ACL 2023

ACL 2023 Cutting-Edge Tutorial: Complex Reasoning over Natural Language

Wenting Zhao, Mor Geva*, Bill Yuchen Lin*,
Michihiro Yasunaga*, Aman Madaan*, Tao Yu*

Tutorial Website

Live Q&A

Plan for the tutorial

- Review recent benchmarks on complex reasoning
- Review promising directions for tackling complex reasoning tasks

By reasoning, what do we mean?

Did Aristotle use a laptop? — Input

1. Aristotle lived from 384-322 BC.
2. The first laptop was made in 1980.
3. 322 BC is before 1980.

— Reasoning

No. — Output

By reasoning, what do we mean?

1. Aristotle lived from 384-322 BC.
2. The first laptop was made in 1980.
3. 322 BC is before 1980.

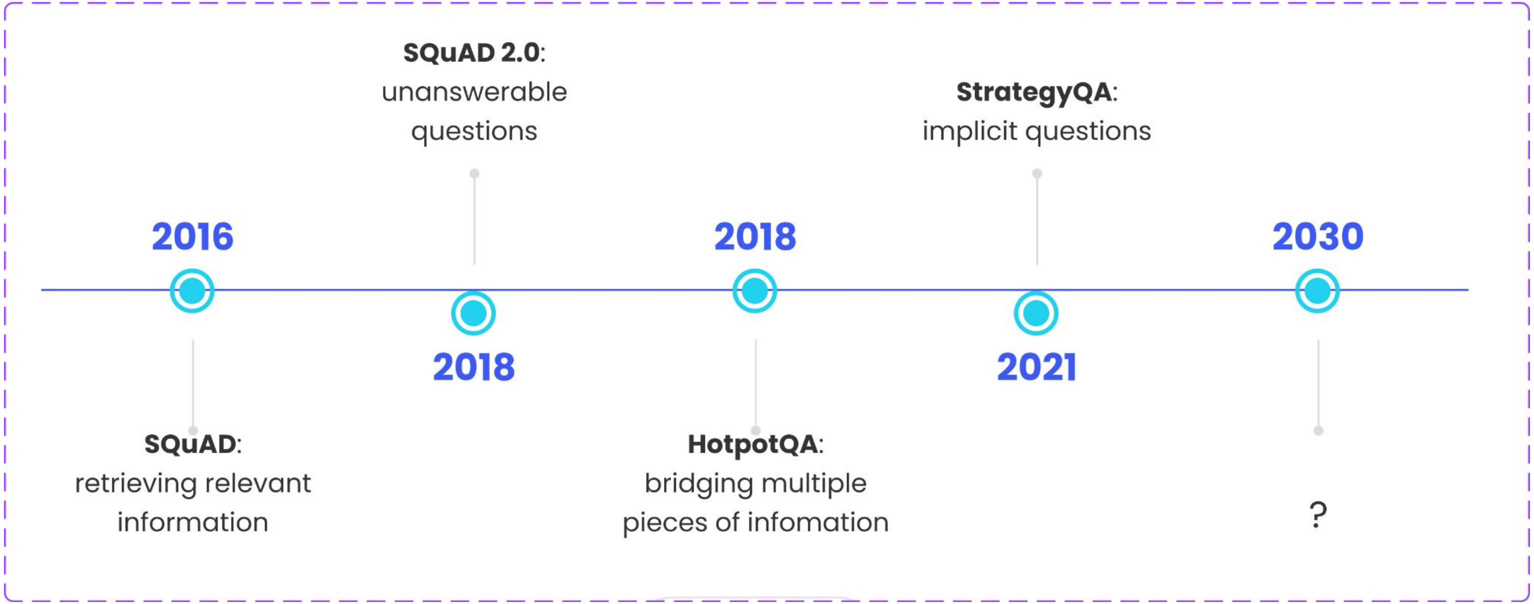
Reasoning

The ***process*** of deriving the output from the input.

Complex reasoning

- Going beyond the surface meaning
    - What can be easily solved by an end-to-end system
- Examples
    - Compositional reasoning
    - Knowledge retrieval
    - Grounding
    - Commonsense reasoning
    - …

Trend of NLP tasks - Question Answering



Trend of NLP tasks - Commonsense Reasoning

Before: Reasoning about common situations

The Smiths went on a vacation without the children.

?

The Smiths brought a souvenir back for Ty.

Ty's face lit up as he ran to the new toy.

Now: Reasoning about uncommon, long-tail situations

She tried sushi for the first time, and really disliked it.

