the book.
5. **5.** Ann heard the telephone ring.

!

***Which story is more plausible?* A**

Why not* B*?

Conflicting sentences: 2 → 5

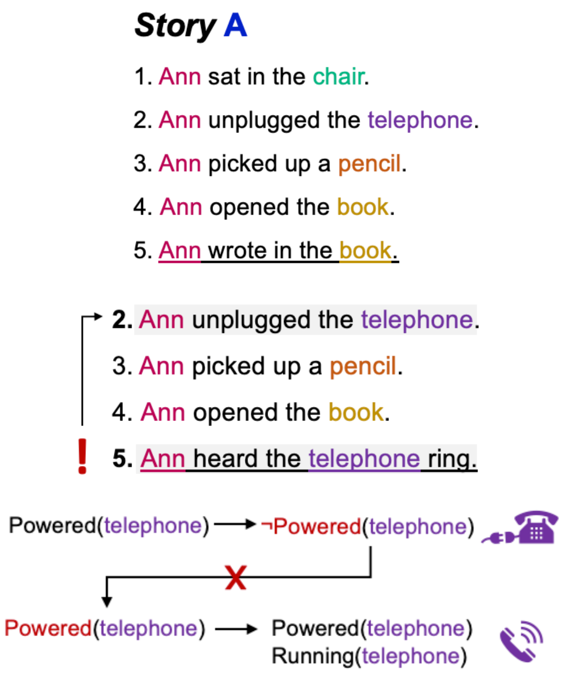
Physical states:

Powered(telephone) → ¬Powered(telephone)

Powered(telephone) → Powered(telephone) Running(telephone)

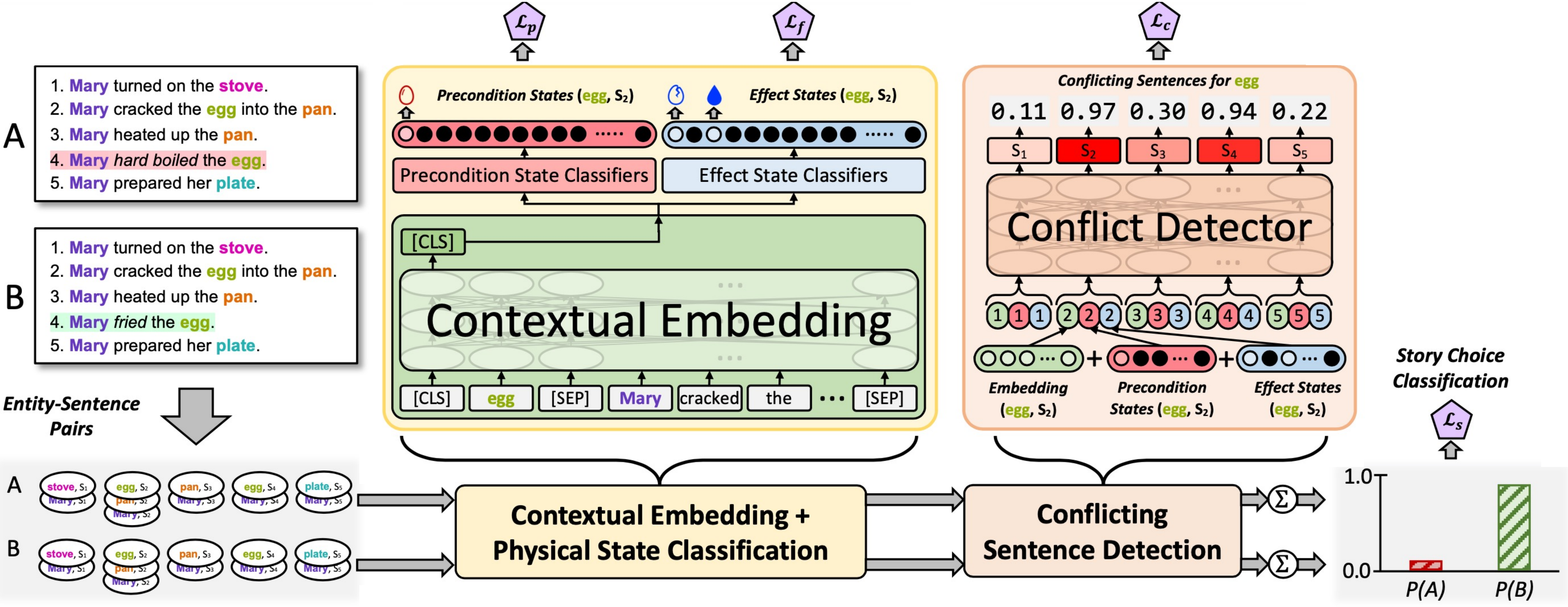
Shane Storks, Qiaozi Gao, Yichi Zhang, & Joyce Chai. 2021. Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding. *Findings of EMNLP 2021.*

