Commonsense Reasoning
in Natural Language Understanding

Shane Storks

(he/him)

Situated Language and Embodied Dialogue

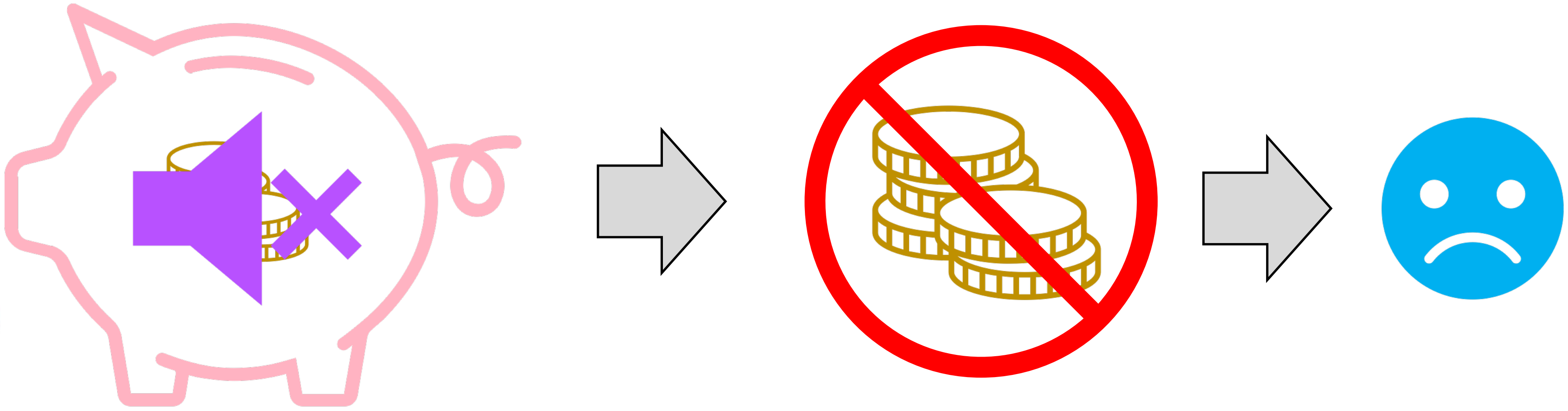
sstorks@umich.edu

EECS 595 (Natural Language Processing) Guest Lecture

Commonsense Reasoning in NLU

"Jack needed some money, so he went and shook his piggy bank. He was disappointed when it made no sound."

(dreamstime)



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

Davis, E. & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. In *Commun. ACM*, 58(9): p. 92-103.

Outline

1. Introduction to Commonsense
2. Benchmarks and Resources for Commonsense
3. State of the Art and Limitations
4. Coherent Commonsense Reasoning

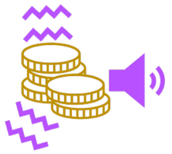
Outline

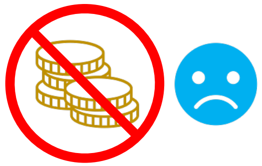
1. **Introduction to Commonsense**
2. Benchmarks and Resources for Commonsense
3. State of the Art and Limitations
4. Coherent Commonsense Reasoning

Commonsense Knowledge

- **Contextual knowledge:** Knowledge situated in specific circumstances
- **Common knowledge**: Factual knowledge about the world
  - Widely agreed upon by a large group of people
  - Can be learned from a book

- **Commonsense knowledge:** Low-level knowledge about how the world works
  - May be widely agreed upon but typically unstated
  - May vary slightly over cultures, regions, time, etc.
  - Learned from life experience (often from very young age)
  - *Especially challenging for machines!*



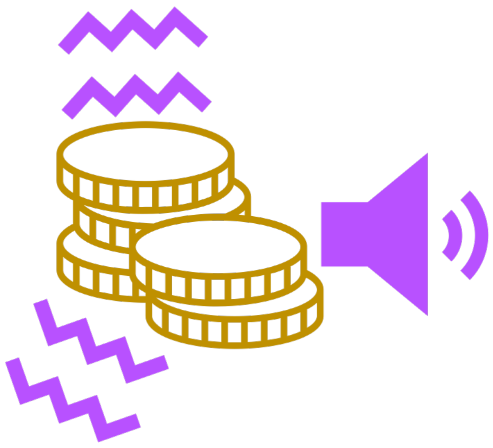


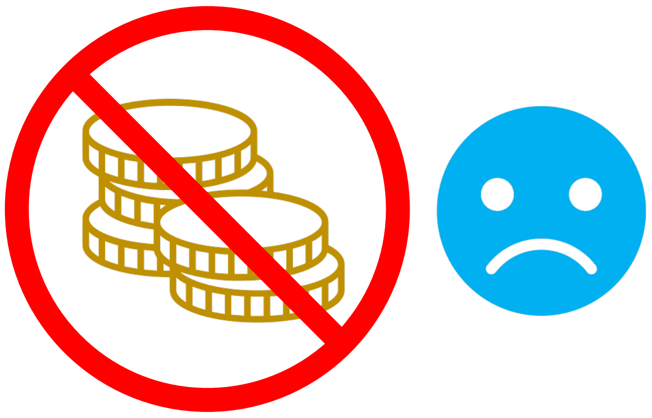
S. Storks, Q. Gao, & J.Y. Chai. (2019). Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. arXiv: 1904.011672 [cs.CL].

E. Davis. (2017). Logical Formalizations of Commonsense Reasoning: A Survey. In *Journal of Artificial Intelligence*, 59.

Types of Commonsense

- No perfect taxonomy of this space
- Two key types of commonsense we develop at a very young age:
  - **Intuitive physics (physical commonsense)**
  - **Intuitive psychology (social commonsense)**





B.M. Lake, T.D. Ullman, J.B. Tenenbaum, & S. Gershman. (2017). Building Machines that Learn and Think Like People. In *Behavioral and Brain Sciences,* 40.

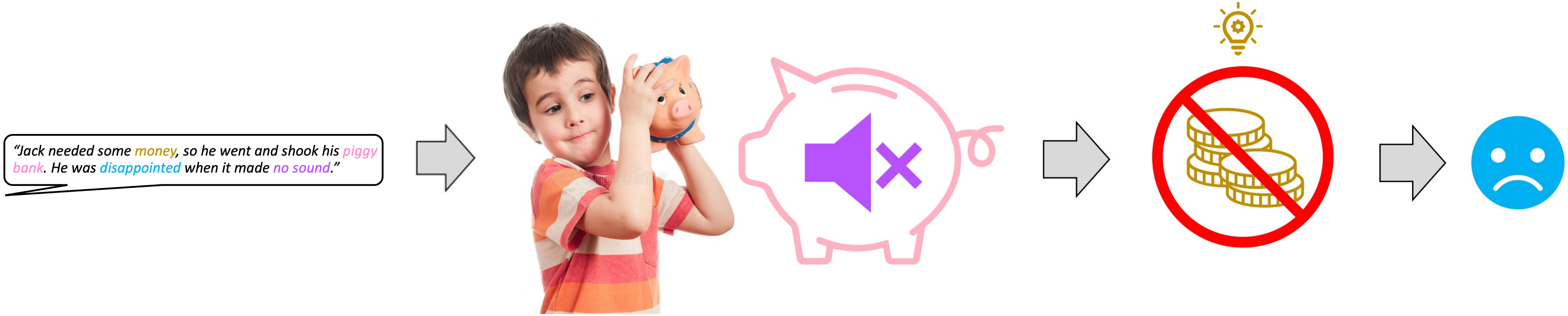
H.M. Wellman & S.A. Gelman. (1992) Cognitive development: Foundational theories of core domains. In *Annual Review of Psychology,* 43.

Example

NBC News

Commonsense Reasoning in NLU

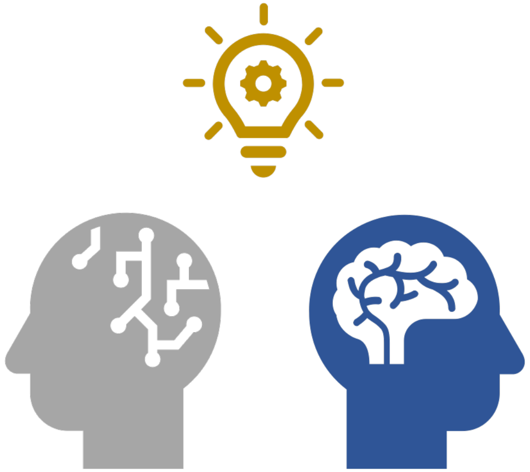
- **Commonsense reasoning:** Connecting pieces of commonsense knowledge together to reach new conclusions.
- **Commonsense reasoning -> natural language understanding (NLU):** *Deep understanding of language that goes beyond what is explicitly expressed, rather relying on new conclusions inferred from commonsense knowledge about how the world works.*



S. Storks, Q. Gao, & J.Y. Chai. (2019). Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. arXiv: 1904.011672 [cs.CL].

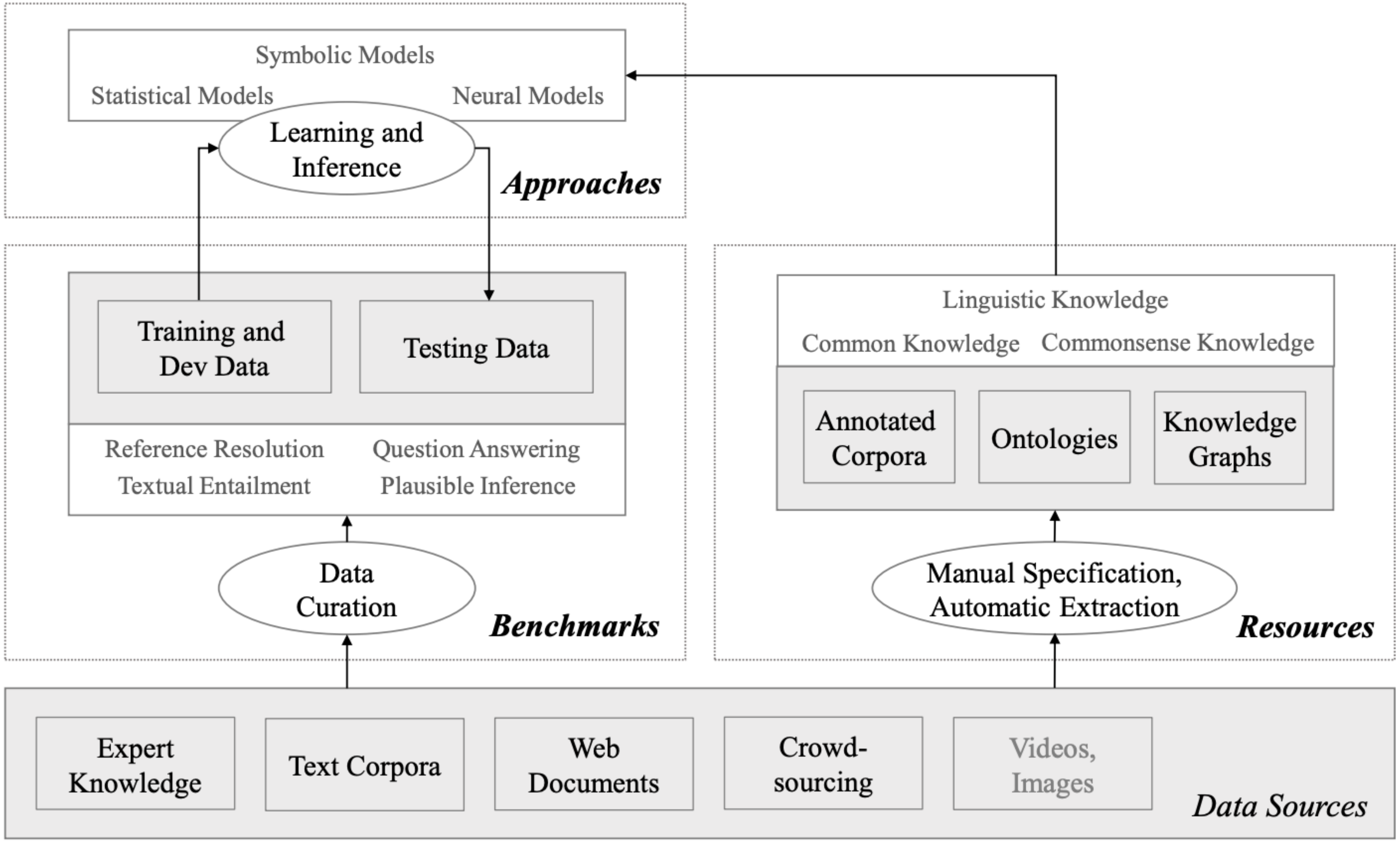
Challenges for Machine Commonsense

- Commonsense knowledge and reasoning comes naturally to us when we think, act, and communicate
- But commonsense has been notoriously hard for machines:
  - A lot of it (estimated 100M axioms in adults)
  - Not often stated explicitly (*reporting bias)*
  - Long tail
  - Subjectivity
    - Regional and cultural variations
    - Value plurality



T. Chklovski. (2003). Learner: A System for Acquiring Commonsense Knowledge by Analogy. In *K-CAP '03*.
E. Davis & G. Marcus. (2015). Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. In *Communications of the ACM*, 58(9).
J. Gordon & B. Van Durme. (2013). Reporting Bias and Knowledge Acquisition. In *AKBC '13*.
E. Davis. (2017). Logical Formalizations of Commonsense Reasoning: A Survey. In *Journal of Artificial Intelligence*, 59.
T. Sorenson, L. Jiang, J. Hwang, et al. (2023). Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. arXiv: 2309.00779.

Overview of Commonsense in NLU Research

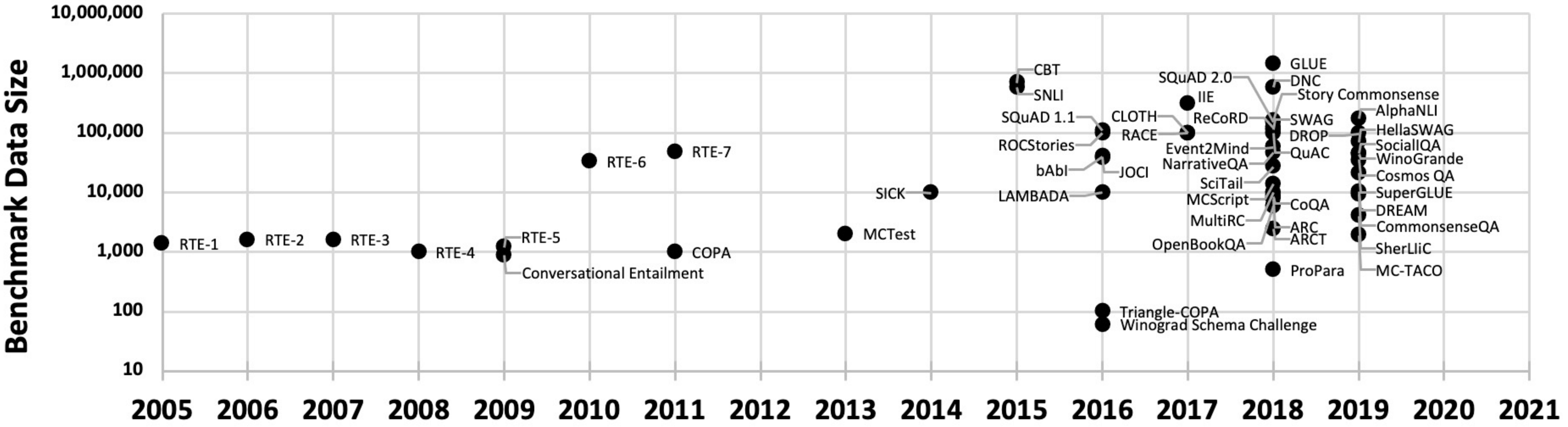


S. Storks, Q. Gao, & J.Y. Chai. (2019). Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. arXiv: 1904.011672 [cs.CL].

Outline

1. Introduction to Commonsense
2. **Benchmarks and Resources for Commonsense**
3. State of the Art and Limitations
4. Coherent Commonsense Reasoning

Benchmark Datasets



S. Storks, Q. Gao, & J.Y. Chai. (2019). Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. arXiv: 1904.011672 [cs.CL].
E. Davis. (2023). Benchmarks for Automated Commonsense Reasoning. arXiv: 2302.04752.

Commonsense Question Answering

(A) MCScript (Ostermann et al., 2018)
Did they throw away the old diaper?

a. **Yes, they put it into the bin.**
b. No, they kept it for a while.

(B) OpenBookQA (Mihaylov et al., 2018)
Which of these would let the most heat travel through?

a. a new pair of jeans.
b. **a steel spoon in a cafeteria.**
c. a cotton candy at a store.
d. a calvin klein cotton hat.

Evidence: Metal is a thermal conductor.

(C) CoQA (Reddy et al., 2018)
The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Who is the democratic candidate?
Terry McAuliffe
Evidence: Democrat Terry McAuliffe

Who is his opponent?
Ken Cuccinelli
Evidence: Republican Ken Cuccinelli

J. McCarthy. (1976). An example for natural language understanding and the AI problems it raises. *Formalizing Common Sense: Papers by John McCarthy*, 355.
S. Ostermann, A. Modi, M. Roth, et al. (2018). MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge. In *LREC-2018.*
T. Mihaylov, P. Clark, T. Khot. & A. Sabharwal. (2018). Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP 2018*.
S. Reddy, D. Chen, & C.D. Manning. (2018). CoQA: A Covnersational Question Answering Challenge. In *Transactions of ACL*, 7.

Commonsense Plausible Inference

(A) COPA (Roemmele, Bejan, & Gordon, 2011)
I knocked on my neighbor's door. What happened as result?

a. **My neighbor invited me in.**
b. My neighbor left his house.

(B) ROCStories (Mostafazadeh et al., 2016)
Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee.

Ending:

a. **Tom asked Sheryl to marry him.**
b. He wiped mud off of his boot.

(C) AlphaNLI (Bhagavatula, Bras, Malaviya, Sakaguchi, Holtzman, Rashkin, Downey, Yih, & Choi, 2019)
Observation 1: There was ten feet of snow outside.
Observation 2: In all that time I was unable to check my mail.
Hypotheses:

a. **I couldn't open my door against a drift for 3 days.**
b. It took 10 minutes for the snow plow to come through.

