

# Language for Knowledge and Reasoning

---

Jeff Dalton  
Associate Professor  
University of Glasgow

ECIR 2023 Tutorial: Neuro-Symbolic  
Representations for IR – Part 3.1



# Outline



## Part I: Knowledge in LLMs

- Accessing Knowledge
- Extracting Knowledge

## Part II: LLM “Reasoning”

- Chain of Thought
- Rationale Generation

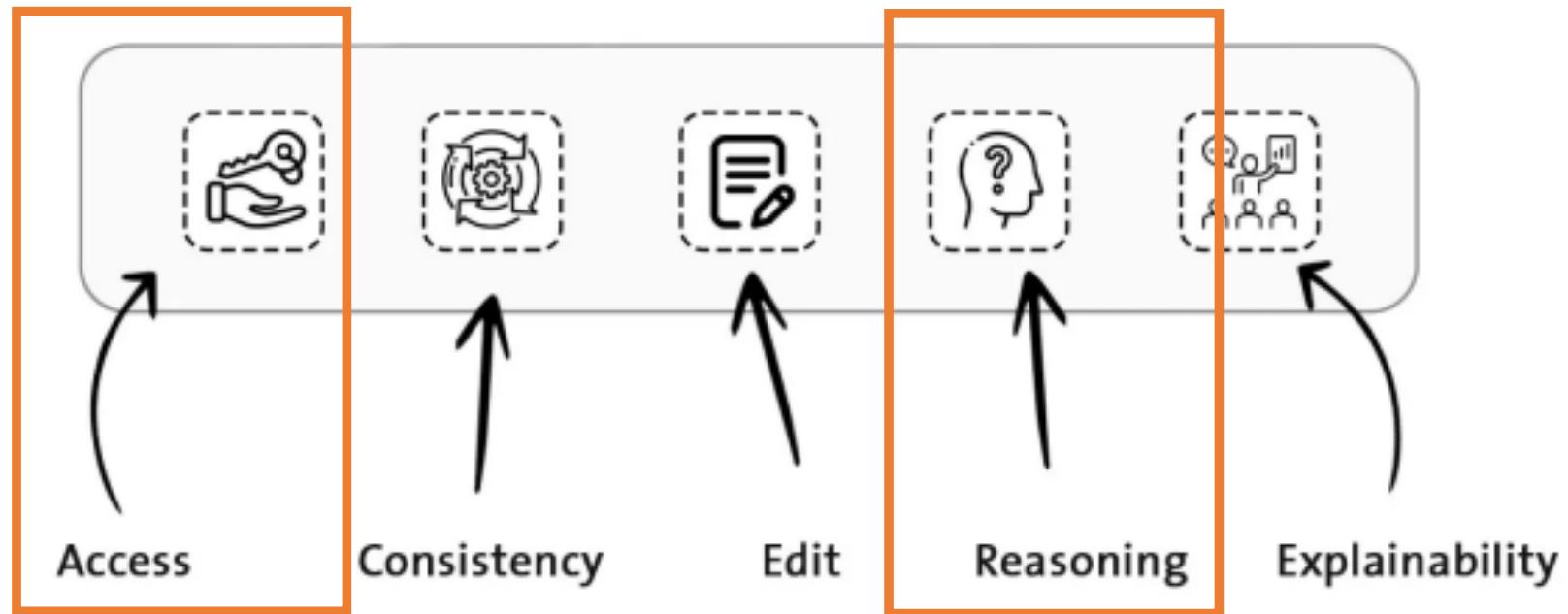
## Part III: Text as Knowledge

- Textual entailment
- Natural logic
- Generating Proofs

## Part VI: Tools Use and Program of Synthesis

- Program of Thought (PoT), PAL
- Demonstrate-Search-Predict

## Part 1: LLMs as KBs

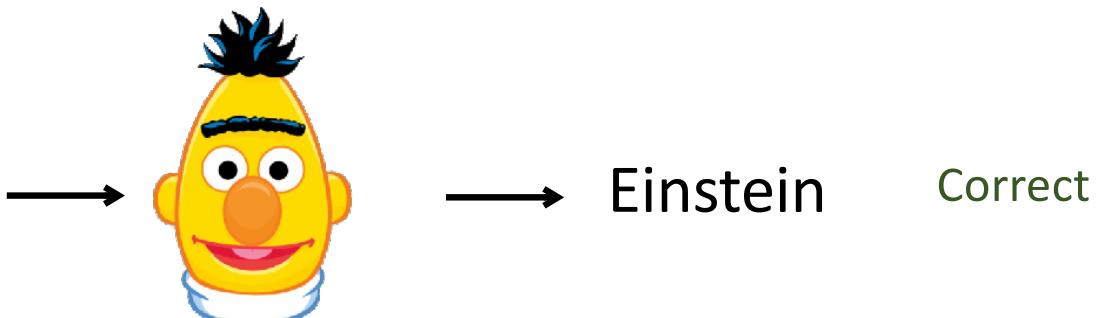


A Review on Language Models as Knowledge Bases, 2022