Evaluation Metrics



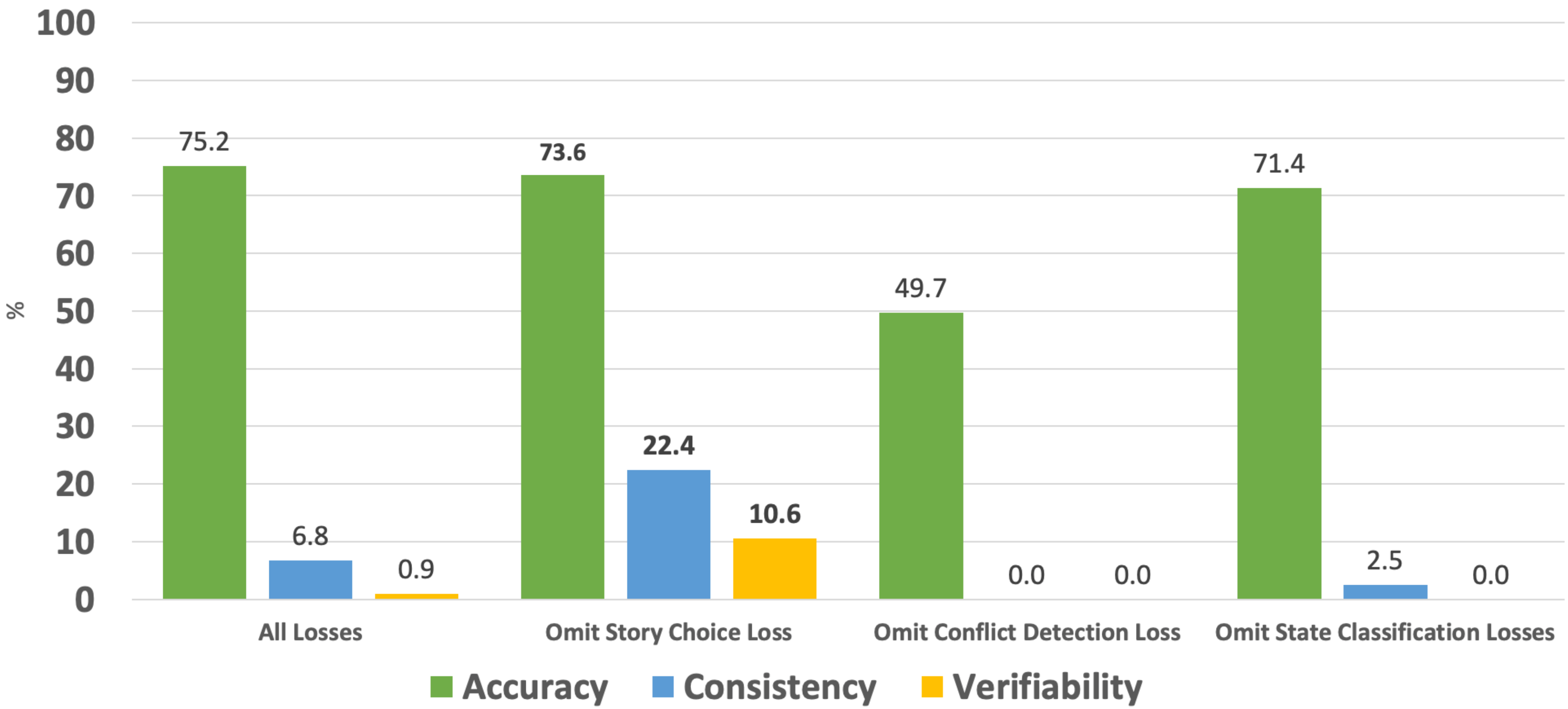
Metric	Story Choice	Conflicting Sentences	Physical States
Accuracy	✓		
Consistency	✓	✓	
Verifiability	✓	✓	✓

Tiered Baseline



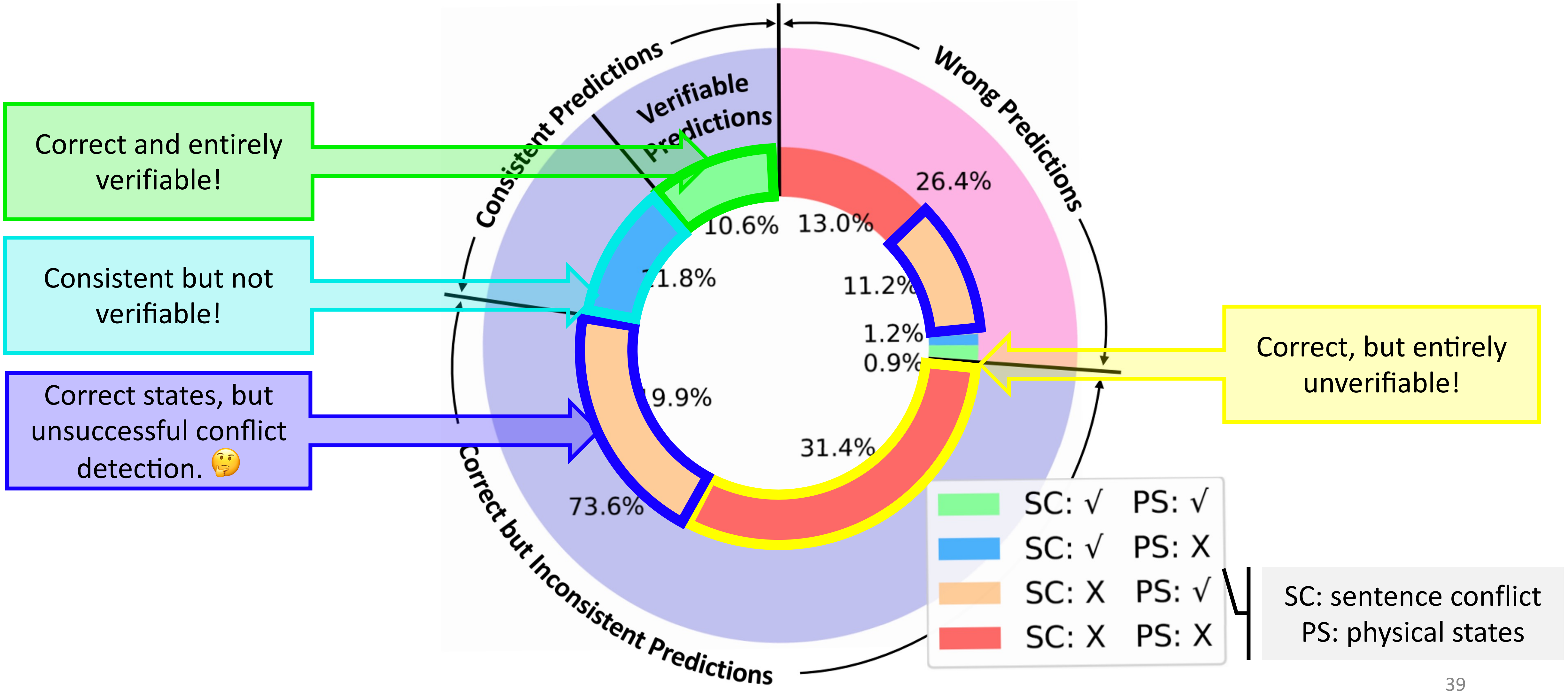
$$\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_f \mathcal{L}_f + \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s$$