E. Davis & G. Marcus. (2015). Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. In *Communications of the ACM*, 58(9).
C.S. Peirce. (1883). A Theory of Probable Inference. In *Studies in Logic by Members of the John Hopkins University.*
M. Roemmele, C.A. Bejan, & A. Gordon. (2011). Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning.*
N. Mostafazadeh, N. Chambers, X. He, et al. (2016). A Corpus and Cloze Evaluation Framework for Deeper Understanding of Commonsense Stories. In *NAACL HLT 2016*.
C. Bhagavatula, Le Bras, R., C. Malaviya, et al. (2019). Abductive commonsense reasoning. In *ICLR 2020.*

Commonsense Reference Resolution

(A) Winograd Schema Challenge (Levesque, 2011)
The trophy would not fit in the brown suitcase because it was too big. What was too big?

a. **The trophy**
b. The suitcase

The trophy would not fit in the brown suitcase because it was too small. What was too small?

a. The trophy
b. **The suitcase**

(B) Winogender (Rudinger et al., 2018a)
The paramedic performed CPR on the passenger even though she knew it was too late. Who knew it was too late?

a. **The paramedic**
b. The passenger

(C) (Rahman & Ng, 2012)
Lions eat zebras because they are predators. Who are predators?

a. **Lions**
b. Zebras

H.J. Levesque. (2011). The Winograd Schema Challenge. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning.*
R. Rudinger, J. Naradowsky, B. Leonard, & B. Van Durme. (2018). Gender Bias in Coreference Resolution. In *NAACL HLT 2018*.
A. Rahman & V. Ng. (2012). Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In *EMNLP-CoNLL 2012.*
K. Sakaguchi, R. Le Bras, C. Bhagavatula, & Y. Choi. (2019). WinoGrande: An Adversarial Winograd Schema Challenge at Scale. arXiv: 1907.10641.

Commonsense Textual Entailment

(A) RTE Challenge (Dagan et al., 2005)
Text: American Airlines began laying off hundreds of flight attendants on Tuesday, after a federal judge turned aside a union's bid to block the job losses.
Hypothesis: American Airlines will recall hundreds of flight attendants as it steps up the number of flights it operates.

Label: **not entailment**

(B) SICK (Marelli, Menini, Baroni, Bentivogli, Bernardi, & Zamparelli, 2014a)[5]
Sentence 1: Two children are lying in the snow and are drawing angels.
Sentence 2: Two children are lying in the snow and are making snow angels.

Label: **entailment**

(C) SNLI (Bowman et al., 2015)
Text: A black race car starts up in front of a crowd of people.
Hypothesis: A man is driving down a lonely road.

Label: **contradiction**

(D) MultiNLI, Telephone (Williams, Nangia, & Bowman, 2017)
Context: that doesn't seem fair does it
Hypothesis: There's no doubt that it's fair.

Label: **contradiction**

I. Dagan, O. Glickman, & B. Magnini. (2005). The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment,* 3944.
M. Marelli, S. Menini, M. Baroni, et al. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *LREC-2014*.
S.R. Bowman, G. Angeli, C. Potts, & C.D. Manning. (2015). A large annotated corpus for learning natural language inference. In *EMNLP 2015*.
A. Williams, N. Nangia, & S.R. Bowman. (2017). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL HLT 2018*.

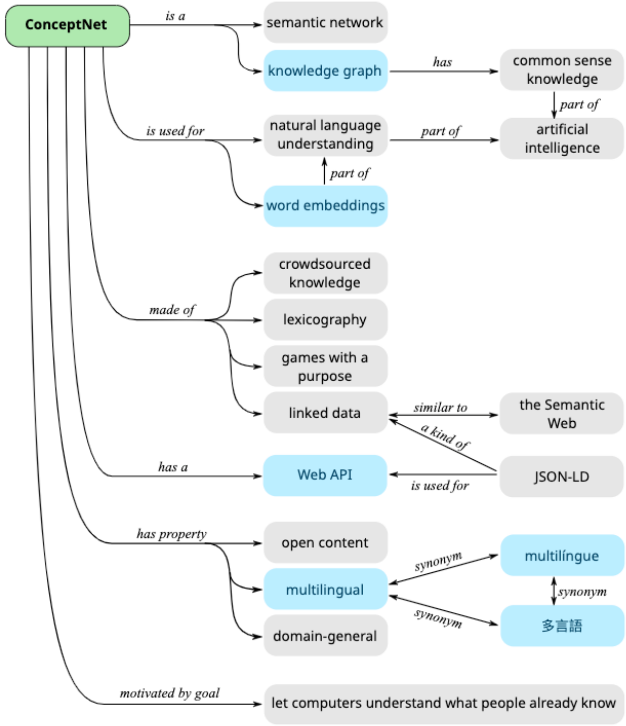
Knowledge Resources

Efforts to collect commonsense knowledge more directly:

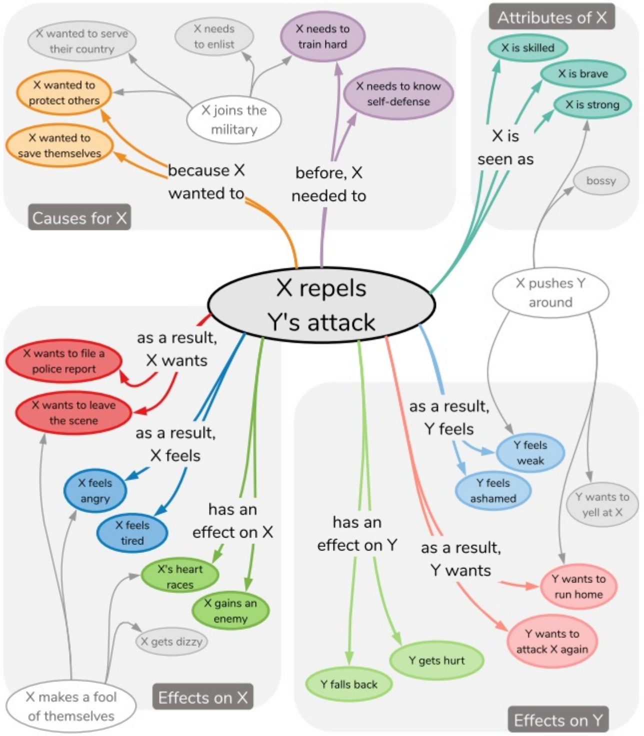
Cyc

Daniel Kornev

ConceptNet



ATOMIC



S. Storks, Q. Gao, & J.Y. Chai. (2019). Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. arXiv: 1904.011672 [cs.CL].
D.B. Lenat & R.V. Guha. (1989). Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project.
R. Speer, J. Chin, & C. Havasi. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI 2017*.
M. Sap, R. Le Bras, E. Allaway, et al. (2019). ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *AAAI 2019*.

Outline

1. Introduction to Commonsense
2. Benchmarks and Resources for Commonsense
3. **State of the Art and Limitations**
4. Coherent Commonsense Reasoning

If commonsense is difficult for AI, what's all this about?

Scientists Developed an AI So Advanced They Say It's Too Dangerous to Release

TECH 19 February 2019 By PETER DOCKRILL

(Science Alert)

A.I. Is Mastering Language. Should We Trust What It Says?

OpenAI's GPT-3 and other neural nets can now write original prose with mind-boggling fluency — a development that could have profound implications for the future.

(New York Times)

AI, ML & DATA ENGINEERING

InfoQ Live (June 22nd) - Overcome Cloud and Serverless Security Challenges

AI Models from Google and Microsoft Exceed Human Performance on Language Understanding Benchmark

LIKE DISCUSS

(InfoQ)

JAN 12, 2021 • 3 MIN READ

Research teams from Google and Microsoft have recently developed natural language

ARTIFICIAL INTELLIGENCE

OpenAI's upgraded ChatGPT reaches shockingly powerful 'human-level performance' (TweakTown)

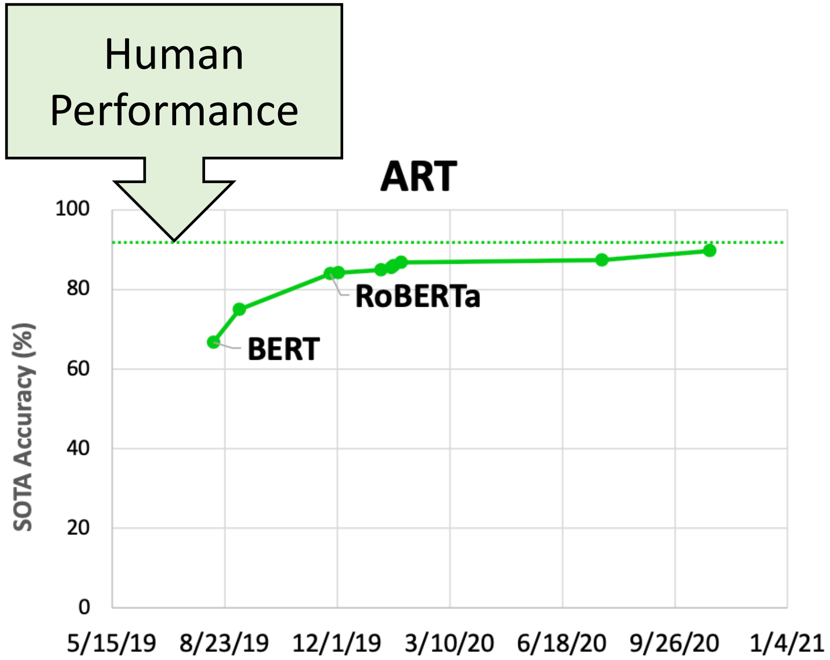
Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang

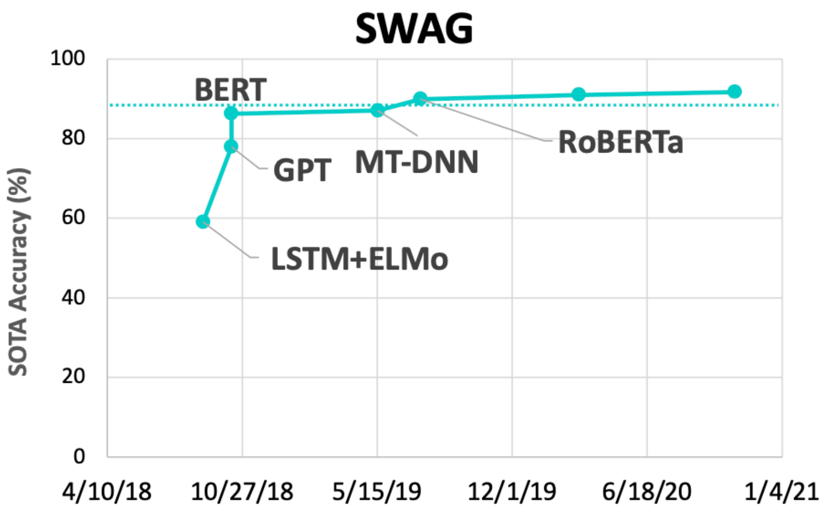
Microsoft Research

S. Bubeck, V. Chandrasekaran, R. Eldan, et al. (2023). Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv: 2303.12712.

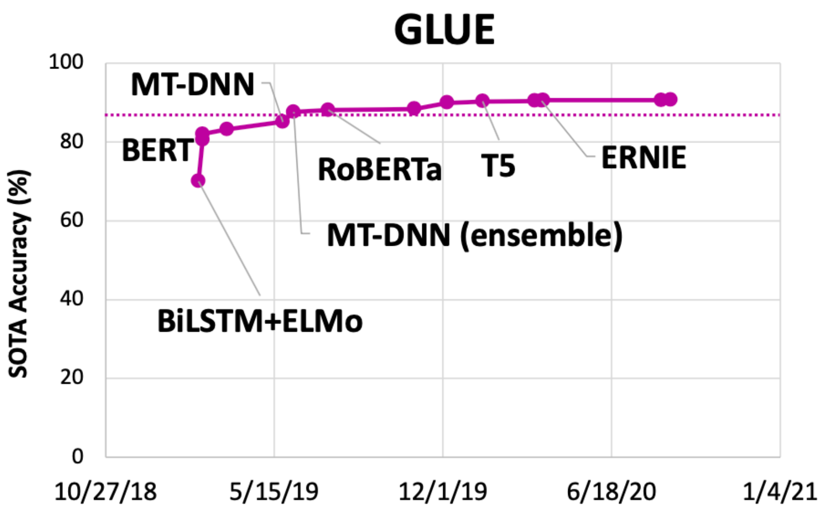
Human-Level Results



https://leaderboard.allenai.org/anli/submissions/public



https://leaderboard.allenai.org/swag/submissions/public



https://gluebenchmark.com/leaderboard

Applying LMs to Classification Tasks

Which sentence is most likely to fill in the blank?

It was a very hot summer day.

____

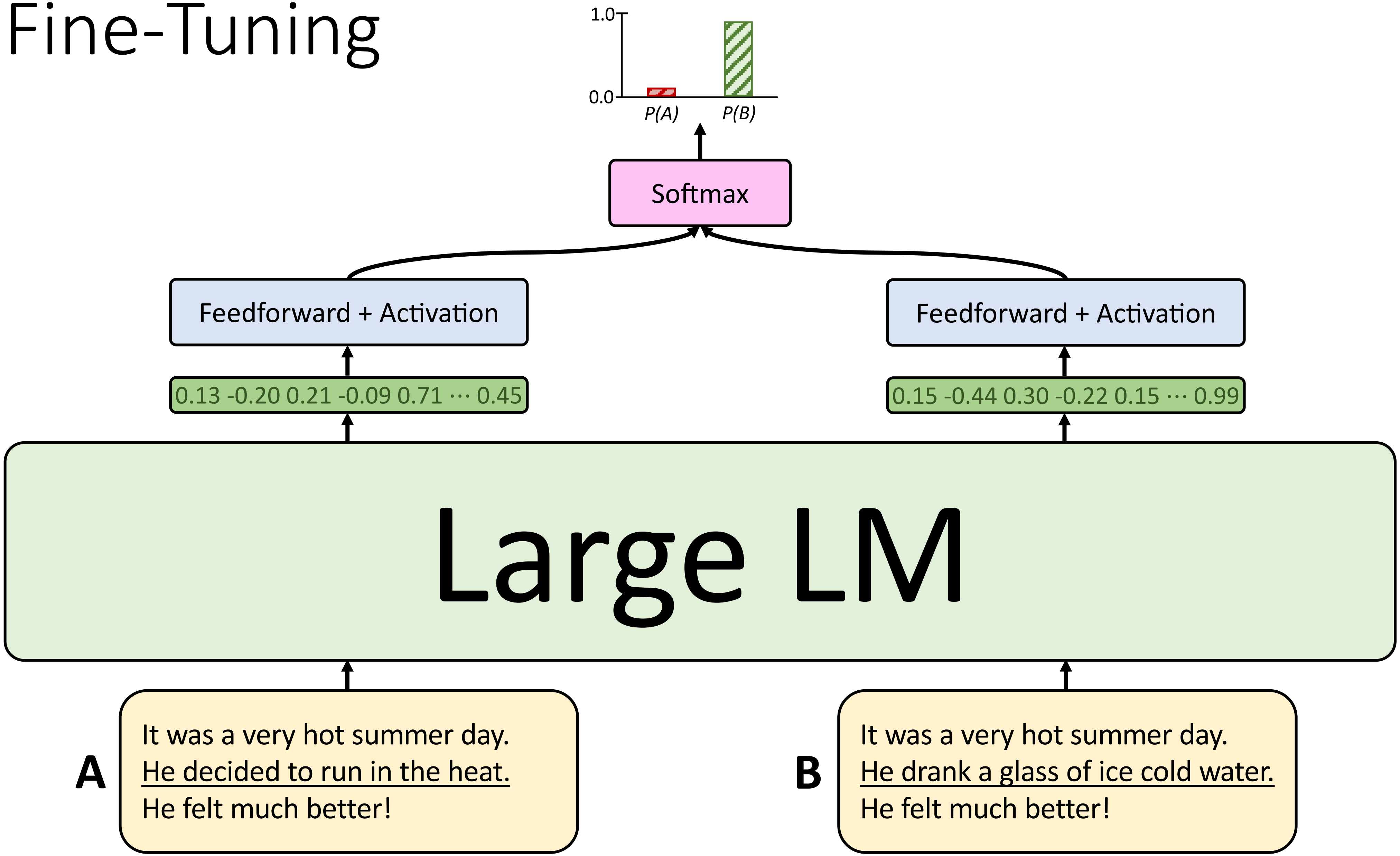
He felt much better!

He decided to run in the heat.

He drank a glass of ice cold water.

[Bhagavatula, C., Le Bras, R., Malaviya, C. et al. (2020). Abductive commonsense reasoning. In ICLR 2020.]

Fine-Tuning



1.0
0.0
P(A)
P(B)
Softmax
Feedforward + Activation
Feedforward + Activation
0.13 -0.20 0.21 -0.09 0.71 ··· 0.45
0.15 -0.44 0.30 -0.22 0.15 ··· 0.99
Large LM
A
It was a very hot summer day.
He decided to run in the heat.
He felt much better!
B
It was a very hot summer day.
He drank a glass of ice cold water.
He felt much better!