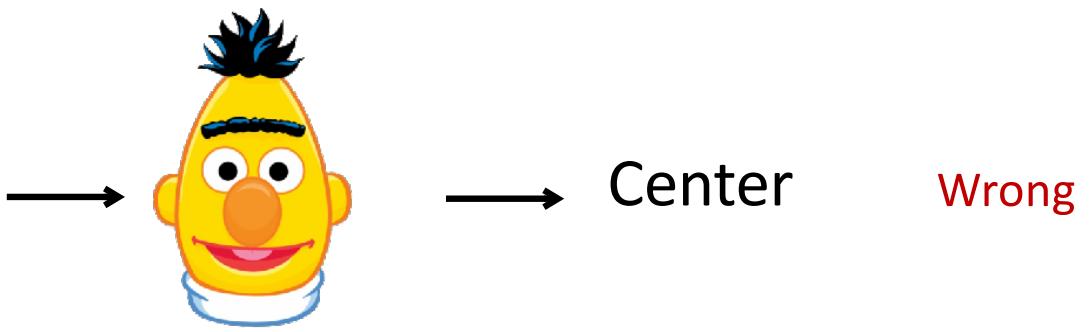
## Access: Natural Language

- Start simple: Zero-shot “fill-in-the-blank” cloze statements

The theory of relativity  
was developed by \_\_\_\_.



Giacomo Tedesco  
plays in \_\_\_\_ position



Language Models as Knowledge Bases?  
EMNLP 2019

## Access: Adding context

- Analogies:

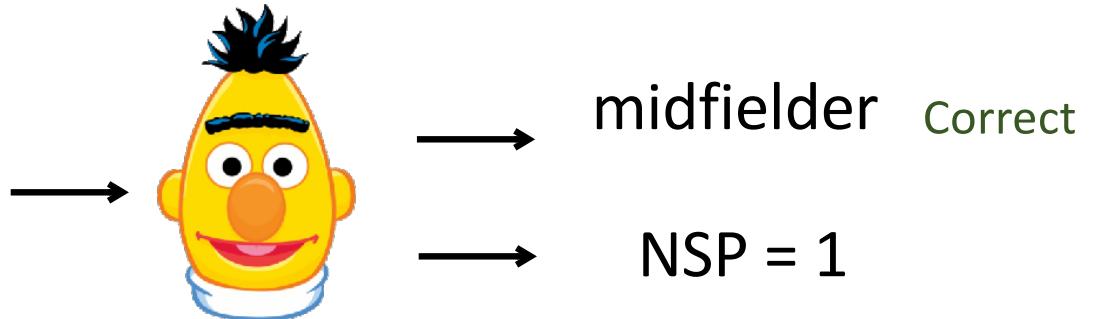
“Obama was born in Hawaii. [SEP] Jobs was born in [MASK].”  
(Brown et al., 2020; Madotto et al., 2020; Gao et al., 2020)

- Entity context from retrieval:

“Jobs lives in California. [SEP] Jobs was born in [MASK].”

Giovanni Tedesco has  
two brothers who are  
also football players,  
Salvatore and Giacomo.  
[SEP]

Giacomo Tedesco plays  
In \_\_\_\_\_ position .