RoBERTa Baseline Results on TRIP

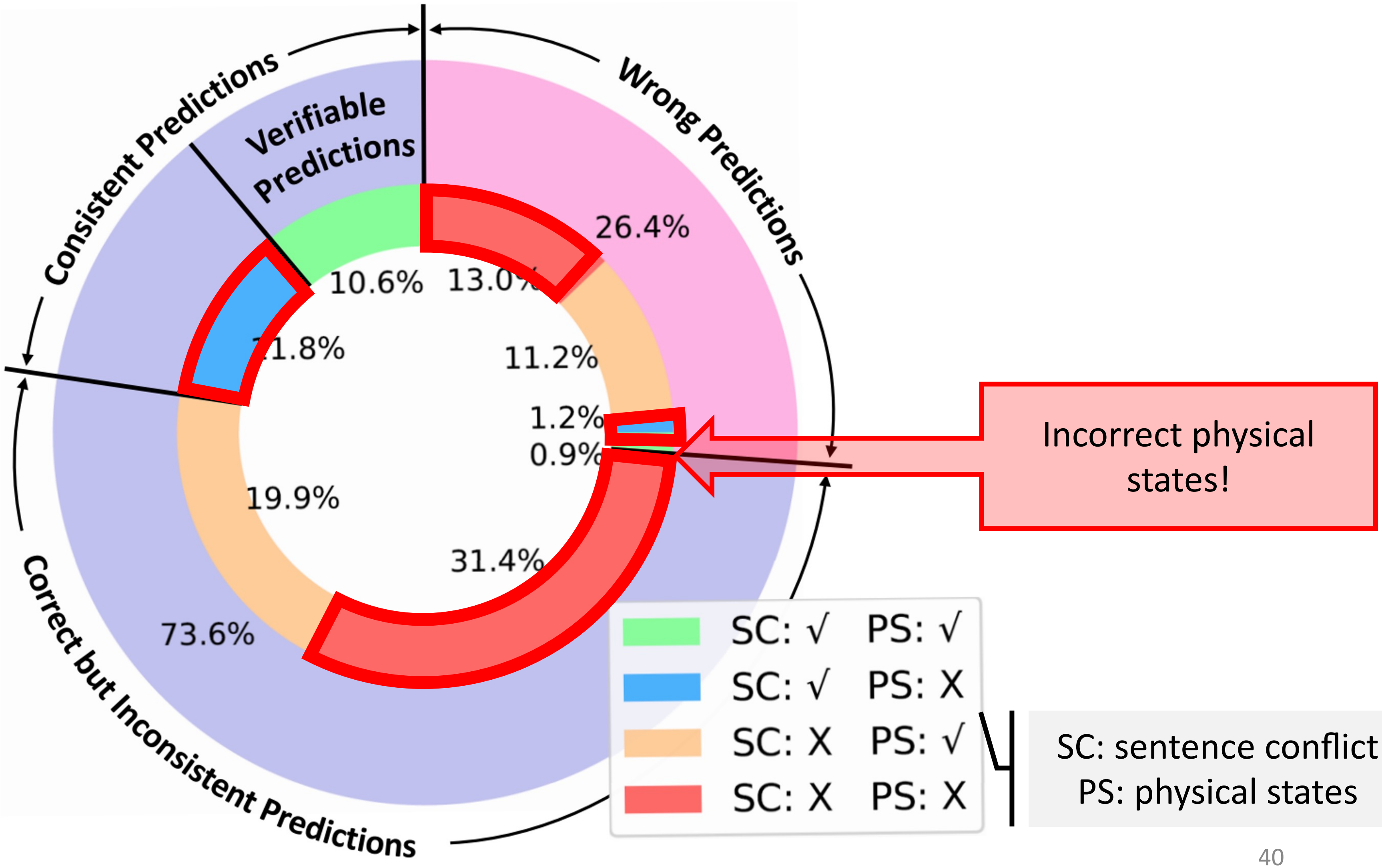


Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692.