Large Language Models (LLMs)

- 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…

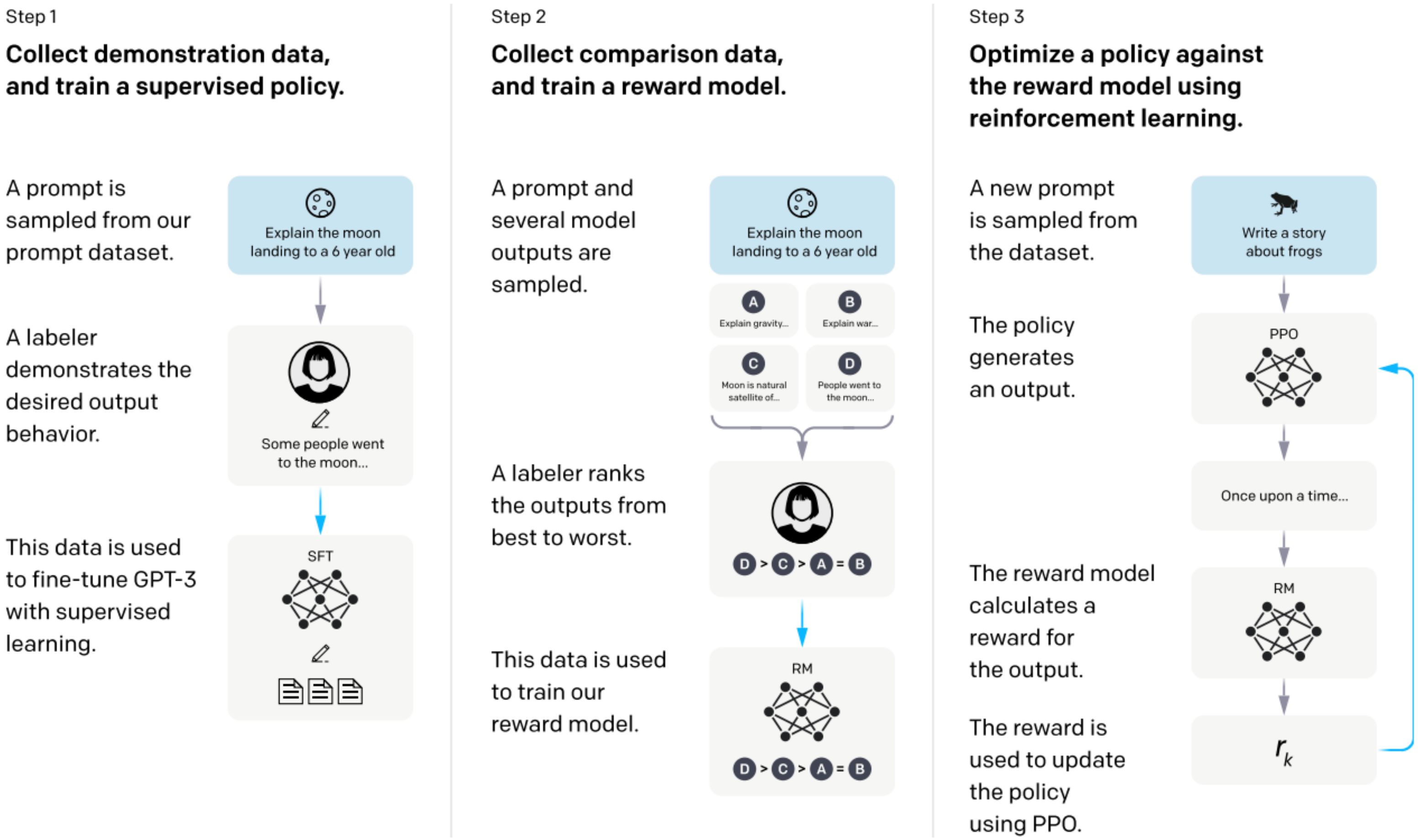
```
Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is
   hungry or if it is hot.

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.
```

Reinforcement Learning from Human Feedback (RLHF)



Long Ouyang, Jeff Wu, Xu Jiang, et al. (2022). "Training Language Models to Follow Instructions with Human Feedback." arXiv: 2203.02155.
https://chat.openai.com/