*How Context Affects Language Models' Factual Predictions, AKBC 2020*

## KB Completion Results

LAMA	Relation	B	B-ADV	<i>open domain sourced context</i>			B-ORA
		B-GEN	DRQA	B-RET			
Google-RE	birth-place	16.1	14.5	8.5	<b>48.6</b>	43.5	70.6
	birth-date	1.4	1.4	1.4	42.9	<b>43.1</b>	98.1
	death-place	14.0	12.6	6.0	<b>38.4</b>	35.8	65.1
	Total	10.5	9.5	5.3	<b>43.3</b>	40.8	78.0
T-REx	1-1	74.5	74.5	71.3	55.2	<b>81.2</b>	91.1
	N-1	34.2	33.8	32.7	30.4	<b>47.5</b>	67.3
	N-M	24.3	23.6	23.8	15.4	<b>32.0</b>	52.4
	Total	32.3	31.8	31.1	25.8	<b>43.1</b>	62.6
SQuAD		17.4	17.4	15.8	<b>37.5</b>	34.3	61.7
<i>weighted average</i>		30.5	30.0	29.0	27.2	<b>42.8</b>	63.6

*How Context Affects Language Models' Factual Predictions, AKBC 2020*

# GPT-3 and ChatGPT

GPT-3

The theory of relativity was developed by \_\_\_\_\_

Albert Einstein.

Giovanni Tedesco has two brothers who  
are also football players, Salvatore and Giacomo..  
Giacomo Tedesco plays in \_\_\_\_\_ position

Guitar.

JE

Giacomo Tedesco plays in \_\_\_\_\_ position



I'm sorry, but I don't have enough information to answer your question accurately. Can you   
please provide more context or specify which Giacomo Tedesco you are referring to?

JE

Giovanni Tedesco has two brothers who are also football players, Salvatore and Giacomo.



Thank you for providing more context. Based on my research, Giacomo Tedesco, who is a brother of Giovanni Tedesco and Salvatore Tedesco, is also a professional football player. However, I couldn't find information about his specific position. It's possible that he plays in different positions depending on the team he's currently playing for or the tactics being employed.

JE

Search for Giacomo Tedesco. And find out what position he plays in.



Finished browsing ^



Searched: "Giacomo Tedesco position"

:≡ Reading content

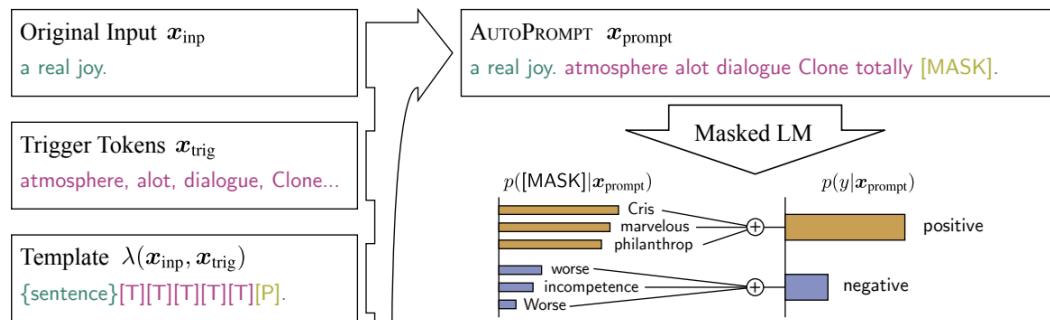
⌚ Finished browsing

Giacomo Tedesco - Player profile | Transfermarkt

Giacomo Tedesco played as a midfielder, specifically in the central midfield position<sup>1</sup>.

# Access: Learning to Ask

- Discrete: AutoPrompt



AutoPrompt: Eliciting Knowledge from Language Models w/ Automatically Generated Prompts, EMNLP 2020

- Continuous: Soft Prompt

Mary Cassatt performed until his death

in \_\_\_\_y.