Error Distribution

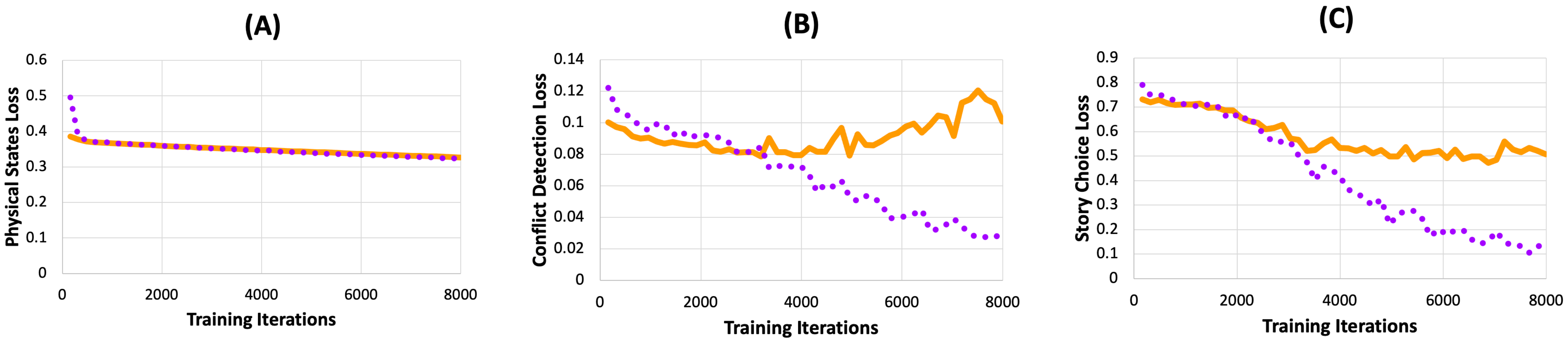


Baseline Results

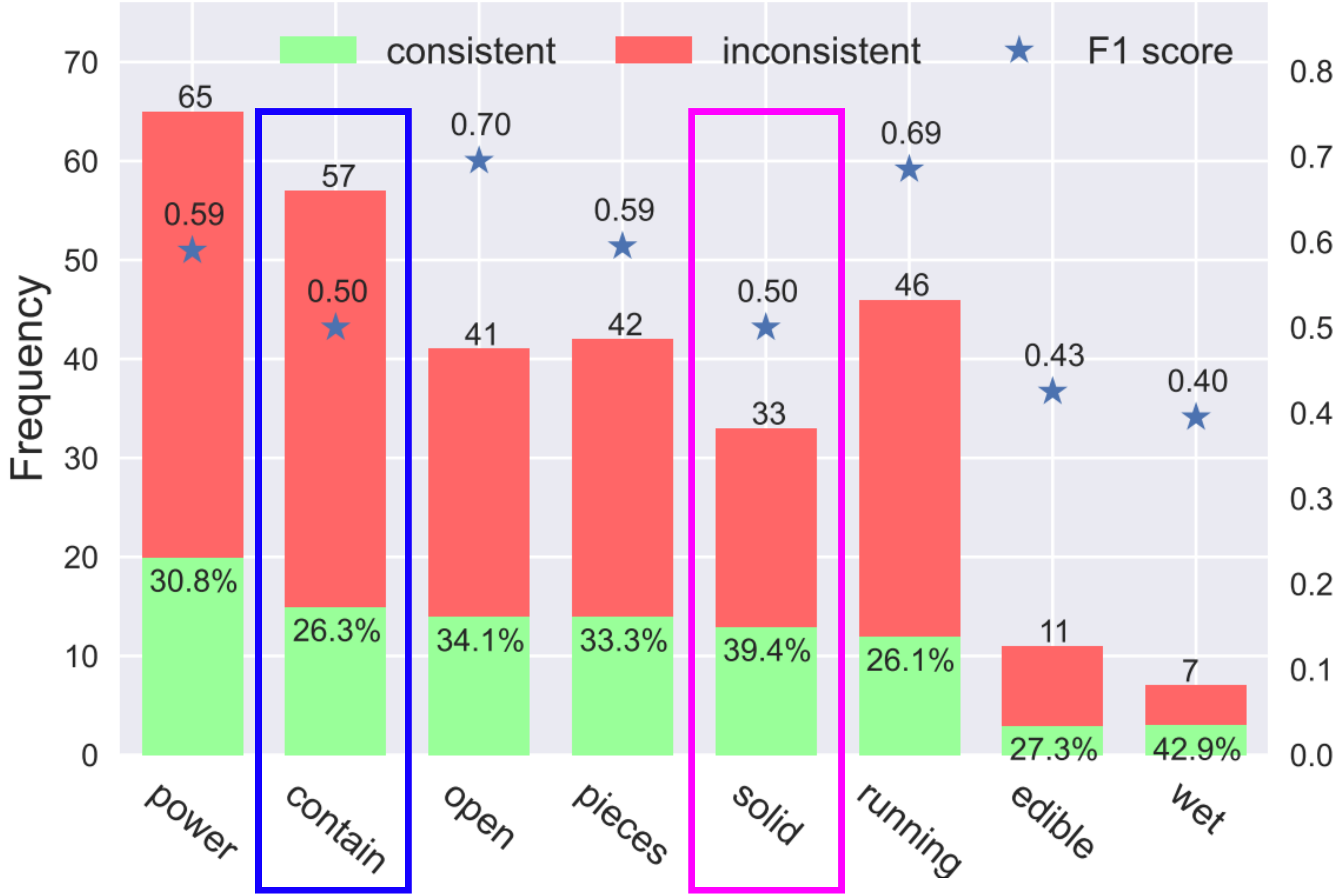


SC: sentence conflict
PS: physical states

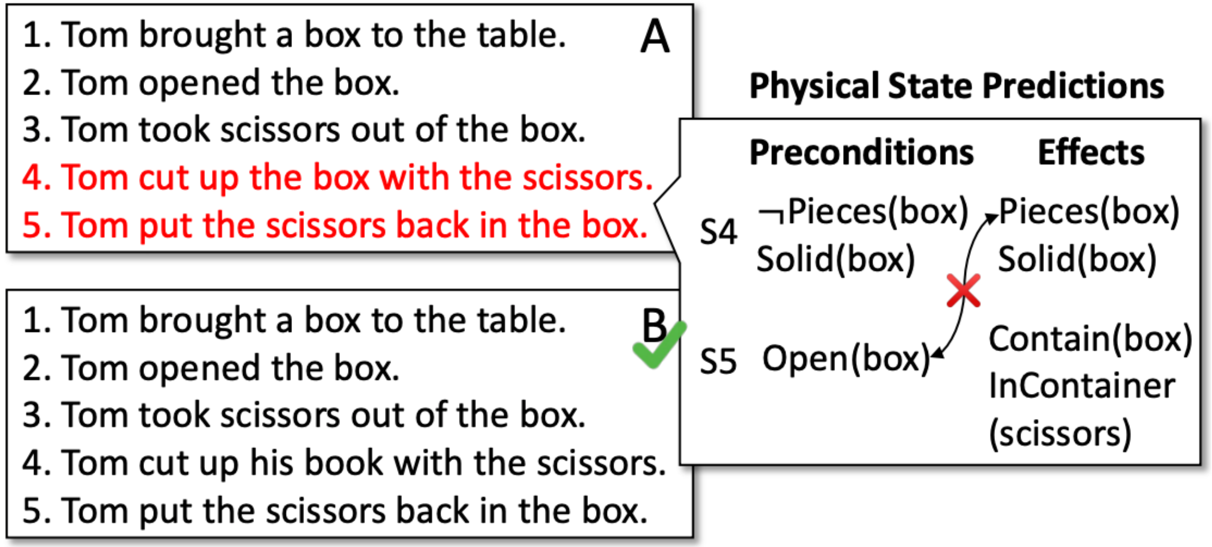
Tiered Task Learning



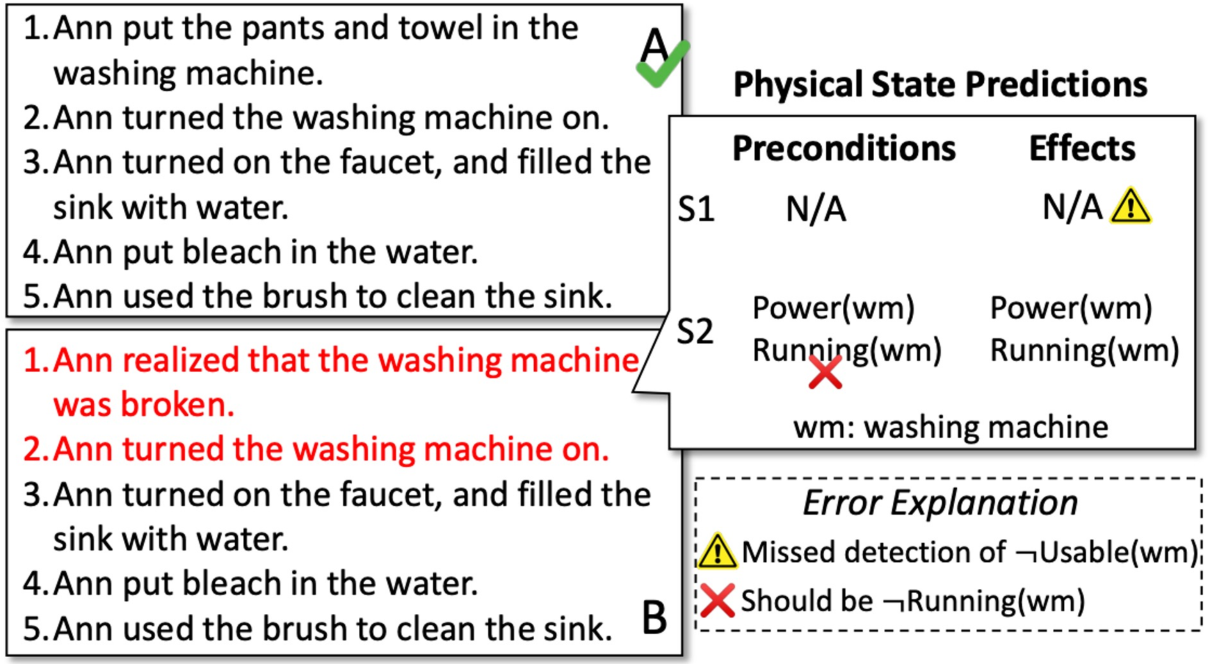
Utility of Attributes



Sample System Outputs



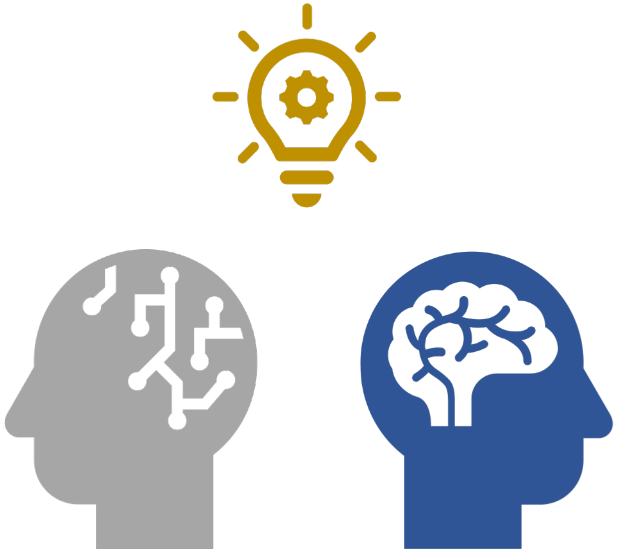
(a) A verifiable prediction.



(b) A consistent but not verifiable prediction.

Conclusion

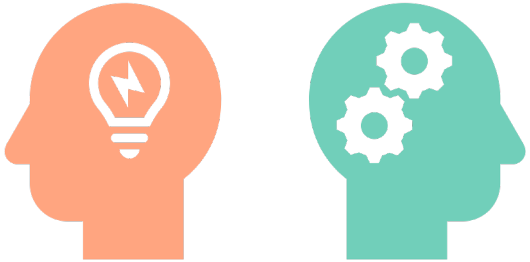
1. Natural language training creates a capacity for abstract analogical reasoning in LLMs!
2. Dual reasoning processes enable LLMs to focus on the correct language context and reason more coherently about the world through language!



Dual Processes of Human Cognition

A line of work theorizes two processes in human reasoning:

- **Heuristic:** fast, intuitive
  - Provides quick intuition for decisions; extracts most relevant info from context
- **Analytic:** slow, deliberative
  - Further operates on relevant info to rationalize and perform inference.
- Can these dual processes similarly strengthen reasoning in PLMs?



P.C. Wason & J.St.B.T. Evans. 1974. Dual processes in reasoning? *Cognition*, 3(2): 141-154.
J.St.B.T. Evans. 1984. Heuristic and analytic processes in reasoning. *British Journal of Psychology,* 75(4): 451-468.
J.St.B.T. Evans. 2010. Intuition and reasoning: A dual-process perspective. *Psychological Inquiry,* 21(4): 313-326.

2 Tasks for Coherent Physical Commonsense

TRIP

Story A:
1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary put the cucumber on the counter.
5. Mary ate the donut.