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:     ← task description
cheese =>                        ← prompt
```

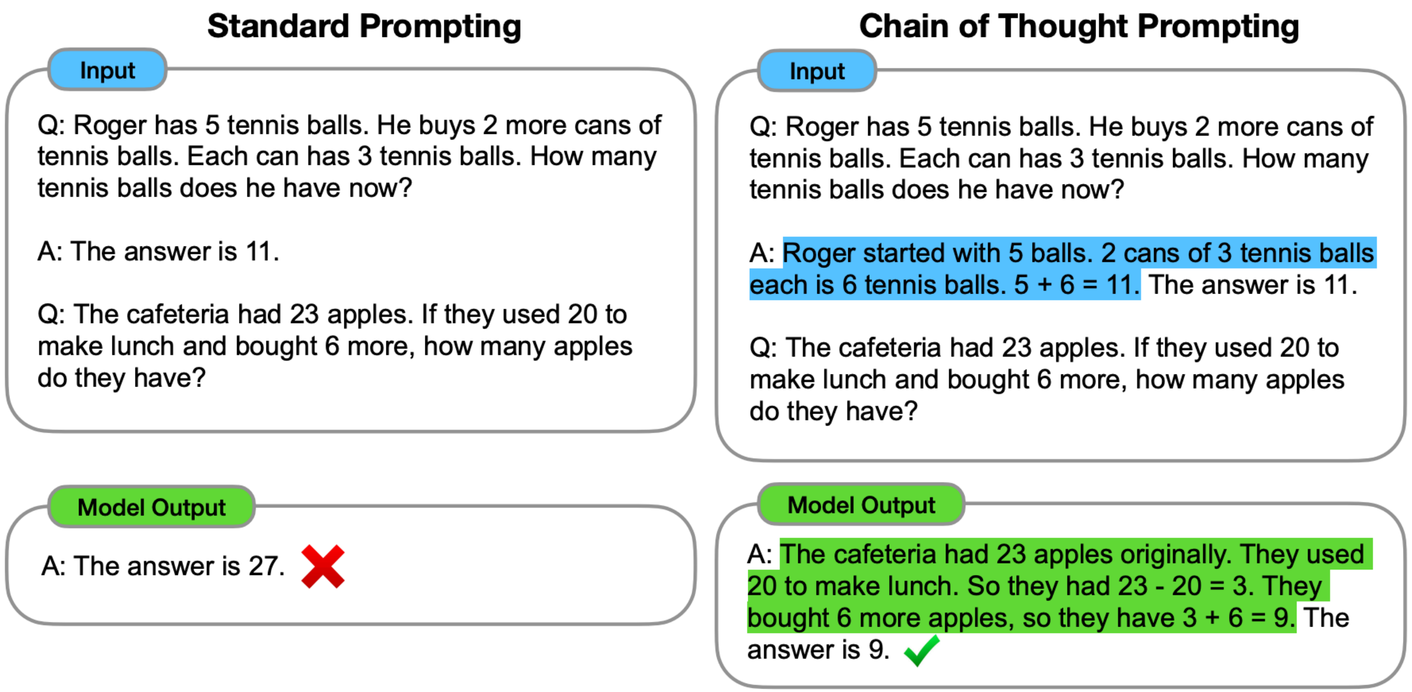
Few-shot

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

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


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

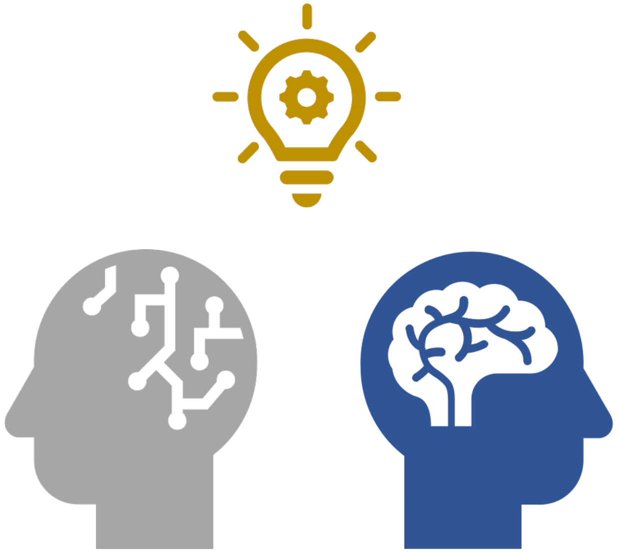
Chain-of-Thought Prompting



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

Limitations of LLMs

- Despite these advancements and impressive capabilities, LLMs still exhibit incoherent behaviors that aren't well aligned with humans
- Related to some key limitations…



Limitations of LLMs: Spurious Cues

Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.

How does the story end?

Karen became good friends with her roommate.

Karen hated her roommate.

[Schwartz, R., Sap, M., Konstas, I., Zilles, L., Choi, Y., & Smith, N.A. (2017). The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. In CoNLL 2017.]
[Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P. & Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In NAACL 2016.]

Limitations of LLMs: Data Contamination

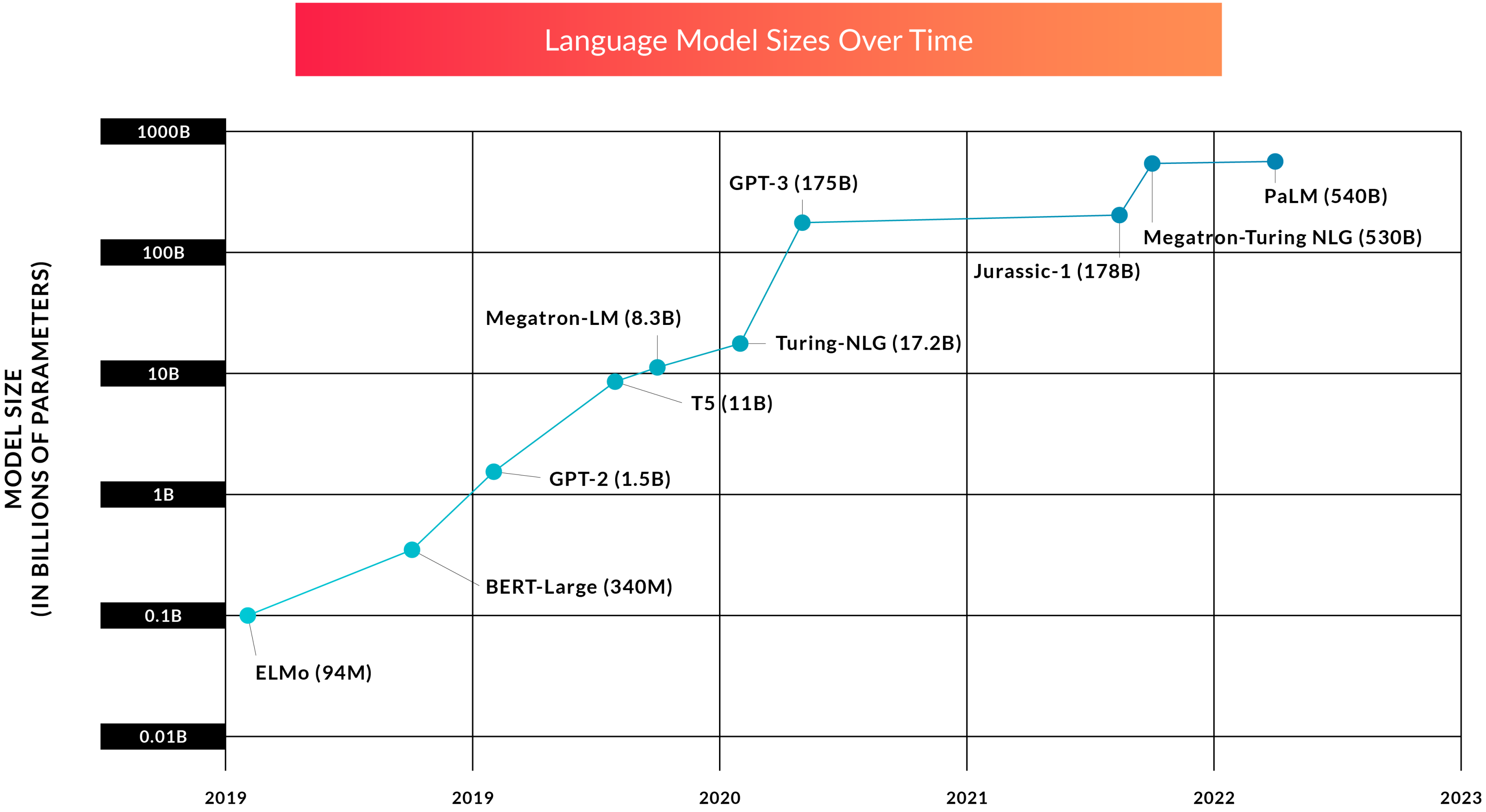
- LLMs have seen so much data in pre-training
- They may have been trained on benchmark datasets…
- Training on the test data is not an objective evaluation!

4 Measuring and Preventing Memorization Of Benchmarks

Since our training dataset is sourced from the internet, it is possible that our model was trained on some of our benchmark test sets. Accurately detecting test contamination from internet-scale datasets is a new area of research without established best practices. While it is common practice to train large models without investigating contamination, given the increasing scale of pretraining datasets, we believe this issue is becoming increasingly important to attend to.

Inbal Magar & Roy Schwartz. (2022). Data Contamination: From Memorization to Exploitation. In *ACL 2022*.
W. Shi, A. Ajith, M. Xia, et al. (2023). Detecting Pretraining Data from Large Language Models. arXiv: 2310.16789.
T.B. Brown, B. Mann, N. Ryder, et al. (2020). Language Models are Few-Shot Learners. arXiv: 2005.14165.

Limitations of LLMs: Complexity



(figure from [Vinay Iyengar])

Limitations of LLMs: Hallucination

- **Hallucination:** generation of text that is factually incorrect, nonsensical, unfaithful to inputs, or otherwise incoherent

ChatGPT

Dr. Joyce Chai is a well-known researcher in the field of artificial intelligence and natural language processing (NLP). As of my last knowledge update in January 2022, here's some information about her:

1. Background: Joyce Chai is a professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. She has made significant contributions to the areas of natural language processing, machine learning, and human-robot interaction.
2. Research Focus: Her research interests include dialogue systems, sentiment analysis, emotion recognition, and the development of AI systems that can understand and generate human language. Her work is particularly relevant to the development of conversational AI systems and chatbots.
3. Contributions: Dr. Chai has published numerous research papers and articles in top conferences and journals in the field of NLP and AI. Her research has helped advance our understanding of how AI systems can better understand and respond to human language in natural and meaningful ways.
4. Education: She holds a Ph.D. in Computer Science from the University of Washington.

Lawyer cites fake cases generated by ChatGPT in legal brief