\_\_\_\_x  $v_1$   $v_2$   $v_3$   $v_4$   $v_5$  \_\_\_\_y  $v_6$



Model	P@1	P@1
	T-REx orig.	T-REx ext.
LAMA (BEb)	31.1	26.4
LPAQA(BEb)	34.1	31.2
AutoPrompt	43.3	45.6
Soft (sin., BEb)	47.7 (+16.6 <sup>?</sup> )	49.6 (+23.2 <sup>?</sup> )
Soft (min., BEb)	<b>50.7<sup>?</sup> (+16.6<sup>?</sup>)</b>	<b>50.5<sup>?</sup> (+19.3<sup>?</sup>)</b>
Soft (par., BEb)	48.4 (+12.8 <sup>?</sup> )	49.7 (+18.5 <sup>?</sup> )
Soft (ran., BEb)	48.1 (+47.4)	50.6 (+49.8)
LAMA (BEI)	28.9 <sup>†</sup>	24.0 <sup>†</sup>
LPAQA(BEI)	39.4 <sup>†</sup>	37.8 <sup>†</sup>
Soft (sin., BEI)	51.1 (+22.2)	51.4 (+27.4)
Soft (min., BEI)	<b>51.6 (+12.2)</b>	<b>52.5 (+14.7)</b>
Soft (par., BEI)	51.1 (+11.7)	51.7 (+13.9)
Soft (ran., BEI)	51.9 (+47.1)	51.9 (+50.5)
AutoPrompt	40.0	-
Soft (min., Rob)	<b>40.6<sup>?</sup> (+39.4)</b>	-

BEb is BERT-base, BEI is BERT-large, Rob is RoBERTa-base.

Learning How to Ask: Querying LMs with Mixtures of Soft Prompts, NAACL 2021

## Trained KG Extraction

- T51.1 XXL
- Learn the semantics of the task/prompt and expected output domain (entity types/subtypes) for each relation

### Zero-shot

**Test Query:** Marat Makhmutov was born in [MASK].

**Correct Answer:** Moscow

**Model Answer:** Moscow

### Finetuned

**Train Data:** Out of all “X was born in [MASK] .” queries:

- the answer to 5 of them is **Moscow**,
- the answer to 5 of them is **Baku**.

**Test Query:** Marat Makhmutov was born in [MASK].

**Correct Answer:** Moscow

**Model Answer:** Baku

Strategy	Hit@1	Hit@3	Hit@5
ZS	35.2	47.9	52.7
FS	47.0	56.1	57.2
RR	39.9	49.9	52.7
FT	<b>51.9</b>	<b>68.4</b>	<b>73.9</b>

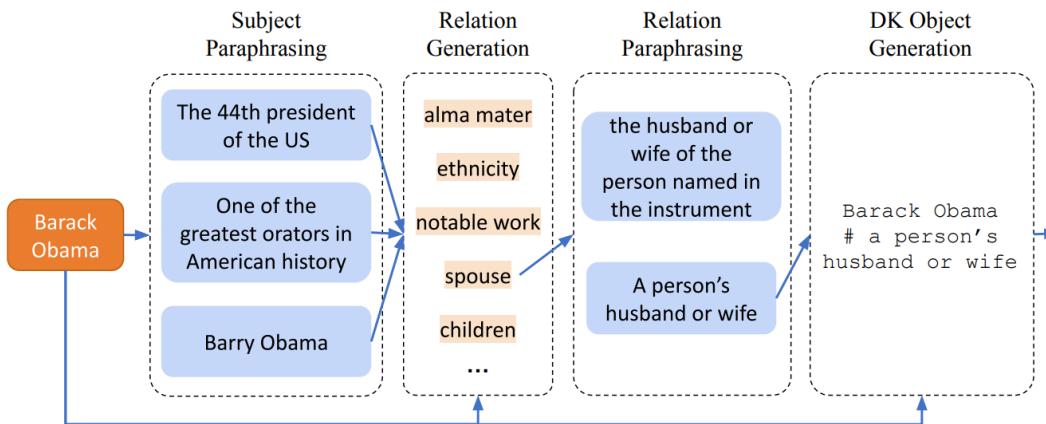
LPAQA as development and LAMA (TREx) as test set