Story B:
1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. **Mary tossed the donut in the trash.**
5. **Mary ate the donut.**

Plausible story: **A**
Conflicting sentences: **(4, 5)**
States: `inedible`(**donut**) → `edible`(**donut**)

Tiered-ProPara

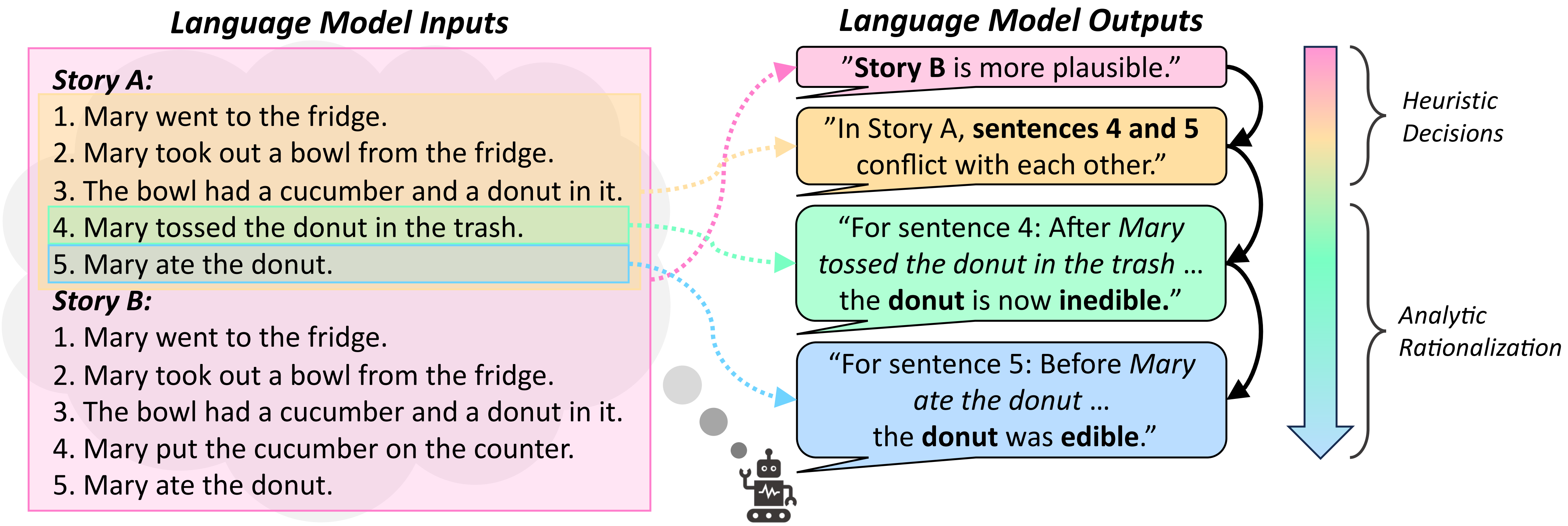
Story A:
1. Air is brought in through the mouth.
2. Passes through the lungs.
3. And into the bronchial tissue.
4. The *carbon dioxide* is removed.
5. The lungs bring the oxygen to the rest of the body.

Story B:
1. *Carbon dioxide* enters the leaves through the stomates by diffusion.
2. Water is transported to the leaves in the xylem.
3. Energy harvested through light reaction is stored by forming ATP.
4. ***Carbon dioxide*** **and energy from ATP are used to create sugar.**
5. Oxygen exits the leaves through the stomata by diffusion. …

Carbon dioxide conversion story: **B**
Carbon dioxide conversion sentence: **4**
Carbon dioxide conversion entity: **sugar**

Shane Storks, Qiaozi Gao, Yichi Zhang, & Joyce Chai. 2021. Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding. *Findings of EMNLP 2021.*
Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, & Peter Clark. 2018. Tracking State Changes in Procedural Text: a Challenge Dataset and Models for Process Paragraph Comprehension. *NAACL 2018.*

Heuristic-Analytic Reasoning (HAR)