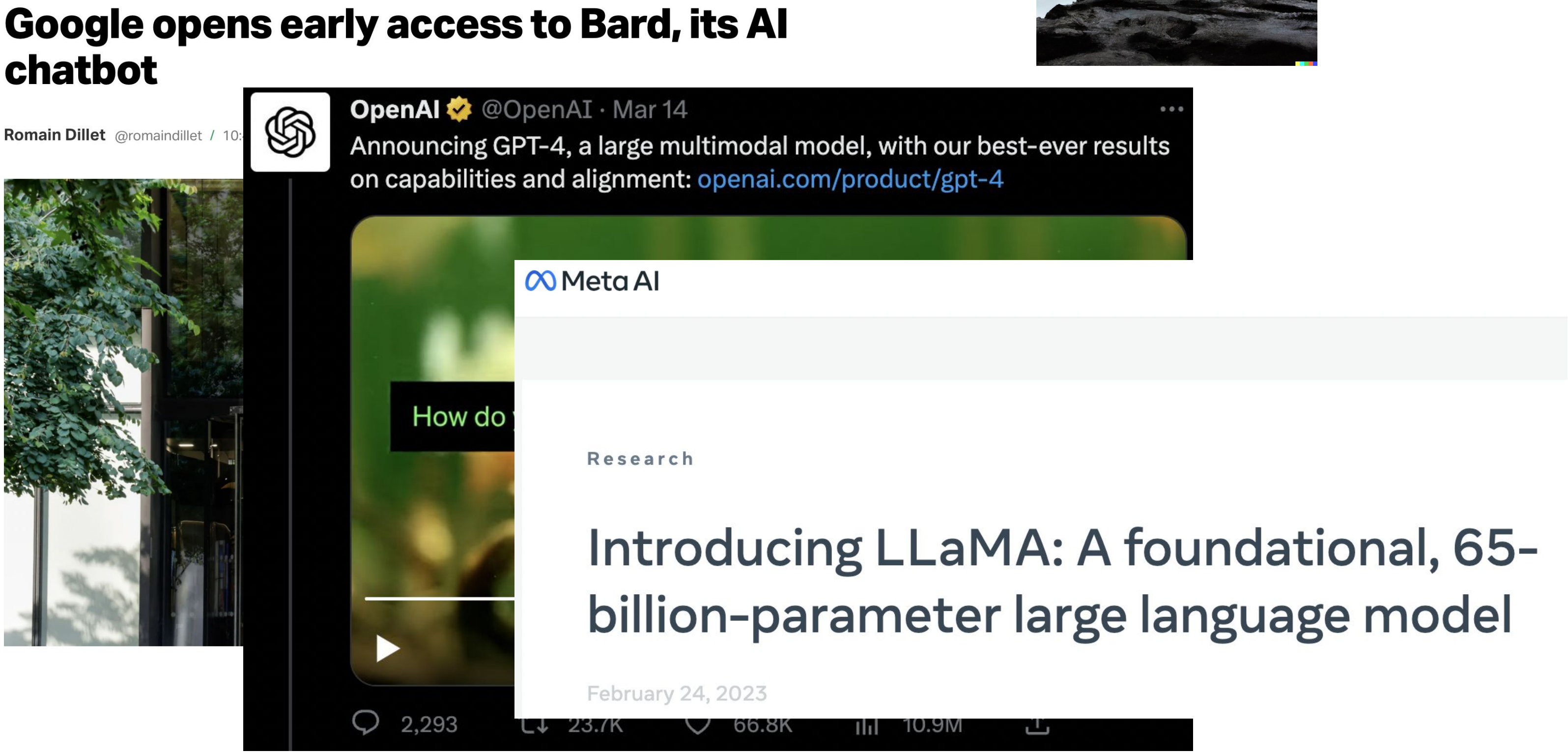
?

She wanted to avoid disappointing her partner.