ZS – zero, FS – Few shot, RR - Reranking, FT - finetuning

Understanding Finetuning for Factual Knowledge Extraction from Language Models, 2022

# Prompt Chaining for KB Extraction

- LMCrawl - Extract a KG from GPT-3



## Series of Manual Prompts

Sub-task	Query	Prompt	Expected Output
Relation Generation	Philippines	Q: René Magritte A: ethnic group, place of birth, place of death, sex or gender, spouse, country of citizenship, member of political party, native language, place of burial, cause of death, residence, family name, given name, manner of death, educated at, field of work, work location, represented by Q: Stryn A: significant event, head of government, country, capital, separated from Q: Philippines A:	leader name # cctld # capital # calling code
Pure Object Generation	Barack Obama # child	Q: Monte Cremasco # country A: Italy Q: Johnny Depp # children A: Jack Depp # Lily-Rose Depp Q: Wolfgang Sauseng # employer A: University of Music and Performing Arts Vienna Q: Barack Obama # child A:	Sasha Obama # Malia Obama
DK Object Generation	Queen Elizabeth II # date of death	Q: Heinrich Peters # occupation A: Don't know Q: Monte Cremasco # country A: Italy Q: Ferydoon Zandi # place of birth A: Don't know Q: Hans Ertl # sport A: mountaineering Q: Queen Elizabeth II # date of death A:	Don't know
Subject Paraphrasing	Alan Turing	Alan Turing is also known as:	The father of computing

	Main Test Set				Head Test Set			
	one-hop		two-hop		one-hop		two-hop	
	Precision	# of Facts	Precision	# of Facts	Precision	# of Facts	Precision	# of Facts
Pure-Greedy	54.6 + 8.2	6.2 + 2.8	43.4 + 6.1	26.1 + 5.5	80.3 + 8.4	14.4 + 3.9	62.1 + 7.3	82.3 + 15.4
LMCRAWL	$83.3 \pm 7.9$	$5.4 \pm 1.1$	$82.0 \pm 7.5$	$21.4 \pm 4.7$	$91.5 \pm 11.4$	$11.0 \pm 4.6$	$90.9 \pm 4.9$	$61.2 \pm 25.1$