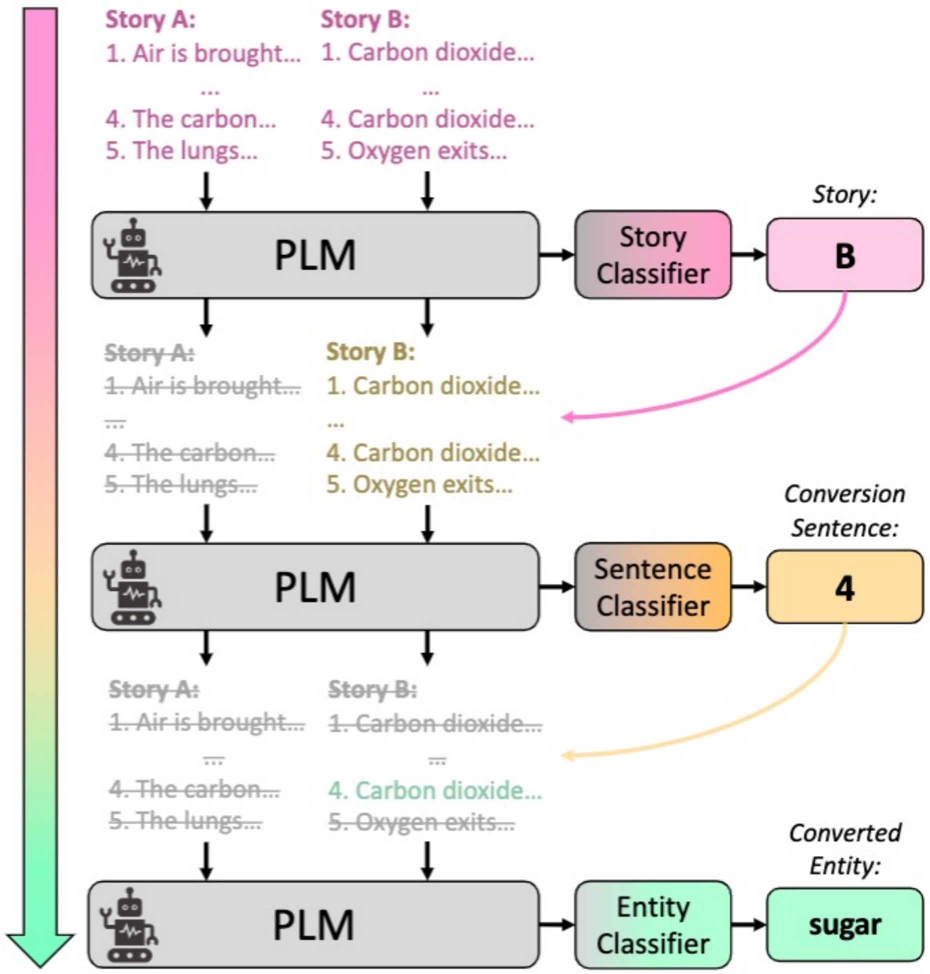


Outline

- **HAR in PLM Fine-Tuning**
- HAR in PLM In-Context Learning

Incorporating HAR into Fine-Tuning

- Coalescing Global & Local Information **(CGLI):**
  - Augments RoBERTa with temporal embedding to capture local information as states change
- Focused CGLI (**FCGLI**):
  - Small improvements to CGLI
- Focused CGLI with Heuristic-Analytic Reasoning (**FCGLI-HAR**):
  - After each prediction is made, delete segments of the context that become irrelevant



Kaixin Ma, Filip Ilievski, Jonathan Francis, et al. 2022. Coalescing Global and Local Information for Procedural Text Understanding. In *COLING 2022*.
Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692.

Fine-Tuning Results

TRIP			
Approach	**Accuracy**	***Consistency***	***Verifiability***
RoBERTa	72.9	19.1	9.1
CGLI	94.1	**77.3**	28.0
Breakpoint	80.6	53.8	32.4
FCGLI	93.7	66.2	33.8
FCGLI-HAR	94.3	75.4	**41.1**

Tiered-ProPara			
Approach	**Accuracy**	***Consistency***	***Verifiability***
FCGLI	94.5	56.7	36.2
FCGLI-HAR	95.1	**83.6**	**57.4**

Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692.
Kaixin Ma, Filip Ilievski, Jonathan Francis, et al. 2022. Coalescing Global and Local Information for Procedural Text Understanding. In *COLING 2022.*
Kyle Richardson, Ronen Tamari, Oren Sultan, et al. 2022. Breakpoint Transformers for Modeling and Tracking Intermediate Beliefs. In *EMNLP 2022.*

Outline

- HAR in PLM Fine-Tuning
- **HAR in PLM In-Context Learning**

Unstructured In-Context Learning (ICL-U)

Story A:

1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary tossed the donut in the trash.
5. Mary ate the donut.

Story B:

1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary put the cucumber on the counter.
5. Mary ate the donut.

Story B is more plausible.

In Story A, **sentences 4 and 5** conflict with each other.

For sentence 4: After *Mary tossed the donut in the trash* … the **donut** is now **inedible.**
For sentence 5: Before *Mary ate the donut* … the **donut** was **edible**.

In-Context Learning with Traditional CoT (ICL-CoT)

Story A:
1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary tossed the donut in the trash.
5. Mary ate the donut.

Story B:
1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary put the cucumber on the counter.
5. Mary ate the donut.

(CoT)
Story B is more plausible.

(CoT)
In Story A, **sentences 4 and 5** conflict with each other.

(CoT)
For sentence 4: After *Mary tossed the donut in the trash* … the **donut** is now **inedible.**
For sentence 5: Before *Mary ate the donut* … the **donut** was **edible**.

In-Context Learning with HAR (ICL-HAR)

Story A:

1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary tossed the donut in the trash.
5. Mary ate the donut.

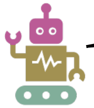
Story B:

1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary put the cucumber on the counter.
5. Mary ate the donut.

Story B is more plausible.

In Story A, **sentences 4 and 5** conflict with each other.

For sentence 4: After *Mary tossed the donut in the trash* … the **donut** is now **inedible.**
For sentence 5: Before *Mary ate the donut …* the **donut** was **edible**.



Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems 35.*

In-Context Learning Results

InstructGPT

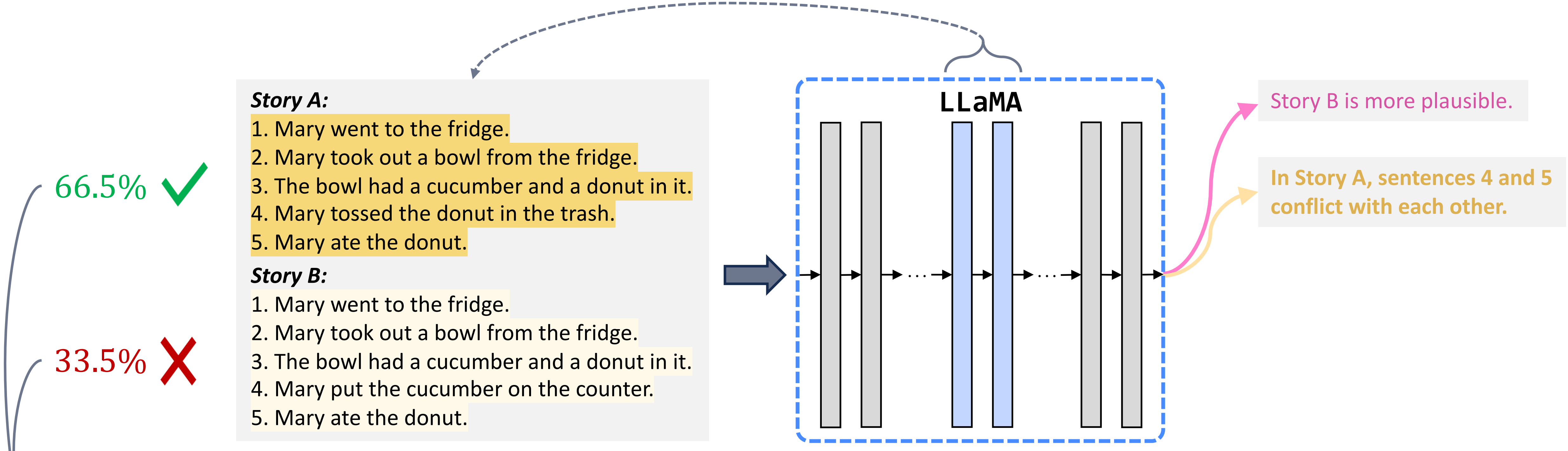
	TRIP			Tiered-ProPara		
Approach	Acc.	*Cons.*	*Ver.*	Acc.	*Cons.*	*Ver.*
ICL-U	70.9	40.7	7.1	54.9	17.4	5.2
ICL-CoT	75.0	40.7	10.8	50.7	19.2	7.5
ICL-HAR	72.6	**47.9**	**23.9**	54.9	**31.5**	**20.7**