The high-profile incident in a federal case highlights the need for lawyers to verify the legal insights generated by AI-powered tools.

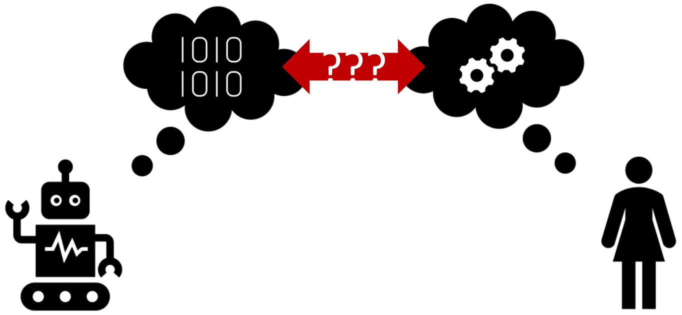
Published May 30, 2023

Lyle Moran
Reporter

Legal Dive

Z. Jio, N. Lee, R. Frieske, et al. (2023). Survey of Hallucination in Natural Language Generation. In *ACM Computing Surveys*, 55.
https://chat.openai.com/

Next Steps



- Is commonsense reasoning solved?
- Impossible to say due to these limitations 🤷‍♀️
    - We no longer have an objective way to evaluate systems’ reasoning
    - We can’t trust LLMs’ reasoning to be consistently human-aligned
- Maybe it’s not productive to focus on such broad questions…
- There’s a need for:
    1. Stronger definitions and understanding of system reasoning
    2. Broader, multidimensional metrics for evaluating system reasoning
    3. Cognitively motivated strategies for more humanlike reasoning

Storks, S., Gao, Q., & Chai, J.Y. (2019). Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. arXiv: 1904.011672 [cs.CL].

Outline

1. Introduction to Commonsense
2. Benchmarks and Resources for Commonsense
3. State of the Art and Limitations
4. **Coherent Commonsense Reasoning**

Tiered Reasoning for Intuitive Physics:

Toward Verifiable Commonsense Language Understanding

Shane Storks, Qiaozi Gao, Yichi Zhang, & Joyce Chai

(he/him)

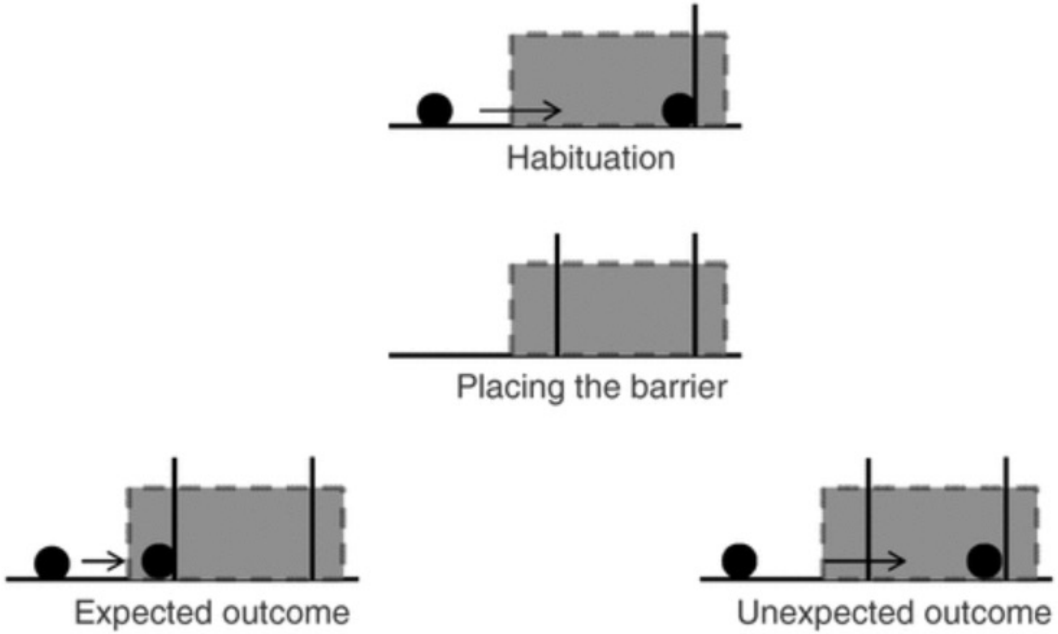
Situated Language and Embodied Dialogue (SLED)

University of Michigan, Computer Science and Engineering Division

sstorks@umich.edu

Findings of EMNLP 2021 Long Paper

Physical Commonsense



(Parents.com)

(dreamstime)

Bliss, J. (2008). Commonsense reasoning about the physical world. In *Studies in Science Education*, 44(2): 123-155.
Lake, B., Ullman, T.D., Tenenbaum, J.B., & Gershman, S.J. (2017). Building machines that learn and think like people. In *Behavioral and Brain Sciences,* 40.
Hespos, S.J. & vanMarle, K. (2011). Physics for infants: characterizing the origins of knowledge about objects, substances, and number.

Tiered Reasoning for Intuitive Physics (`TRIP`)

- We can't trust LLM outputs are coherent - need to show their work!
- Introduce a dataset providing multi-tiered, human-annotated reasoning processes for physical commonsense:
  - Low-level, concrete physical states
  - High-level end task of plausibility classification

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

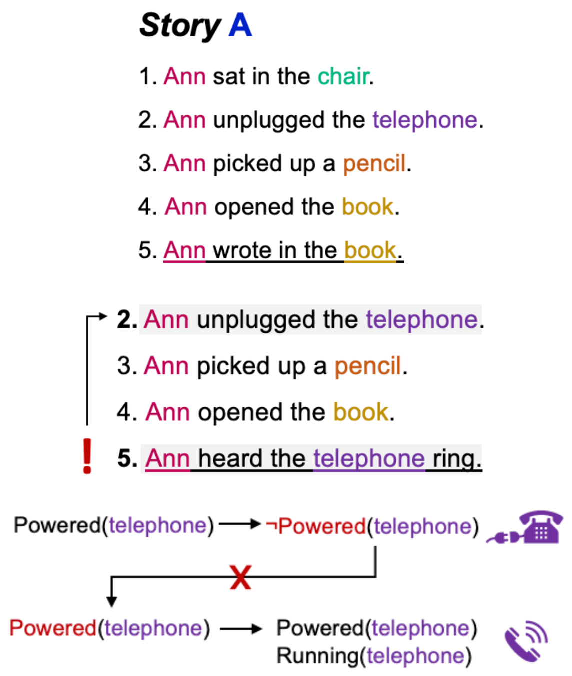
Physical states:

Powered(telephone) → ¬Powered(telephone)

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

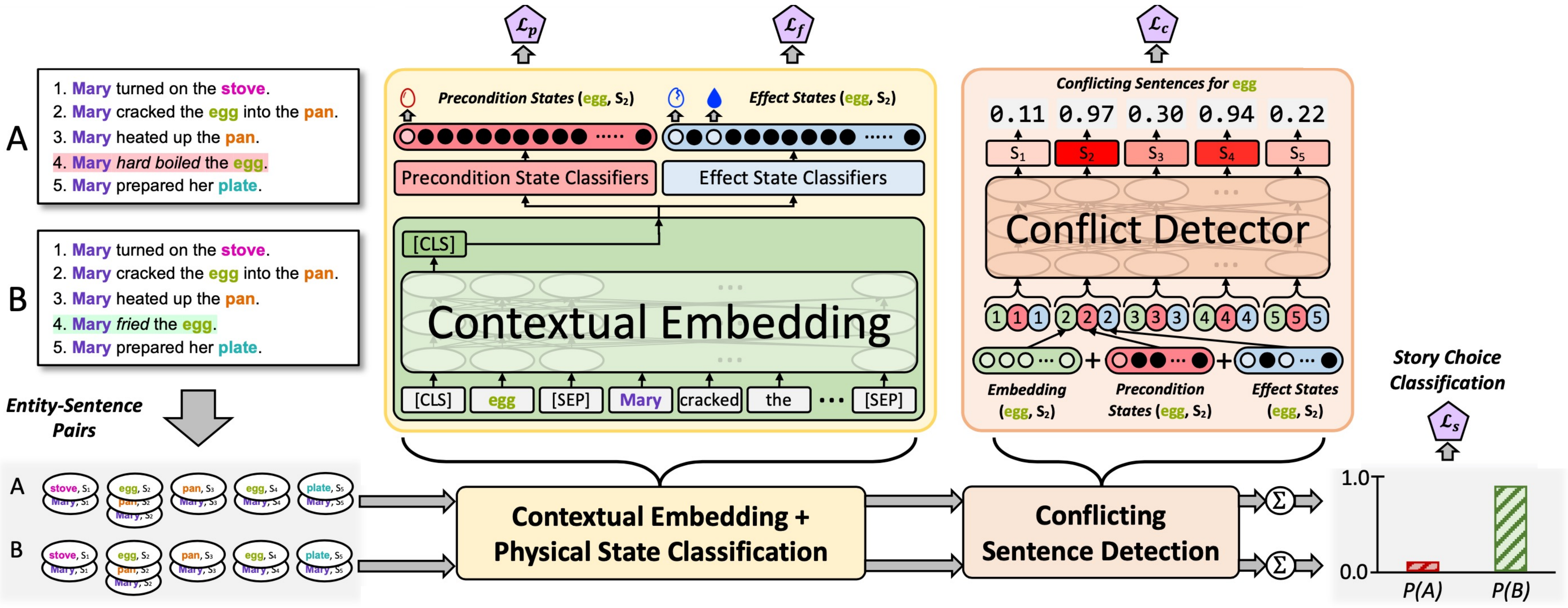
X

Evaluation Metrics



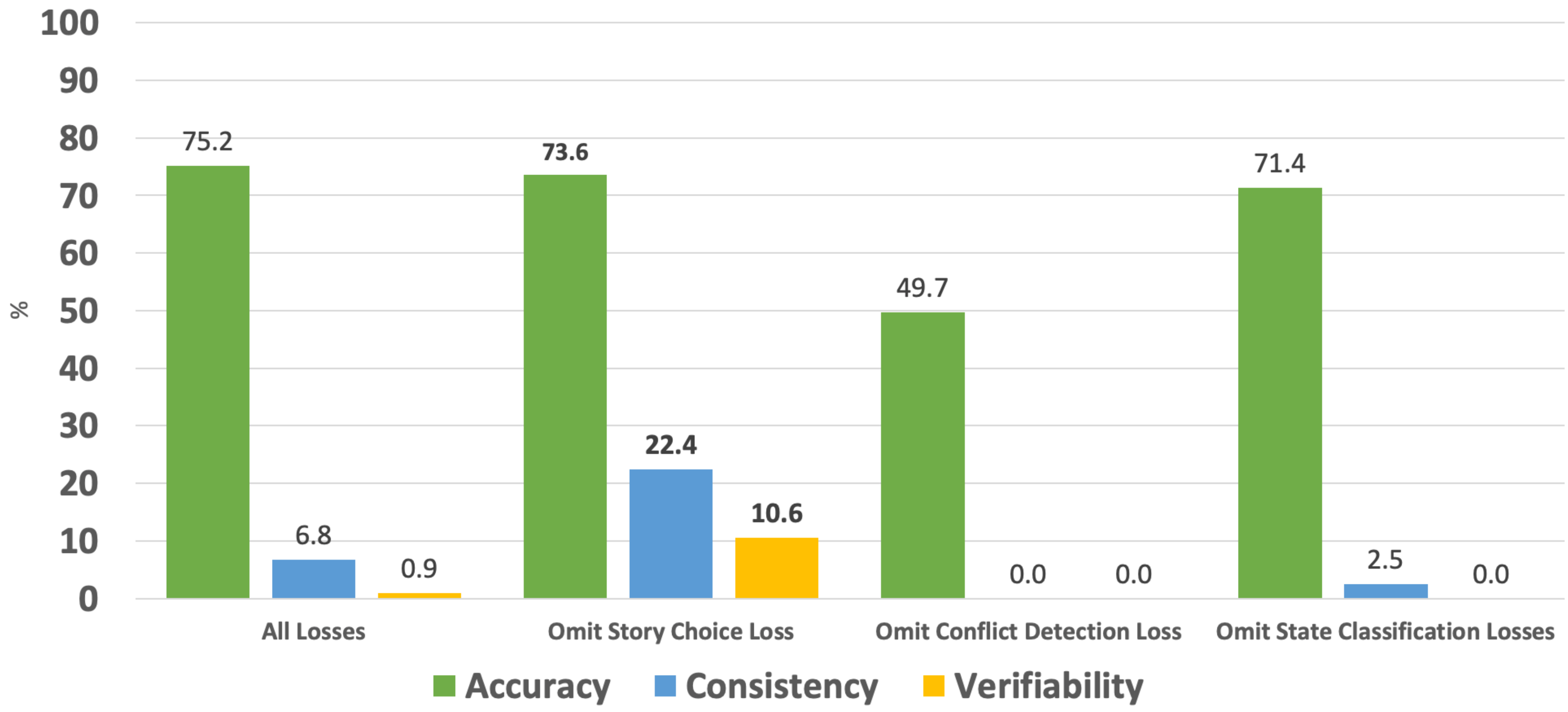
Metric	Story Choice	Conflicting Sentences	Physical States
Accuracy	✓		
Consistency	✓	✓	
Verifiability	✓	✓	✓

Tiered Baseline



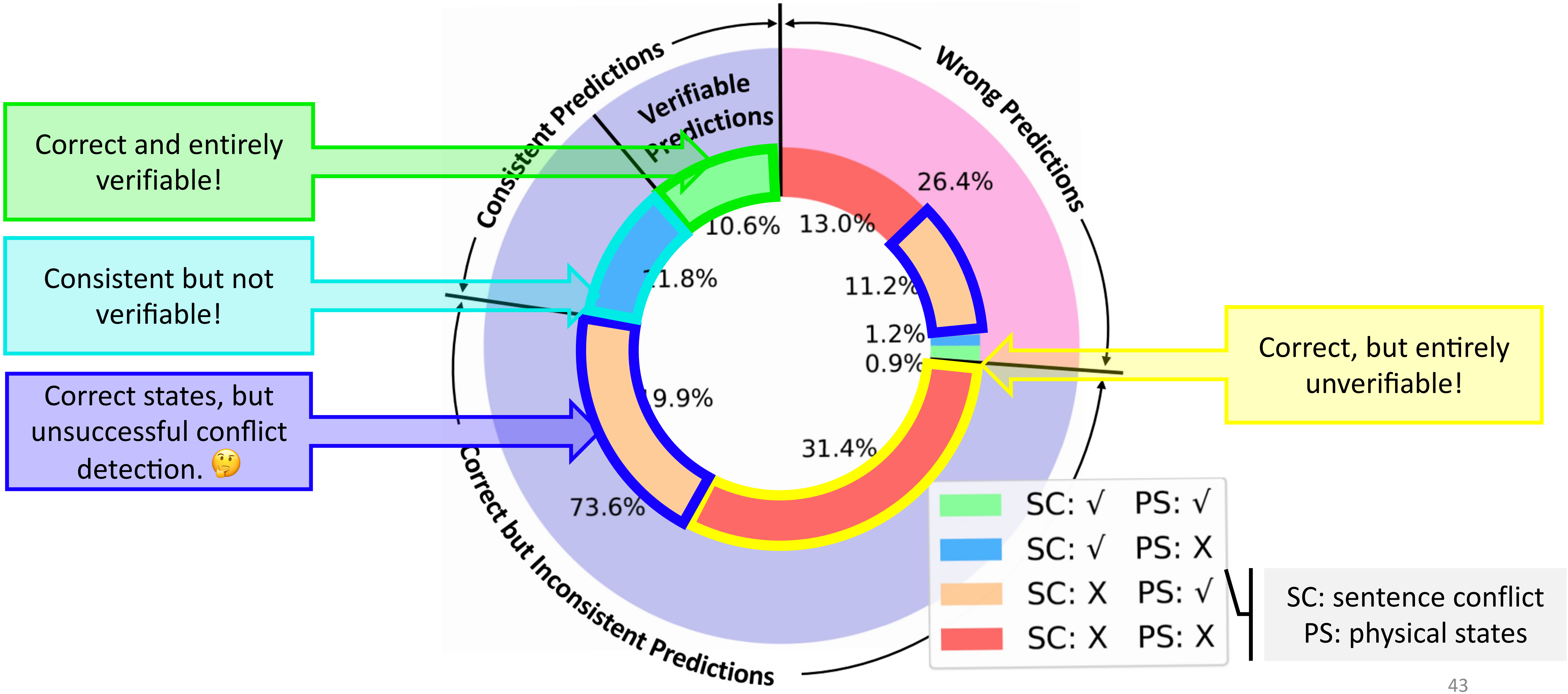
$$\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_f \mathcal{L}_f + \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s$$

RoBERTa Baseline Results on TRIP

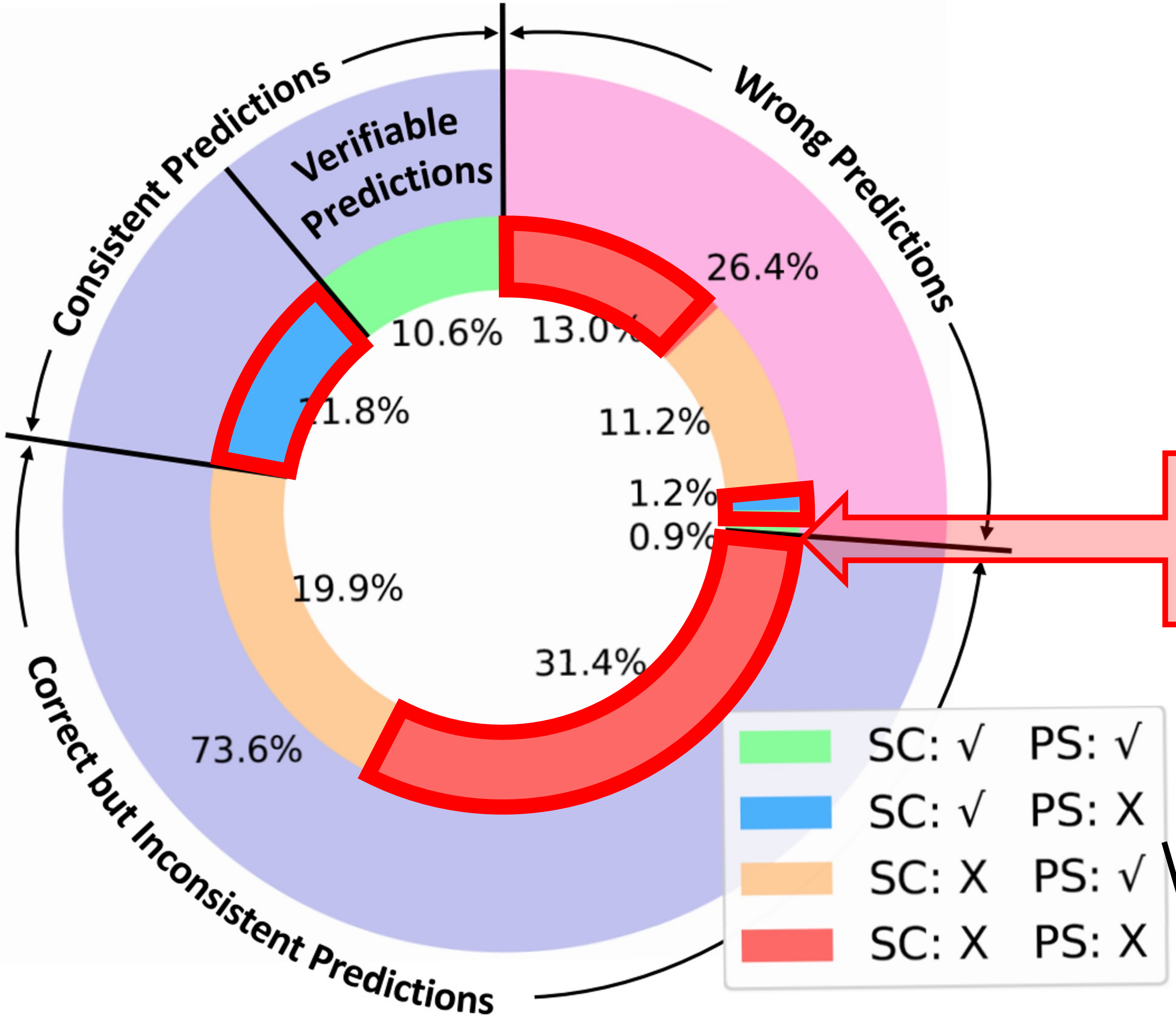


Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692.

Error Distribution



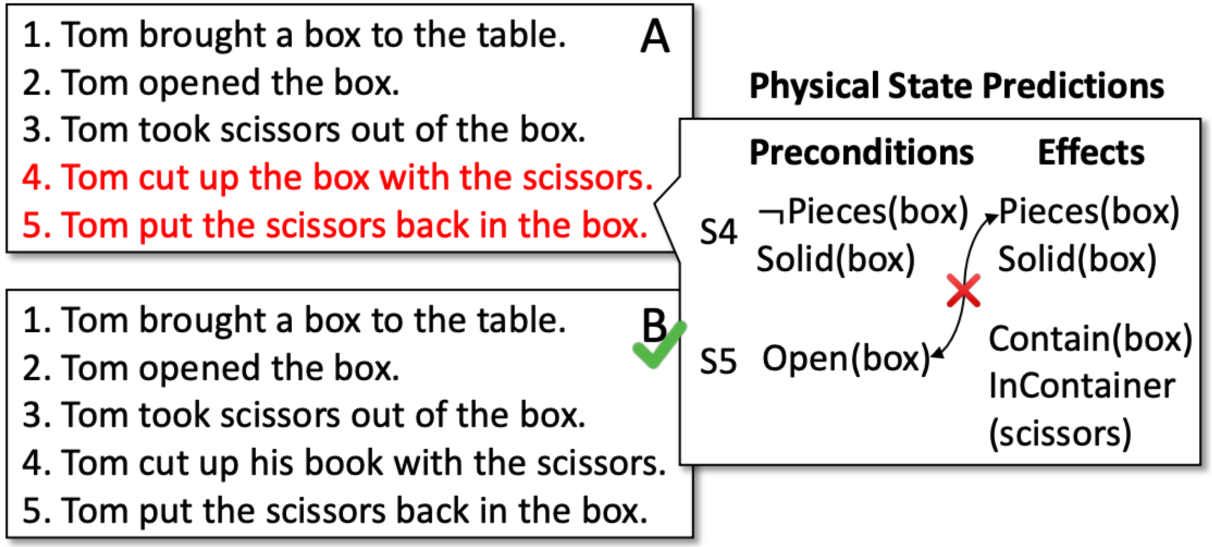
Baseline Results



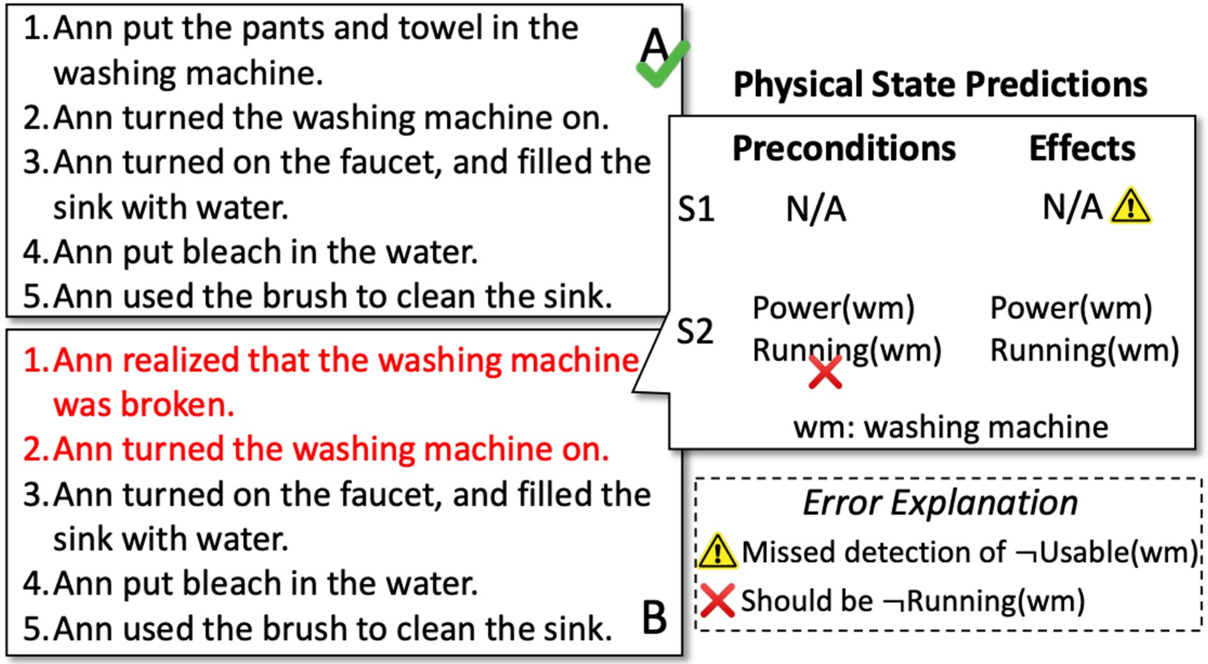
Incorrect physical states!

SC: sentence conflict
PS: physical states

Sample System Outputs



(a) A verifiable prediction.



(b) A consistent but not verifiable prediction.

Summary

- LMs can easily get high accuracy when fine-tuned on TRIP
- But they **struggle to learn verifiable reasoning strategies** when trained as tiered, verifiable reasoning systems!

From Heuristic to Analytic:

Cognitively Motivated Strategies for Coherent Physical Commonsense Reasoning

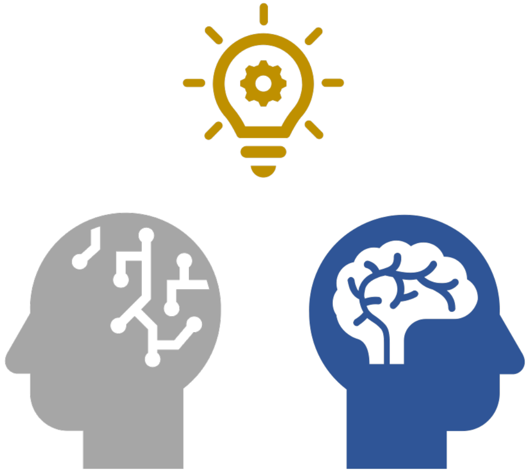