Results on 100 seed entities from WikiData

Crawling The Internal Knowledge-Base of Language Models, 2022

## LLMs and KB Summary

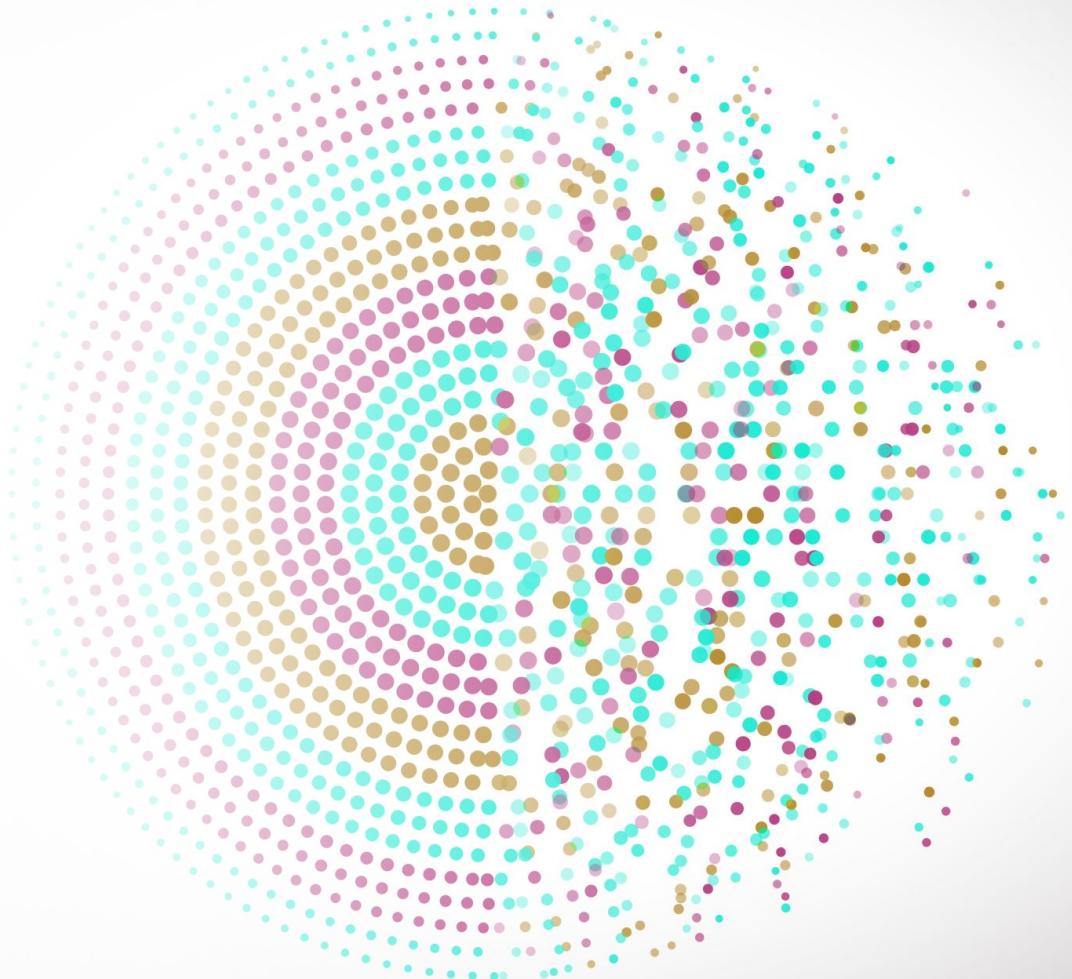
- LLMs encode significant knowledge in parametric memory
- Automatic prompting or fine-tuning approaches outperform manual
- Extracting or “crawling” is still evolving
- Lots of open challenges!

	LM-as-KB	Structured KB
<b>Construction</b>	Self/Unsupervised	Manual or semi-automatic
<b>Schema</b>	Open-ended	Typically fixed
<b>Maintenance</b> - adding facts - correcting/deleting	Difficult, unpredictable side effects Difficult	Easy Easy
<b>Knows what it knows</b>	No, assigns probability to everything	Yes, content enumerable
<b>Entity disambiguation</b>	No/limited	Common
<b>Provenance</b>	No	Common