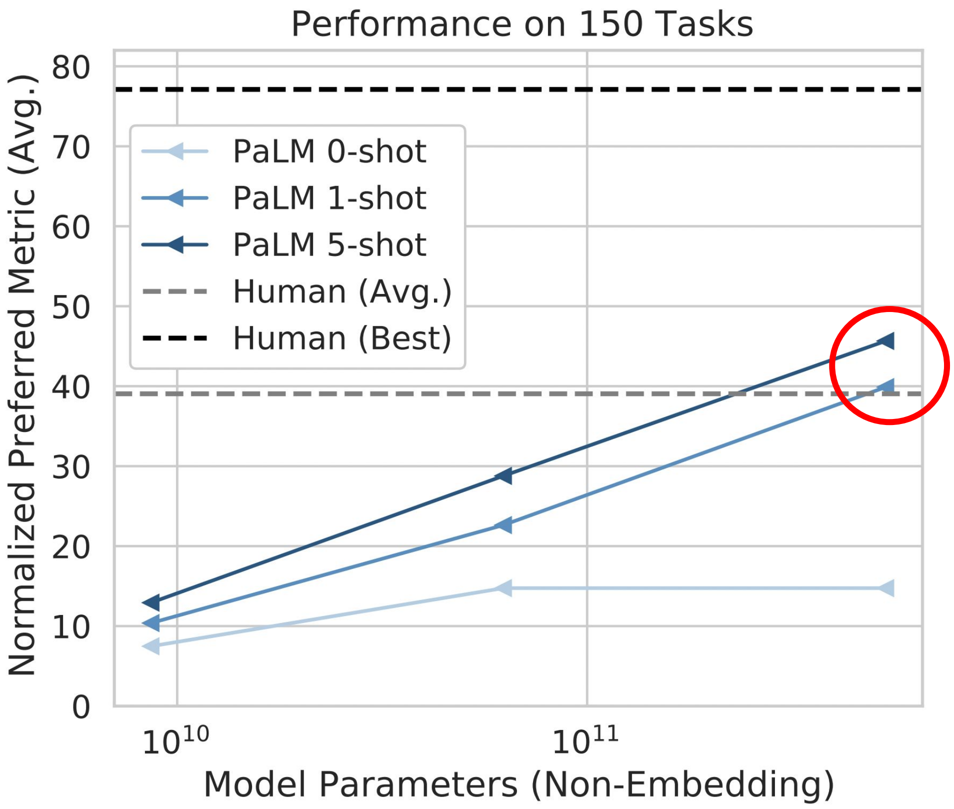
She stayed and ate more sushi.

[Bhagavatula et al., 2019; Uncommonsense, 2023]

In 2023, LLMs are coming in a flood



LLMs have made amazing progress on complex tasks



[Chowdhery et al., 2022]

Are the models really this good?

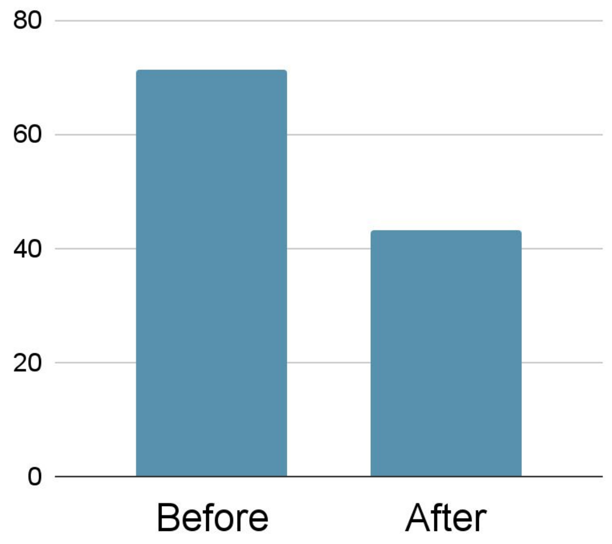
ted.com

Yejin Choi: Why AI is incredibly smart and shockingly stupid

Computer scientist Yejin Choi is here to demystify the current state of massive artificial intelligence systems like ChatGPT, highlighting three key problems ...

Data contamination

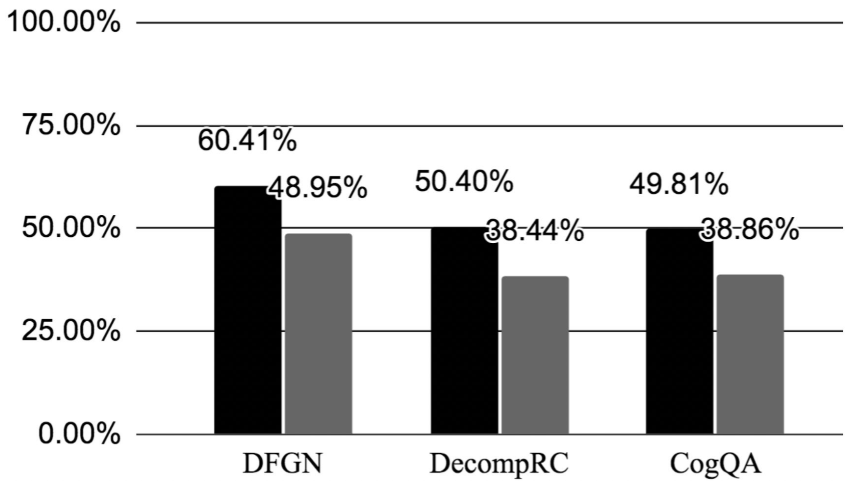
- Before and after removing test data that has n-gram overlap with train data
- Dataset: ARC
- Model: RoBERTa-large



Branco et al., 2021

Generalization

- Multi-hop QA systems are less good at answering single-hop sub-questions
- Dataset: HotpotQA
- Model: RoBERTa-large



Left bar: Multi-hop accuracy
Right bar: Single-hop accuracy

Tang et al., 2021

Limitation

- Standard fine-tuning / prompting methods only maximizes the task accuracy without explicitly considering the underlying reasoning

Why care about taking the correct reasoning route?