Zheyuan Zhang[1*] Shane Storks[1*] Fengyuan Hu[1] Sungryull Sohn[2]
Moontae Lee[2] Honglak Lee[1,2] Joyce Chai[1]

[1]University of Michigan
[2]LG AI Research
*Equal Contribution

EMNLP 2023 Long Paper

Strengths and Weaknesses of PLM Cognition

- Pre-trained language models (PLMs) have recently attracted attention for seemingly human-like reasoning capabilities
- Spurious behaviors like hallucination lead to incoherent behaviors
- Using them to reason about the physical world, e.g., in embodied AI, may be especially dangerous!
- How to enable more coherent, humanlike reasoning?

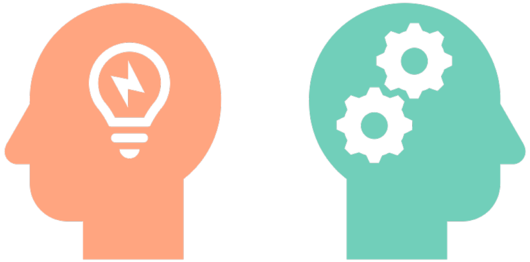


https://chat.openai.com/
OpenAI. GPT-4 Technical Report. arXiv: 2303.08774.

Dual Processes of Human Cognition

A line of work theorizes two processes in human reasoning:

- **Heuristic:** fast, intuitive
  - Extract most relevant info from context, provide quick intuition for decisions
- **Analytic:** slow, deliberative
  - Further operate on relevant info to perform inference and rationalize
- 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

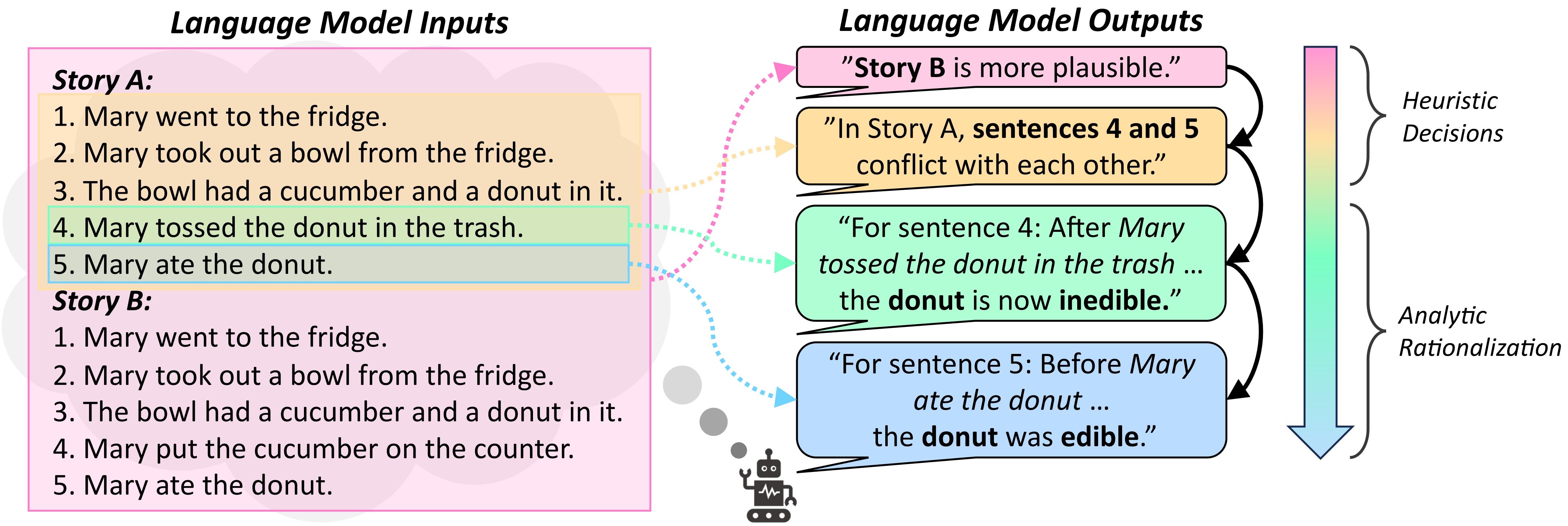
Story A:
1. Air is brought in through the mouth.
2. Passes through the lungs.
3. And into the bronchial tissue.
4. The *carbon dioxide* is removed.
5. The lungs bring the oxygen to the rest of the body.

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**

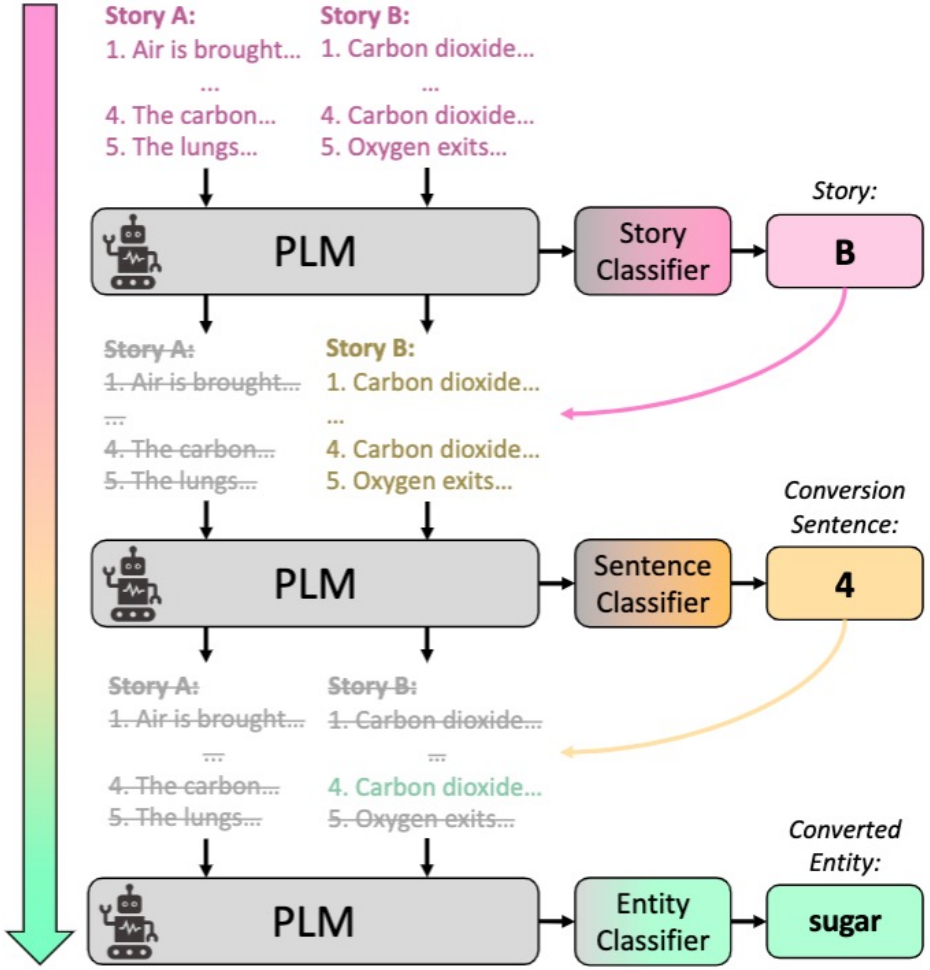
Shane Storks, Qiaozi Gao, Yichi Zhang, & Joyce Chai. 2021. Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding. *Findings of EMNLP 2021.*
Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, & Peter Clark. 2018. Tracking State Changes in Procedural Text: a Challenge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

Heuristic-Analytic Reasoning (HAR)



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**

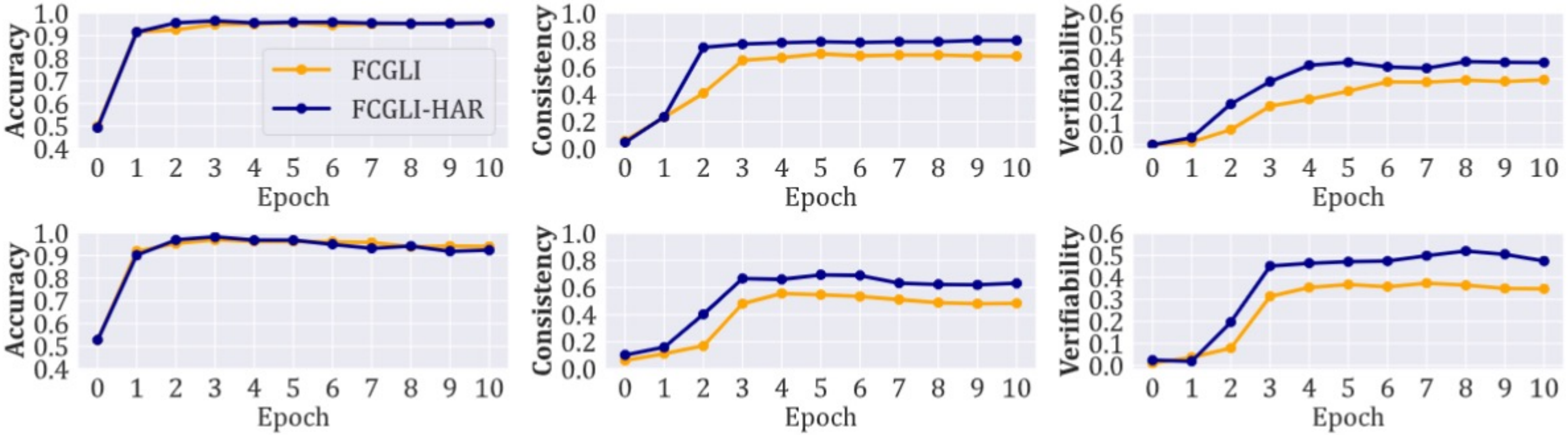
Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692.
Kaixin Ma, Filip Ilievski, Jonathan Francis, et al. 2022. Coalescing Global and Local Information for Procedural Text Understanding. In *COLING 2022.*
Kyle Richardson, Ronen Tamari,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

Learning Curves in Fine-Tuning

Consistency and verifiability converge 1-2 epochs faster in FCGLI-HAR.

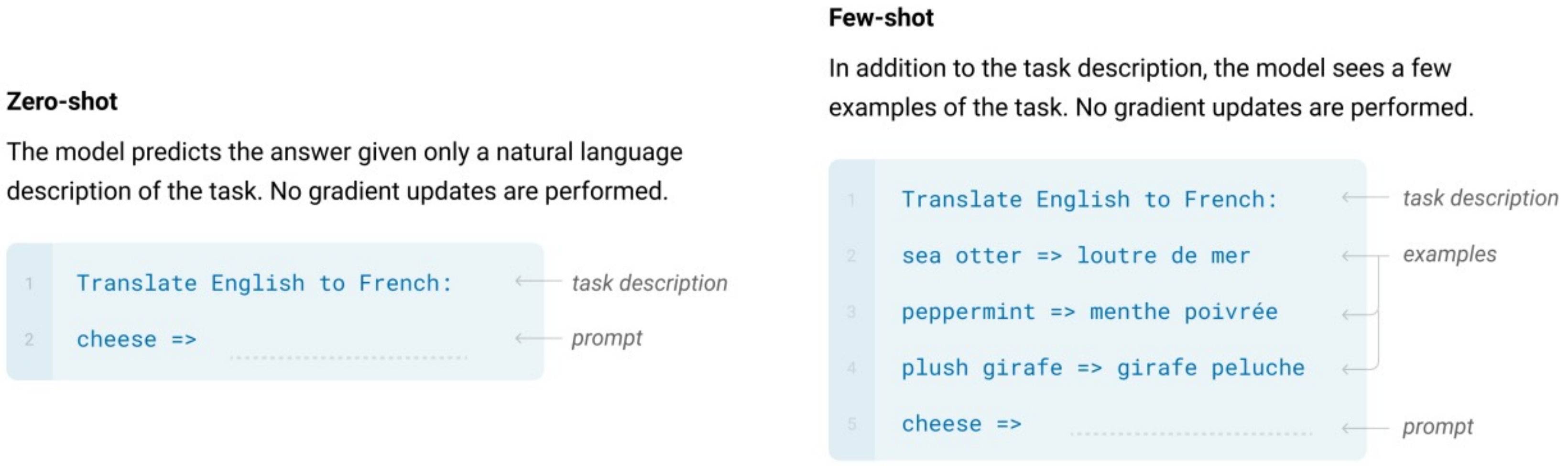
FCGLI

FCGLI-HAR



Limitations of PLM Fine-Tuning

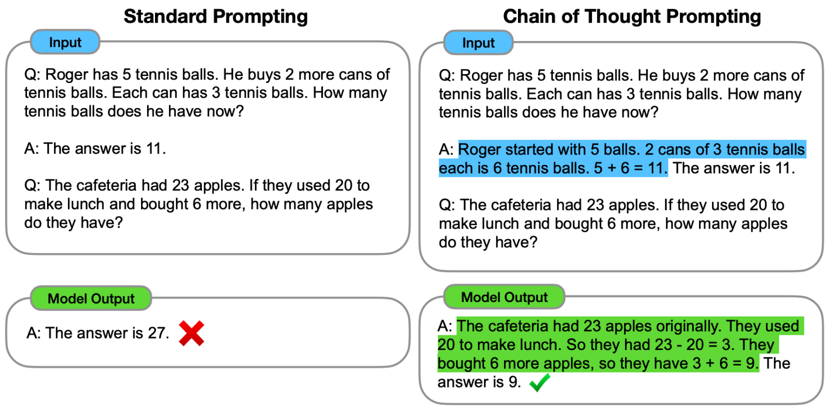
- PLM fine-tuning requires expensive training on a large amount of in-domain data, which may sacrifice generalizability
- Instead, recent work applies PLMs directly to downstream tasks through zero-shot prompting and in-context learning



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

Limitations of In-Context Learning Methods

- We have tricks like chain-of-thought (CoT) to help break down complex tasks into separate reasoning steps
  - Condition PLM with these to reach final answer
- Physical state prediction (most complex step) is difficult to break down further
- Can (heuristic) story and sentence selections serve as useful information for (analytic) physical state prediction in this setting?



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

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 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.

Let's think step by step...

… In Story A, Mary takes a bowl out of the fridge that has a cucumber and a donut in it. She then proceeds to put the cucumber… Therefore, …

Story B is more plausible.

… In Story B, Mary also took the bowl out of the fridge and found a cucumber and a donut inside. However, she did not toss the donut… Therefore, …

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