Language Models As or For Knowledge Bases, 2021

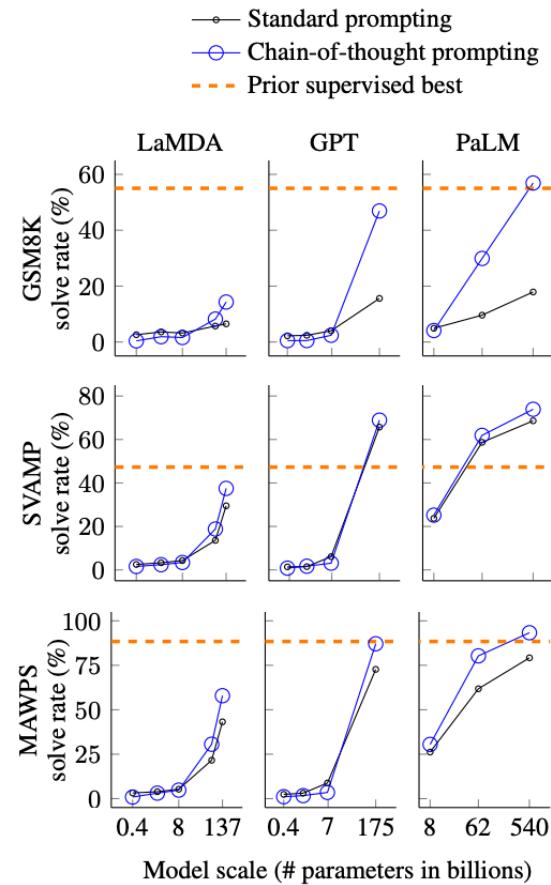
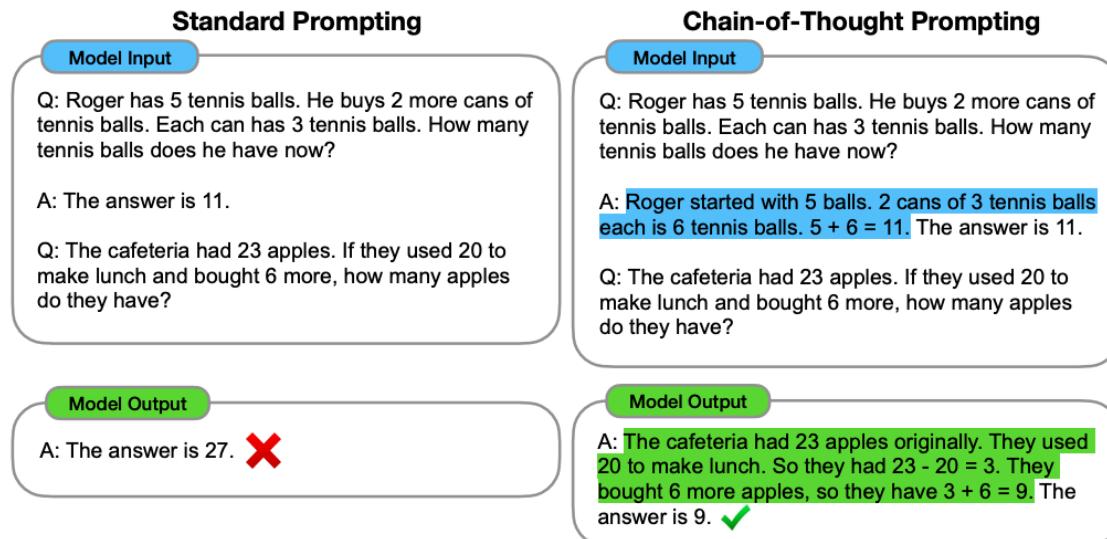
# Part II: Chain of Thought and Explanations

ECIR 2023 Tutorial: Neuro-Symbolic  
Representations for IR – Part 3.1



# The Power of How: Chain of Thought (CoT)

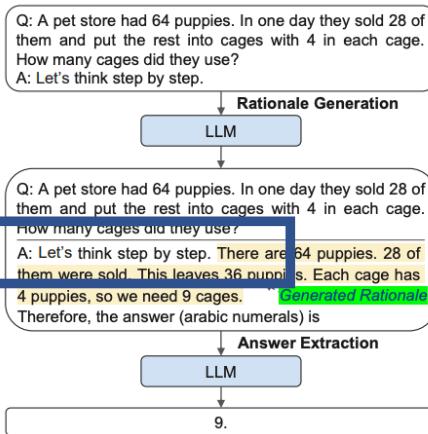
- Generating explanation and reasoning “chains”.
- CoT only works on large LMs



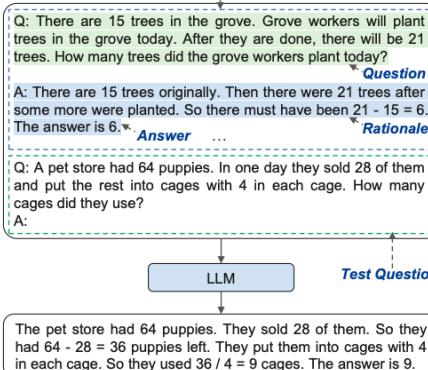
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