- Deployment in critical domains / building trust with users
- Generalization

Goal of this tutorial

- Explore ways to augment language models with methods that make the reasoning process ***explicit***
  - Can we explicitly incorporate knowledge?
  - Can we explicitly specify rules?
  - Can we integrate symbolic reasoning?

Tutorial Schedule

Benchmarks & Evaluation

Mor Geva
Visiting Researcher at Google

"What are the types of complex reasoning abilities recent NLP benchmarks are focused on? And how do we evaluate such abilities?"

9:15-9:40 EST

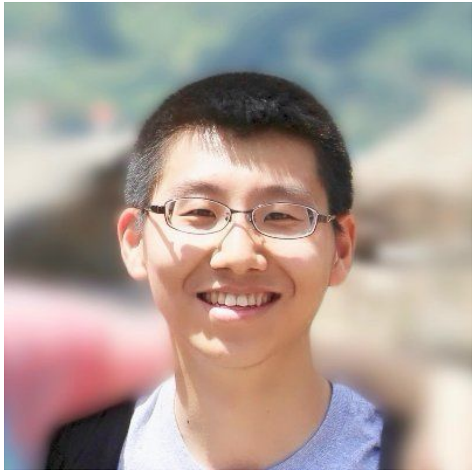
1(a). Knowledge-augmention after pretraining

Yuchen Lin
Postdoc at AI2

“What are the ways to incorporate external knowledge when learning specific NLP tasks?”

9:40-10:05 EST

1(b). Knowledge-augmented pretraining



Michihiro Yasunaga
PhD student at Stanford

“We often incorporate knowledge in a task-specific manner 🤔 Can we do this during pretraining to help a broader range of downstream tasks?

10:05 -10:30 EST

2. Few-shot prompting approaches

Aman Madaan
PhD student at CMU

"What are the clever ways to perform few-shot prompting so that it's more robust and requires less prompt engineering efforts?"

11:00-11:30 EST

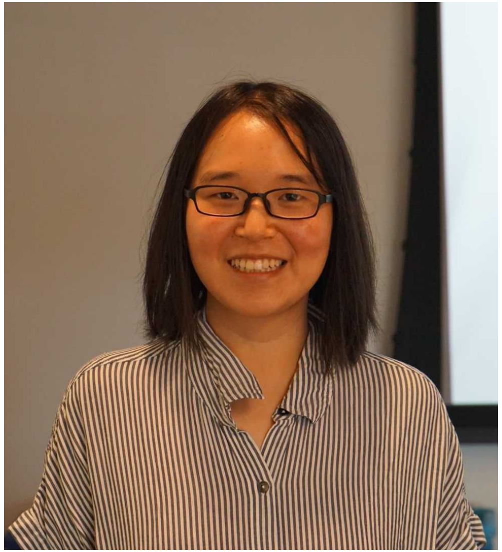
3. Neuro-symbolic approaches: LLMs + tool use

Tao Yu
Assistant Professor at HKU

"Can LLMs utilize external tools to not only expand their capacities but also to make our NLP systems more robust, scalable, and interpretable?"

11:30-12:00 EST

4. Rationale-based approaches & Conclusion



Wenting Zhao
PhD student at Cornell

"Let's think about ways to produce rationales and how can they improve the existing NLP systems."

12:00-12:30 EST

Paper / Workshop Highlights at ACL’23

Abductive Commonsense Reasoning Exploiting Mutually Exclusive Explanations
Wenting Zhao, Justin Chiu, Claire Cardie, Alexander Rush
11:45-12:00 (Metropolitan West) on July 10

Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions
Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, Ashish Sabharwal
09:00-10:30 (Frontenac Ballroom and Queen’s Quay) on July 11

Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios
Jiaxuan Li, Lang Yu, Allyson Ettinger
11:00-12:30 (Frontenac Ballroom and Queen’s Quay) on July 10

Single Sequence Prediction over Reasoning Graphs for Multi-hop QA
Gowtham Ramesh, Makesh Narsimhan Sreedhar, Junjie Hu
09:00-10:30 (Frontenac Ballroom and Queen’s Quay) on July 11

Workshop: Natural Language Reasoning and Structured Explanations
July 13

Paper list

[Bhagavatula et al., 2019] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In International Conference on Learning Representations

[Chowdhery et al., 2022] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S. and Schuh, P., 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

[Branco et al., 2021] Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted Commonsense: Data Spuriousness in Deep Learning of Commonsense Reasoning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1504–1521, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Tang et al., 2021] Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 2021. Do Multi-Hop Question Answering Systems Know How to Answer the Single-Hop Sub-Questions?. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3244–3249, Online. Association for Computational Linguistics.