LLaMA

	TRIP			Tiered-ProPara		
Approach	Acc.	*Cons.*	*Ver.*	Acc.	*Cons.*	*Ver.*
ICL-U	70.4	42.3	14.8	51.2	3.8	1.4
ICL-CoT	74.6	42.3	19.7	57.3	9.4	4.2
ICL-HAR	55.6	**44.4**	**35.2**	41.8	**17.8**	**13.1**

Long Ouyang, Jeff Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. arXiv: 2203.02155.
Hugo Touvran et al. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv: 2302.13971.

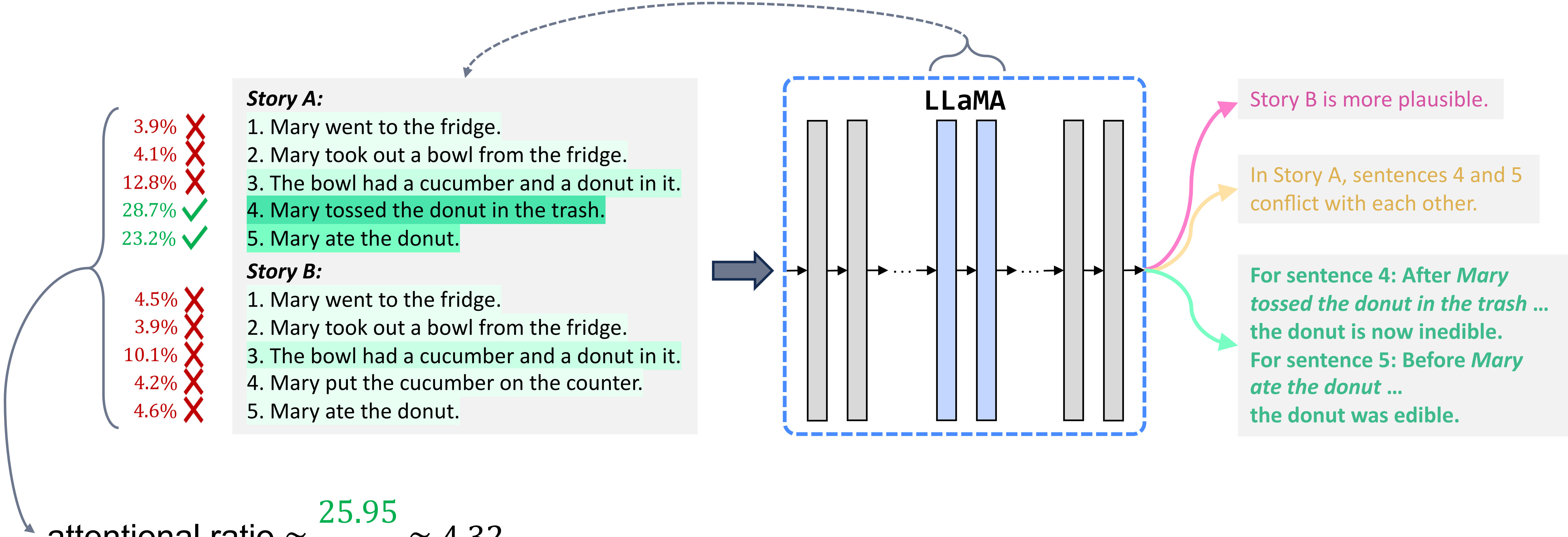
Attention Analysis



$$\text{attentional ratio} = \frac{66.5}{33.5} \approx 1.99$$

Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv: 2302.13971*.

Attention Analysis



$$\text{attentional ratio} \approx \frac{25.95}{6.01} \approx 4.32$$

Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv: 2302.13971*.

Attentional Precision and Recall

- To measure how attended context and correct predictions correlate, we use **attentional precision** and **attentional recall**
  - *True/false positive*: Correct attention, and correct/incorrect prediction
  - *True/false negative*: Incorrect attention, and correct/incorrect prediction

Attention Analysis Results

- PLMs focus better on the correct language context during each step of reasoning
- Faithful attention and coherent reasoning go hand in hand!

Sentence Selection Step

	TRIP			Tiered-ProPara		
Approach	Ratio	Prec.	Rec.	Ratio	Prec.	Rec.
ICL-U	0.96	42.6	39.6	0.90	14.8	30.6
ICL-HAR	**1.07**	**75.2**	**48.7**	**1.80**	**51.1**	**58.2**

Physical State Prediction Step

	TRIP			Tiered-ProPara		
Approach	Ratio	Prec.	Rec.	Ratio	Prec.	Rec.
ICL-U	1.23	43.0	35.4	1.21	14.6	25.9
ICL-HAR	**1.95**	**79.8**	**98.2**	**2.20**	**72.1**	**83.3**

Conclusion

- Human-inspired heuristic-analytic reasoning helps PLMs reason more coherently when applied to downstream tasks
- Successful because it helps PLMs focus on the correct language context at each step of reasoning
- Check out our paper for more details and results!

From Heuristic to Analytic: Cognitively Motivated Strategies for Coherent Physical Commonsense Reasoning

Zheyuan Zhang[1*] **Shane Storks**[1*] **Fengyuan Hu**[1] **Sungryull Sohn**[2]
Moontae Lee[2,3] **Honglak Lee**[1,2] **Joyce Chai**[1]
[1]University of Michigan [2]LG AI Research
[3]University of Illinois at Chicago
{zheyuan, sstorks, hufy, chaijy}@umich.edu
{srsohn, moontae.lee, honglak}@lgresearch.ai

Thank you!