…The conflicting physical states are that in Story A, Mary throws the donut in the trash, and then, Mary eats the donut… Therefore, …

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)

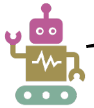
Story A:
1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary tossed the donut in the trash.
5. Mary ate the donut.

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

Approach	TRIP			Tiered-ProPara		
	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

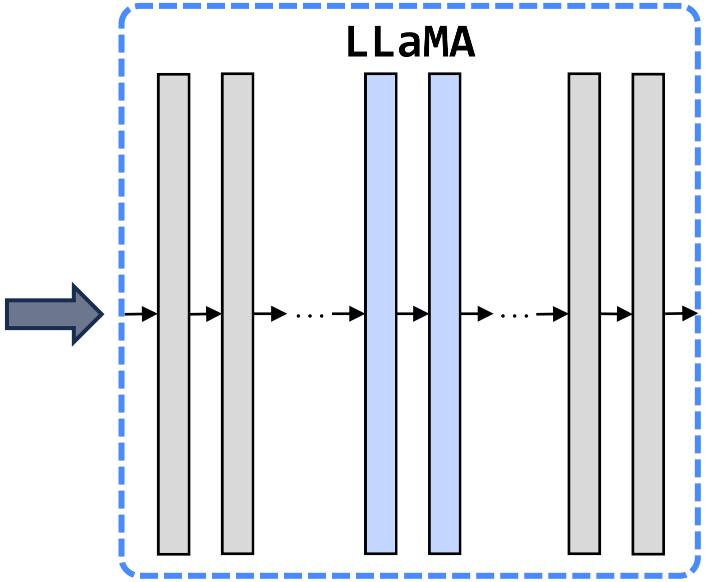
Approach	TRIP			Tiered-ProPara		
	Acc.	*Cons.*	*Ver.*	Acc.	*Cons.*	*Ver.*
ICL-U	70.4	42.3	14.8	51.2	3.8	1.4
ICL-CoT	74.6	42.3	19.7	57.3	9.4	4.2
ICL-HAR	55.6	**44.4**	**35.2**	41.8	**17.8**	**13.1**

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

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.

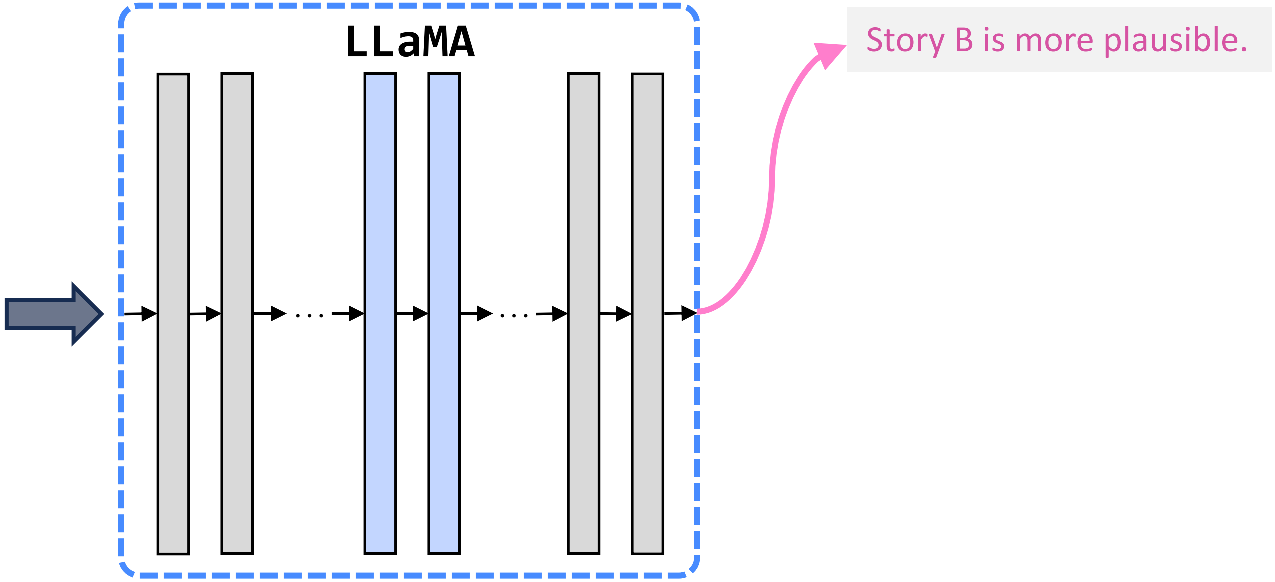


Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo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

Attention Analysis

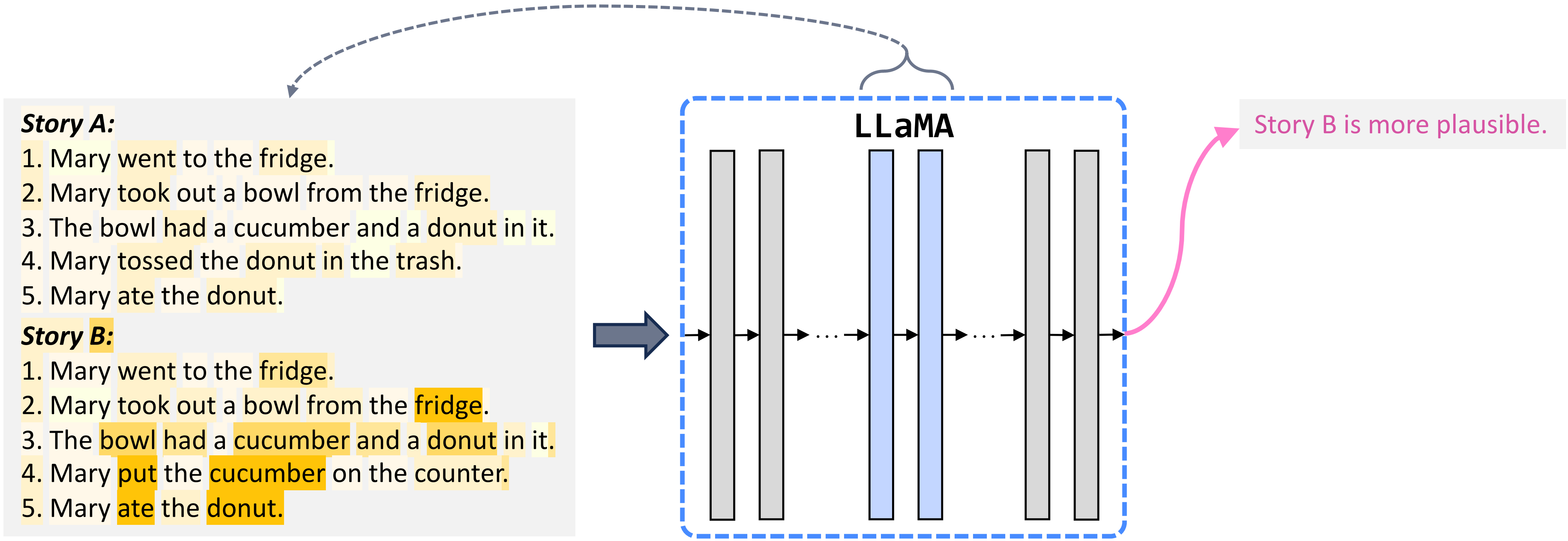
Story A:
1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary tossed the donut in the trash.
5. Mary ate the donut.

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.



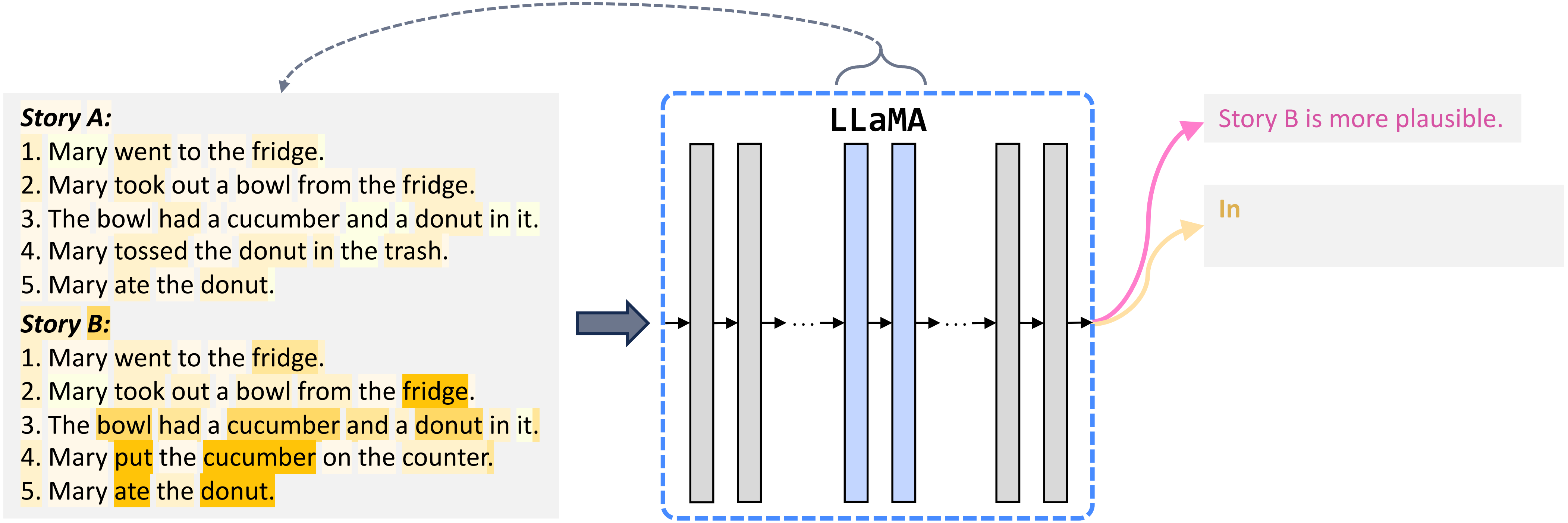
Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo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

Attention Analysis



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

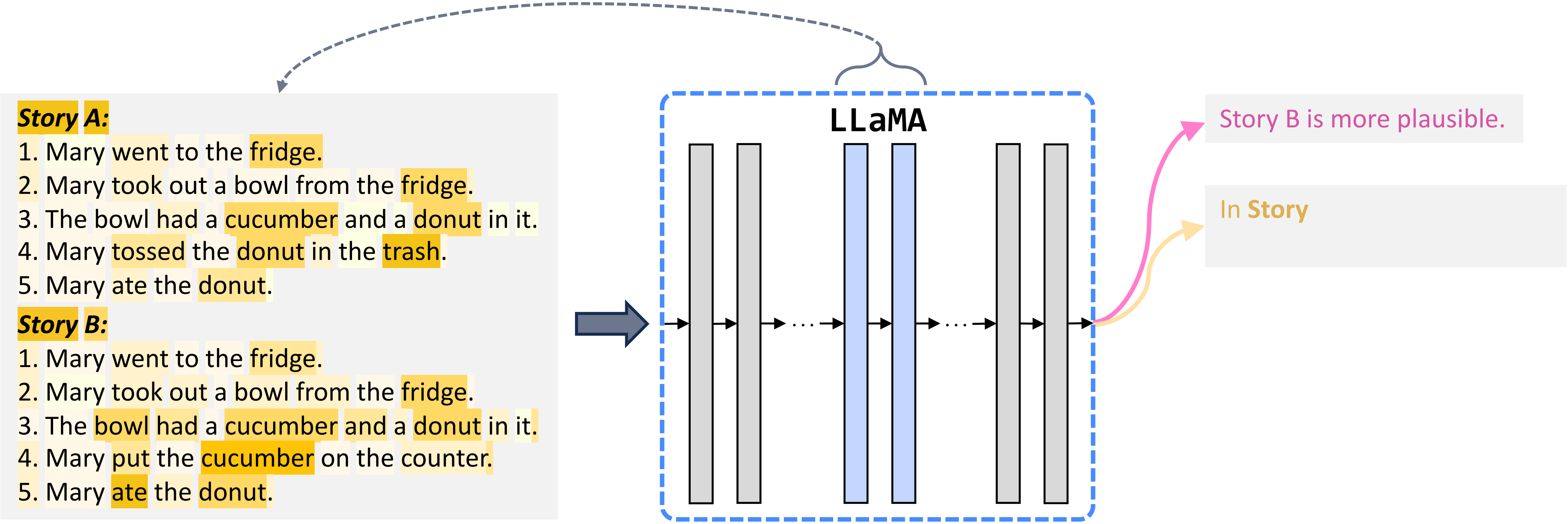


Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*

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



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*.

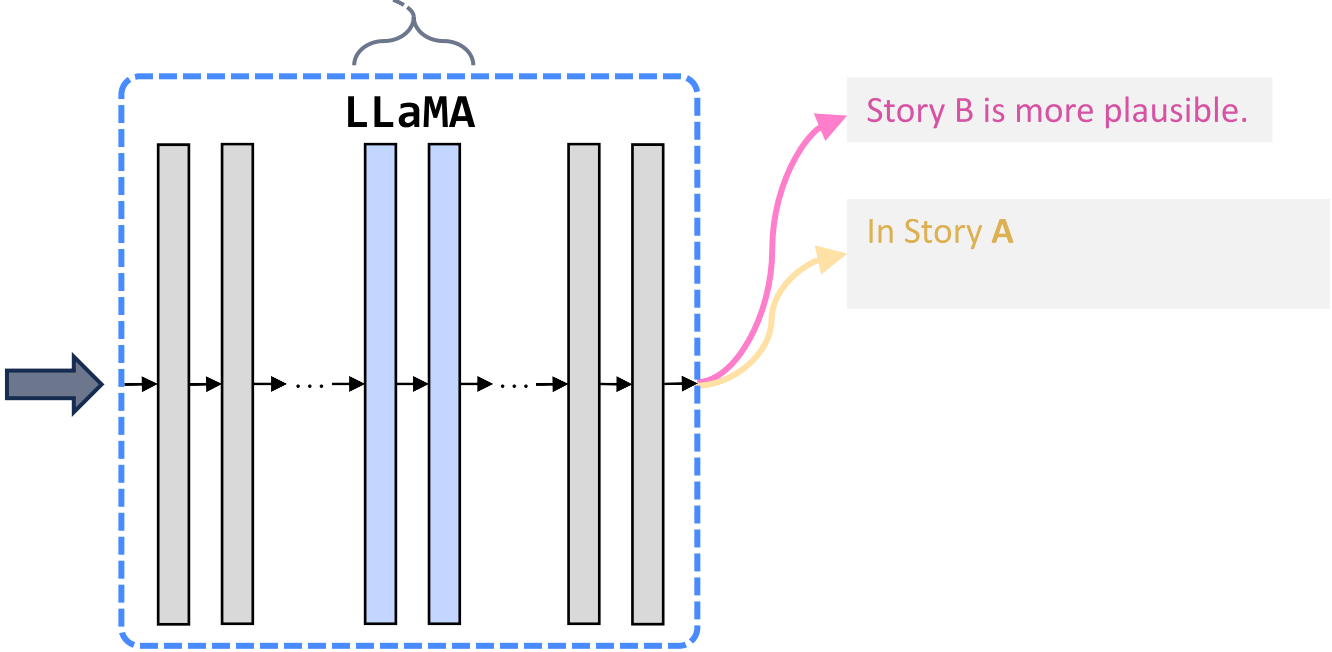
Attention Analysis

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.

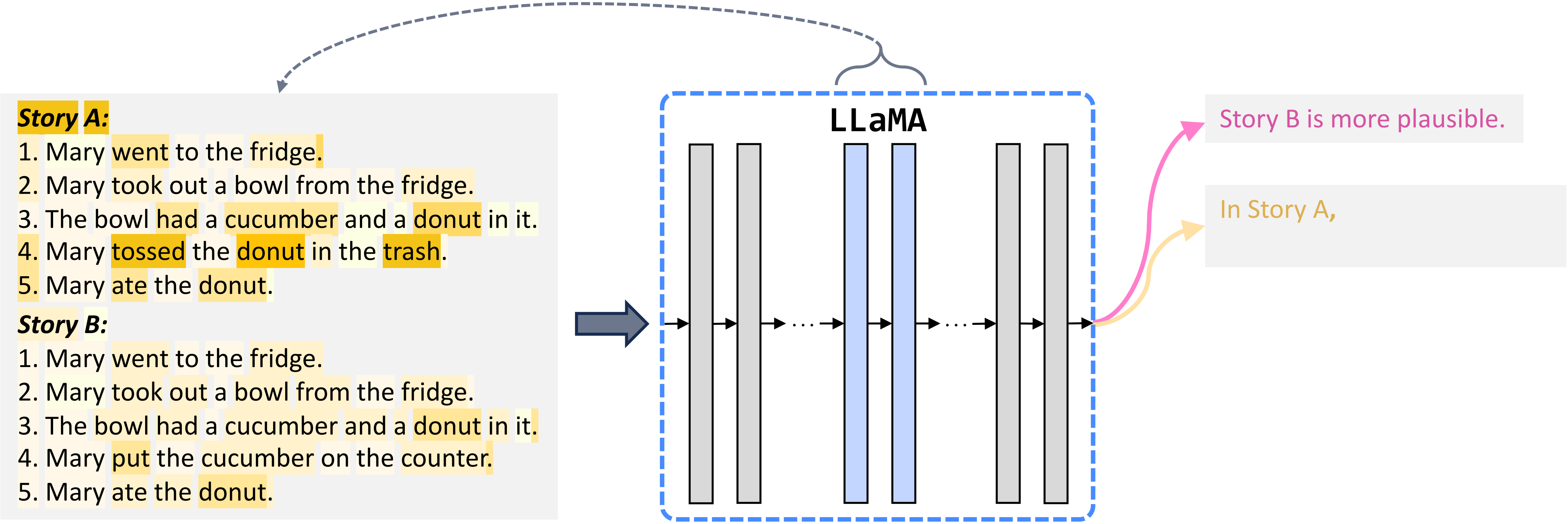
Story B:

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.



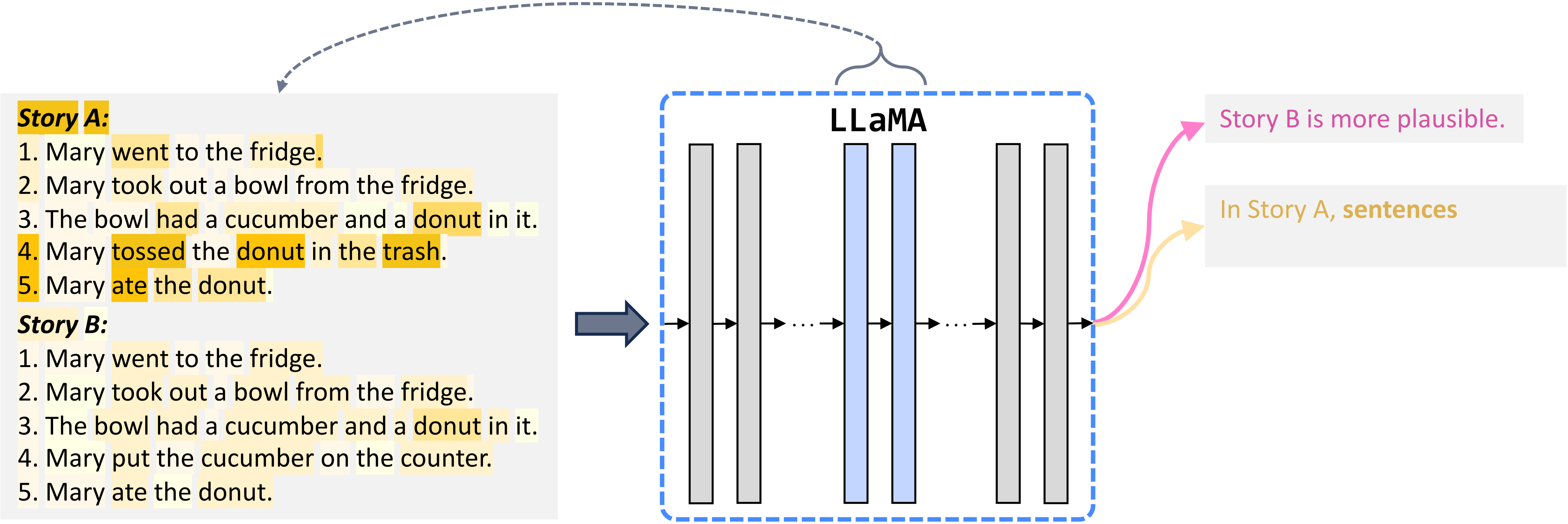
Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo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

Attention Analysis



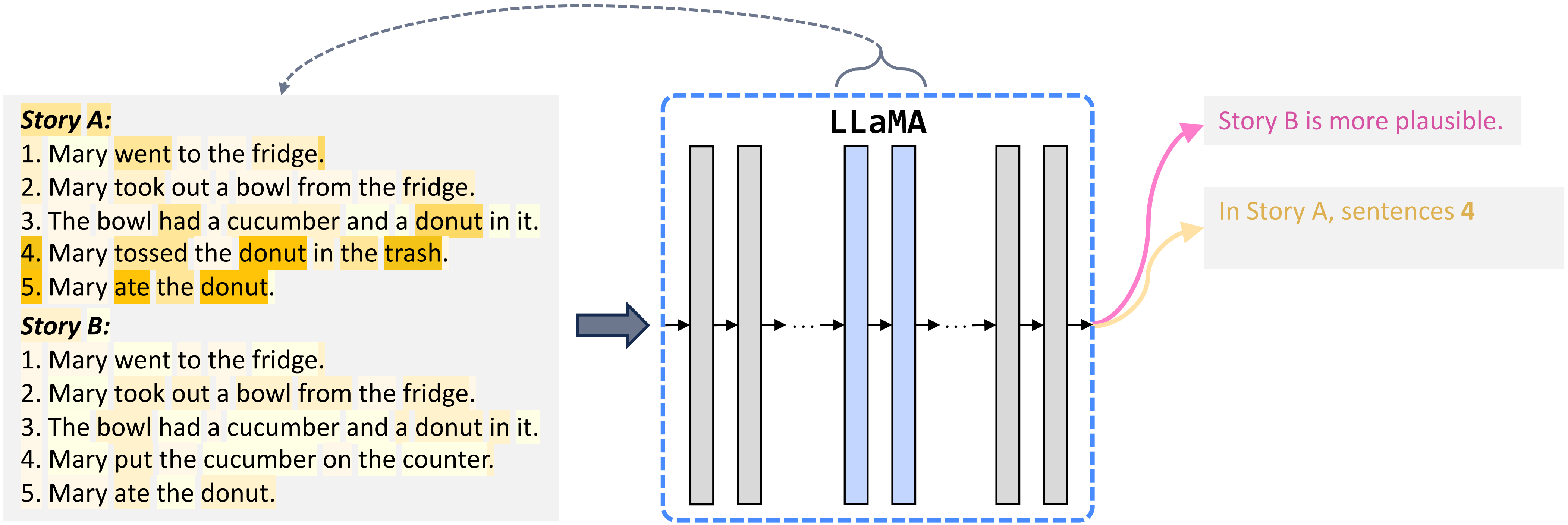
Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo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

Attention Analysis



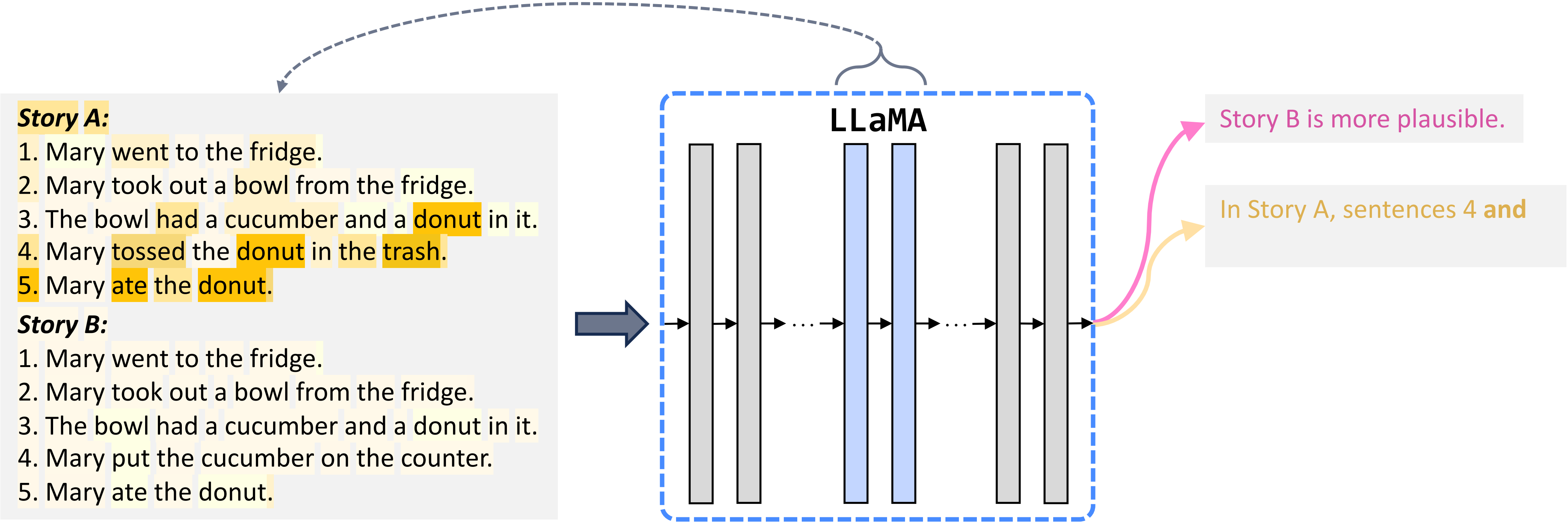
Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo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

Attention Analysis



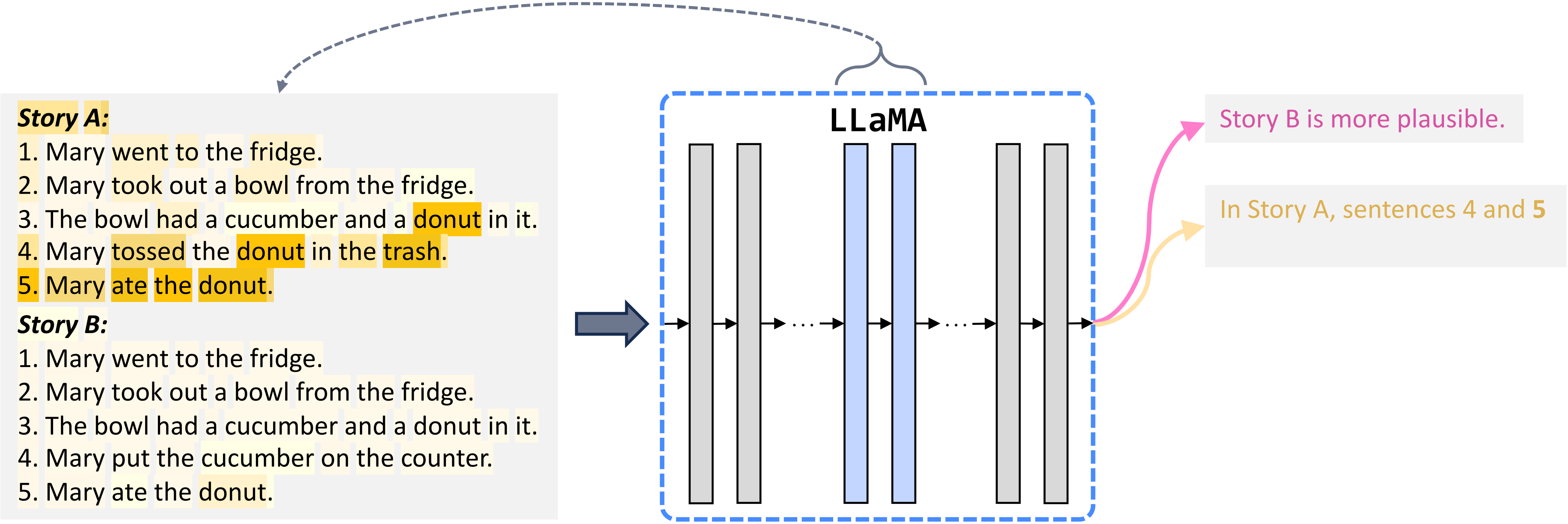
Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo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

Attention Analysis



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



Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*

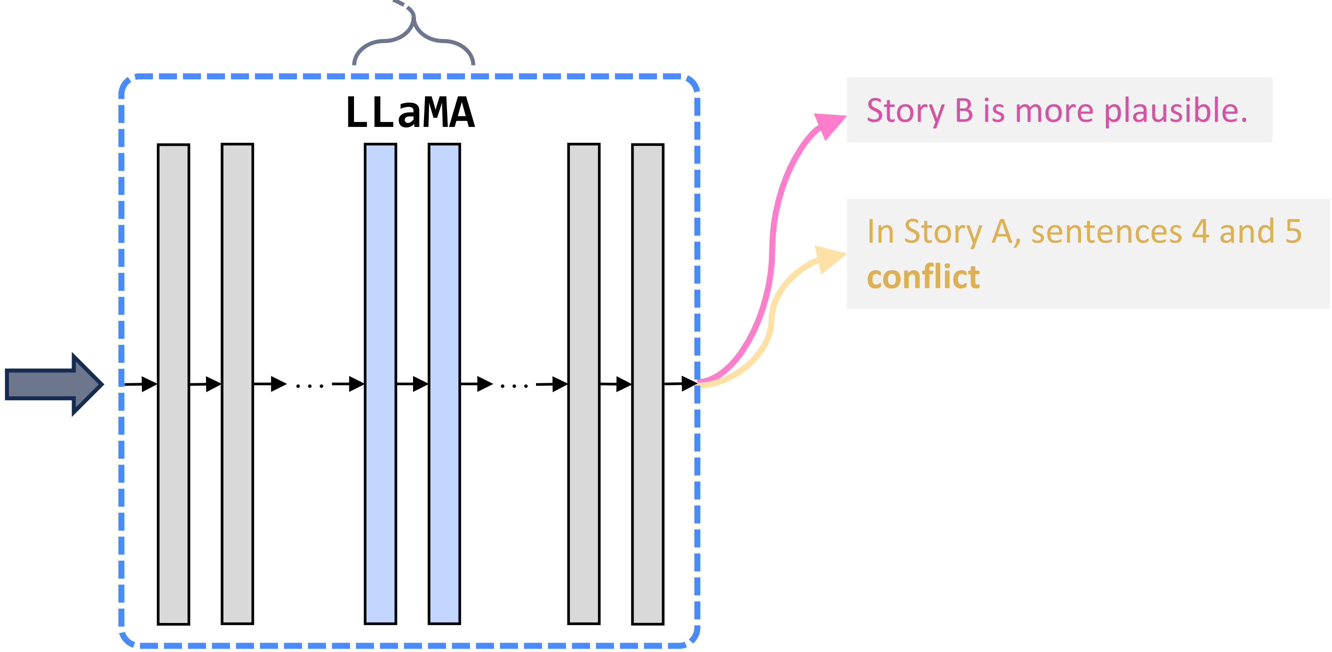
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv: 2302.13971*.

Attention Analysis

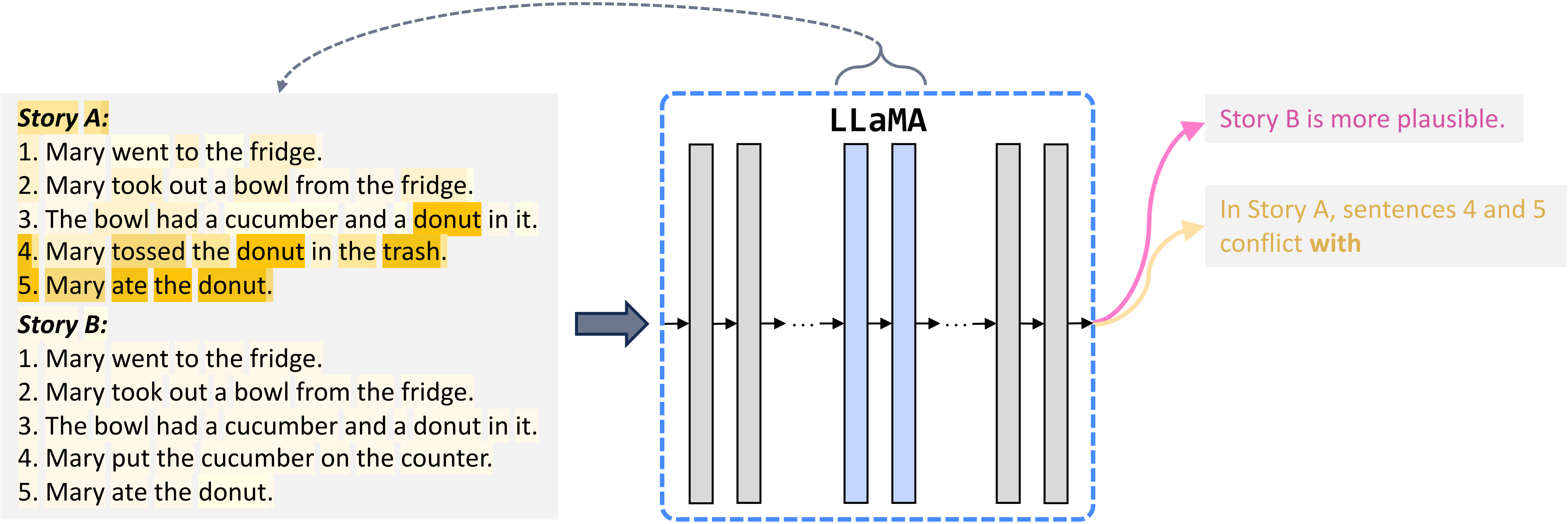
Story A:
1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary tossed the donut in the trash.
5. Mary ate the donut.

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.



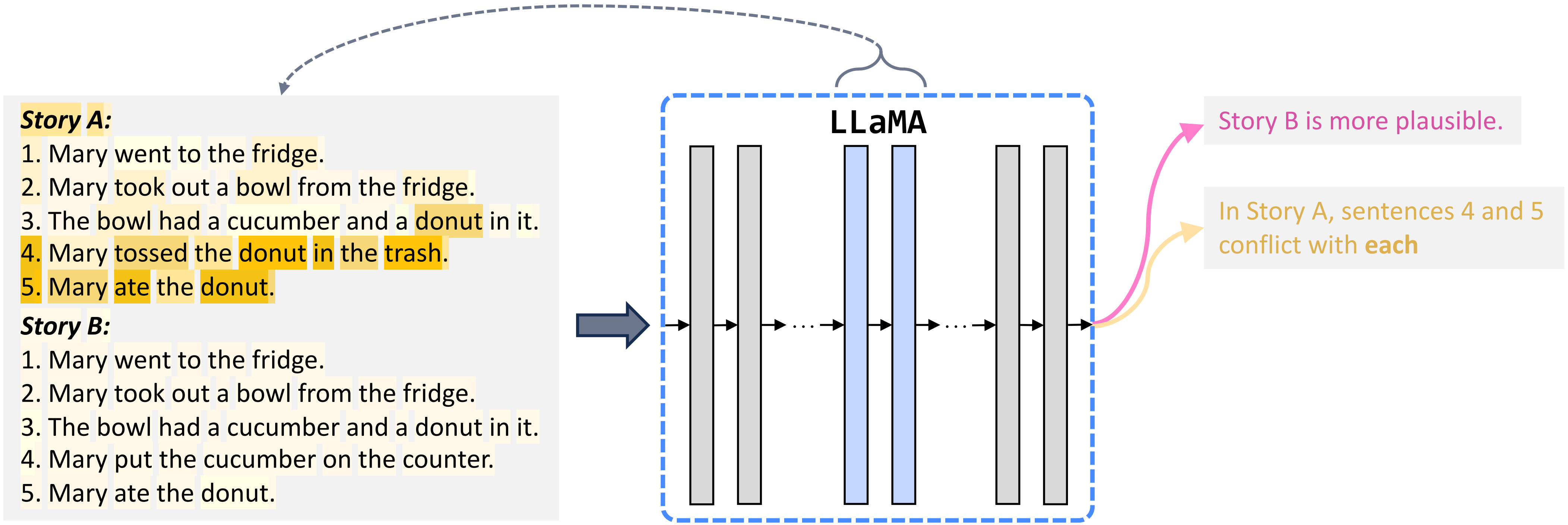
Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo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

Attention Analysis



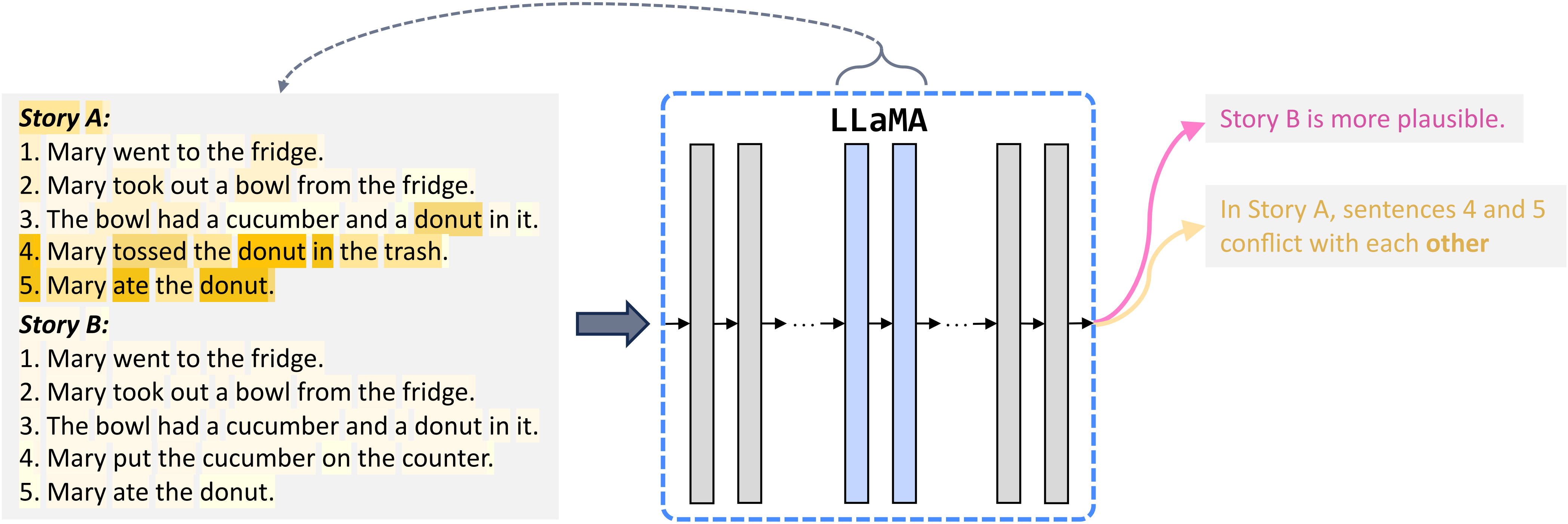
Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo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

Attention Analysis



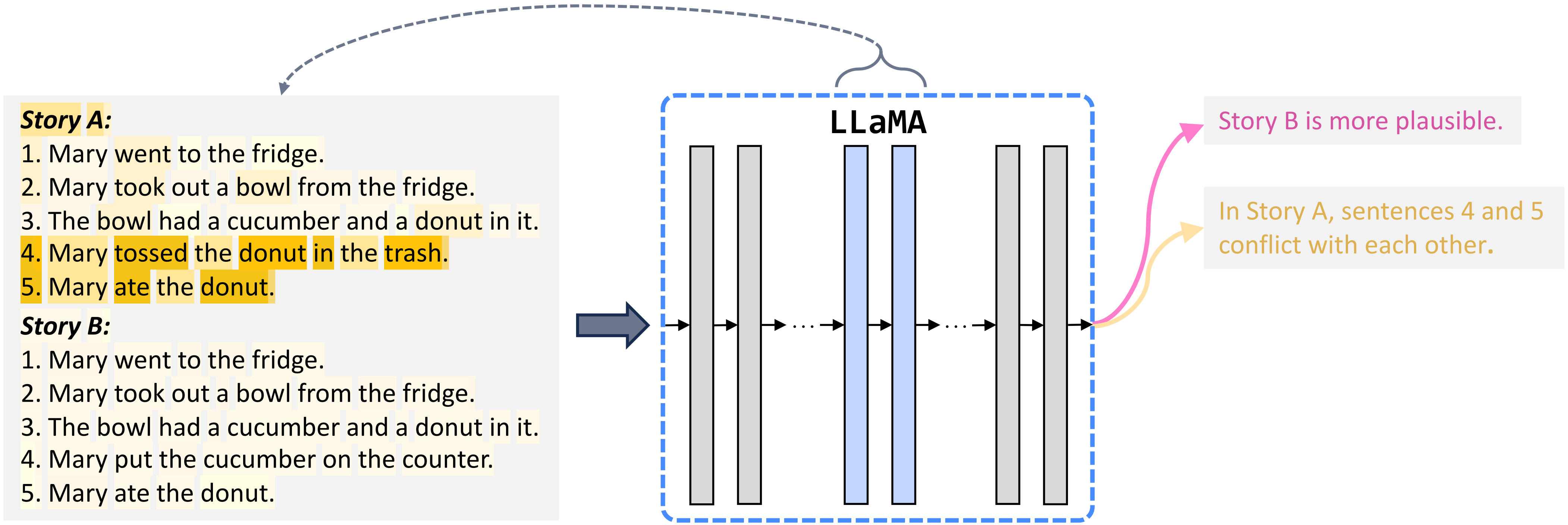
Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*.
Hugo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

Attention Analysis



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

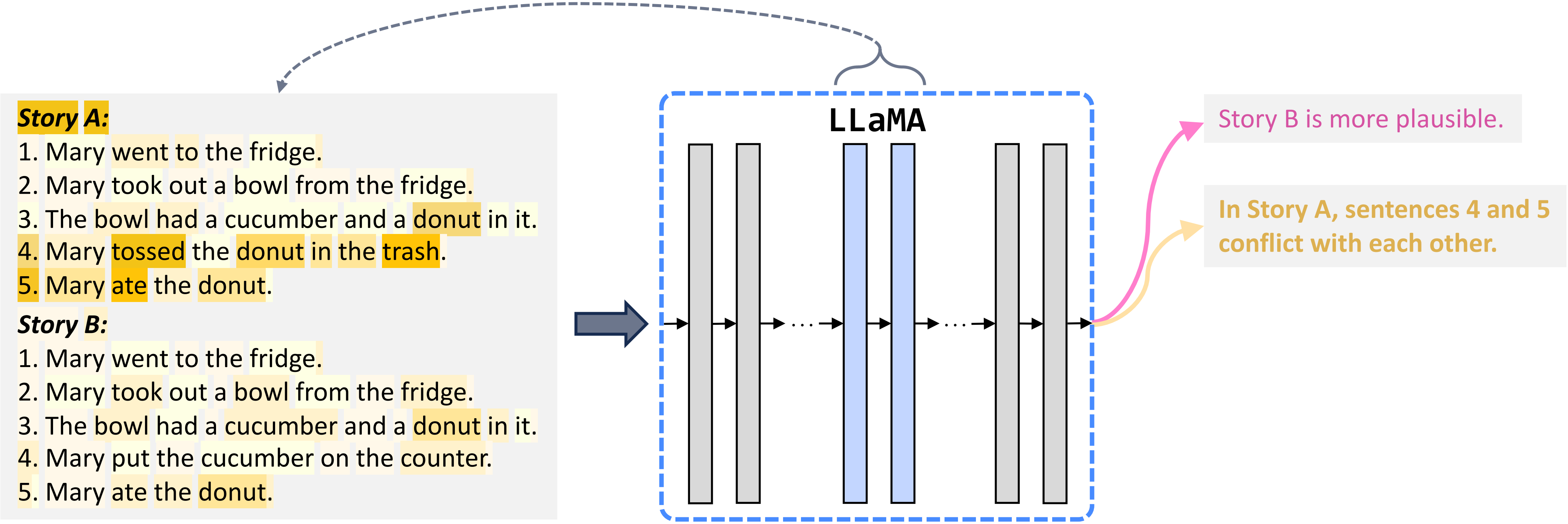


Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*

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



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

Story A:

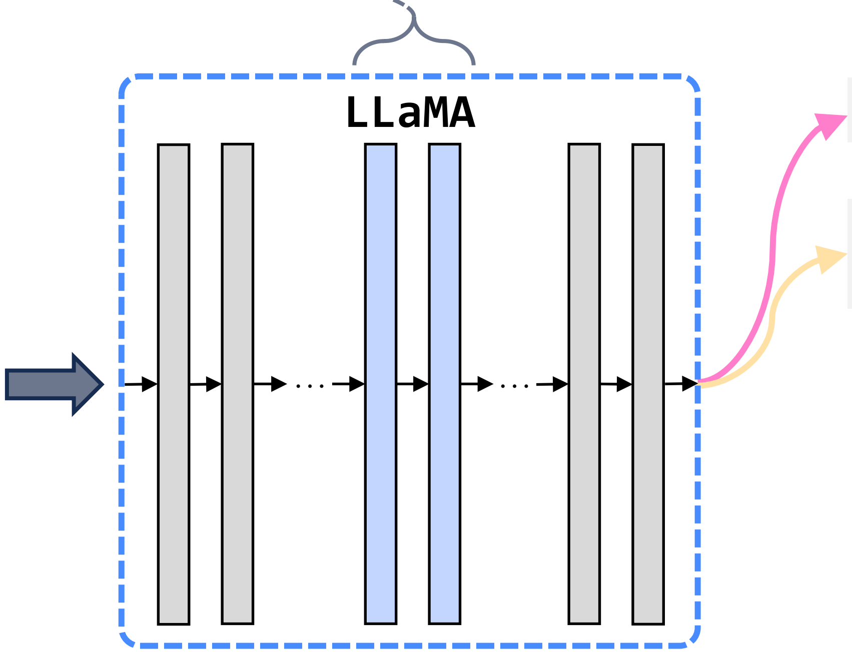
1. Mary went to the fridge.
2. Mary took out a bowl from the fridge.
3. The bowl had a cucumber and a donut in it.
4. Mary tossed the donut in the trash.
5. Mary ate the donut.

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.

66.5% ✓

33.5% ✗



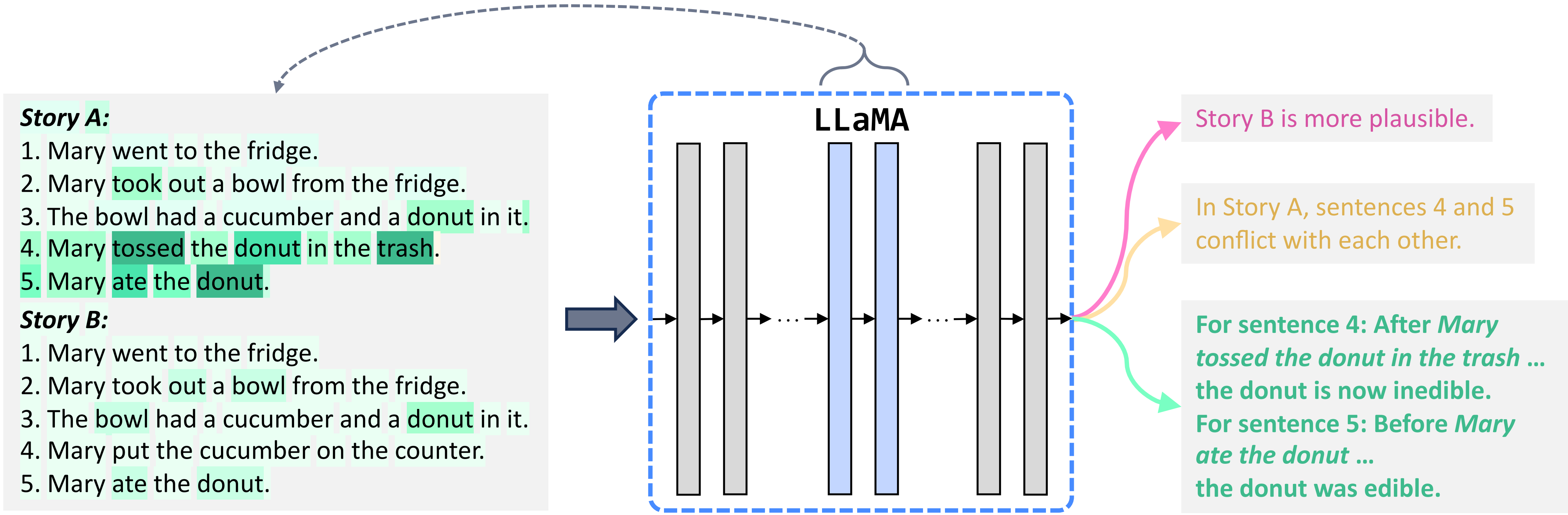
Story B is more plausible.

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

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

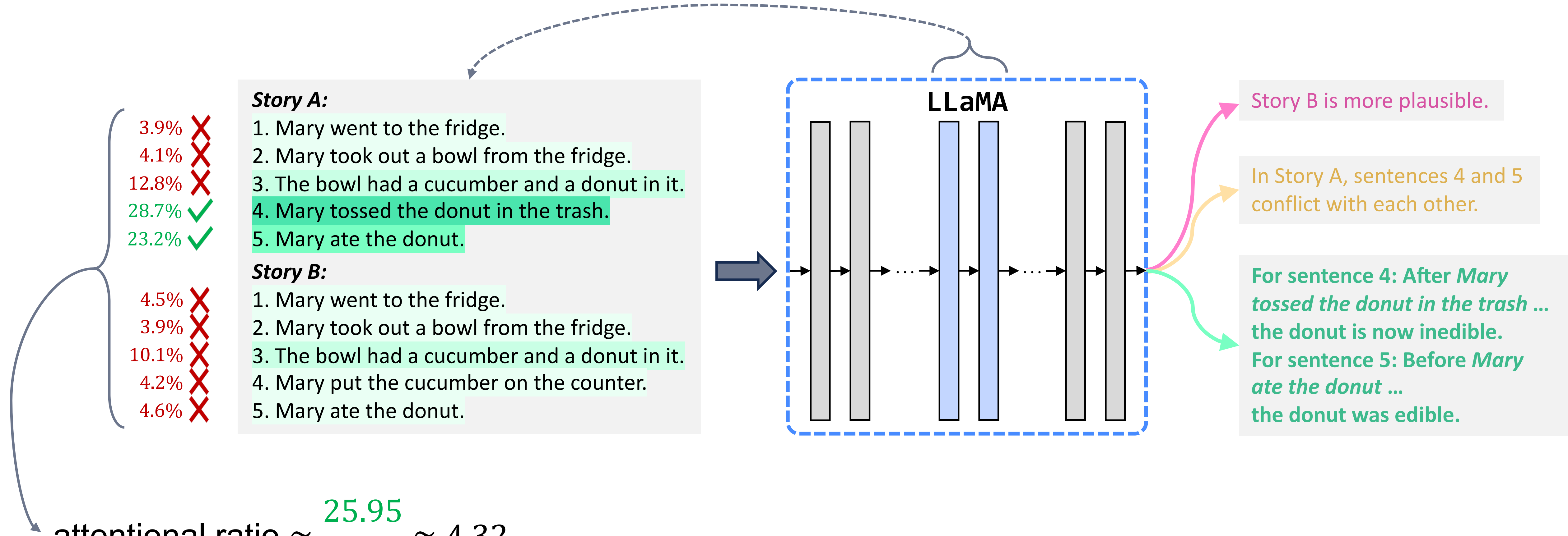
Matthew E. Peters, Mark Neumann, et al. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP 2018.*
Ian Tenney, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR.*
Hugo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

Attention Analysis



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

- We use attentional ratio to measure how **attended context** aligns with the **true context** (which should be used to make predictions)
- 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**

Story Attention Visualization

ICL-U

Story A: 41.0%

1. Coal is heated in the boiler.
2. The water tank over the boiler is heated.
3. Creates steam.
4. The steam is funneled to the piston.
5. Piston uses the steam as energy.
6. The piston causes the crankshaft to move.

Story B: 59.0%

1. Plates on the Earth's crust move slowly past each other.
2. As the plates move, they exert a great force.
3. When the force is large enough, the crust breaks.
4. The stress is released as energy.
5. The energy moves through the Earth in the form of waves.
6. We feel the earthquake.

Attending too much to wrong story!

ICL-HAR

Story A: 16.3%

1. Coal is heated in the boiler.
2. The water tank over the boiler is heated.
3. Creates steam.
4. The steam is funneled to the piston.
5. Piston uses the steam as energy.
6. The piston causes the crankshaft to move.

Story B: 83.7%

1. Plates on the Earth's crust move slowly past each other.
2. As the plates move, they exert a great force.
3. When the force is large enough, the crust breaks.
4. The stress is released as energy.
5. The energy moves through the Earth in the form of waves.
6. We feel the earthquake.

Focusing on correct story!

Sentence Attention Visualization

ICL-U

Story A:

1. Tom found he is out of ice cream. 9.0%
2. Tom peeled a hard boiled egg. 5.5%
3. Tom sliced the egg with a knife. 4.6%
4. Tom washed the knife in the sink. 4.4%
5. Tom ate ice cream for dessert. 8.6%

Story B: %

1. Tom poured a glass of milk. 10.4%
2. Tom peeled a hard boiled egg. 25.4%
3. Tom sliced the egg with a knife. 3.3%
4. Tom washed the knife in the sink. 16.2%
5. Tom ate ice cream for dessert. 12.5%

Attending to wrong sentences!

ICL-HAR

Story A:

1. Tom found he is out of ice cream. 21.3%
2. Tom peeled a hard boiled egg. 7.1%
3. Tom sliced the egg with a knife. 5.3%
4. Tom washed the knife in the sink. 4.4%
5. Tom ate ice cream for dessert. 15.4%

Story B:

1. Tom poured a glass of milk. 7.2%
2. Tom peeled a hard boiled egg. 8.2%
3. Tom sliced the egg with a knife. 2.4%
4. Tom washed the knife in the sink. 20.8%
5. Tom ate ice cream for dessert. 7.9%

Attending to correct sentences!

Summary

- 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
- Still room for improvement…

Conclusion

- Commonsense reasoning in natural language understanding is a longstanding challenge for AI
- While LLMs continue to get closer to achieving this, still a long way before we can completely trust the coherence of their understanding