# CoT Types: Zero to Automatic

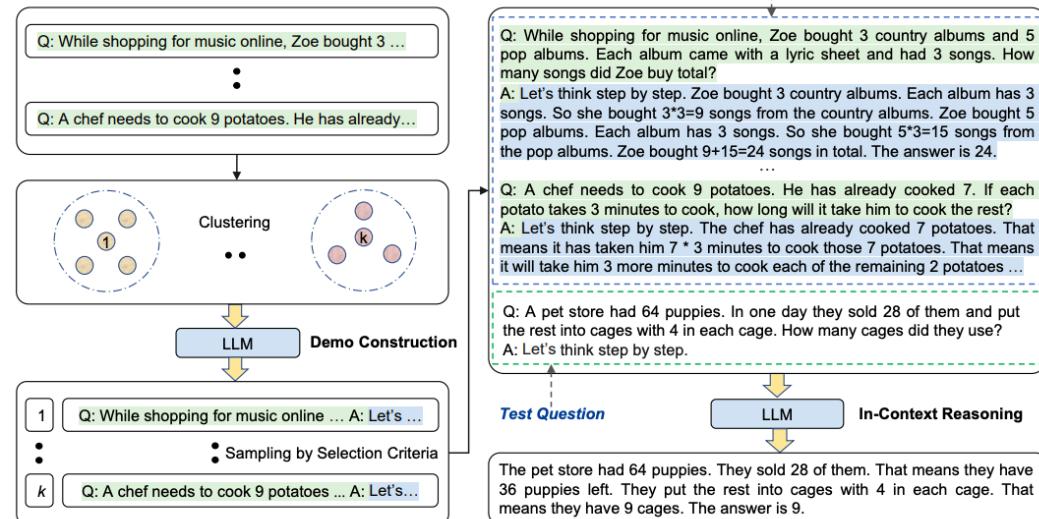
## Zero-shot CoT



## Manual CoT



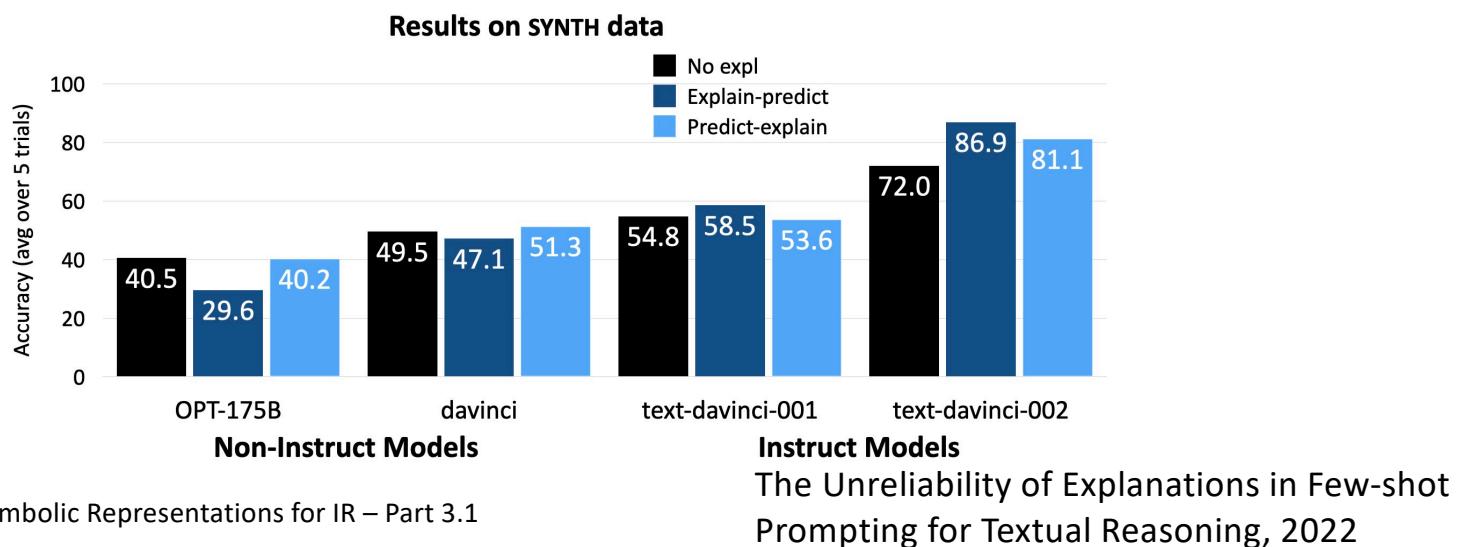
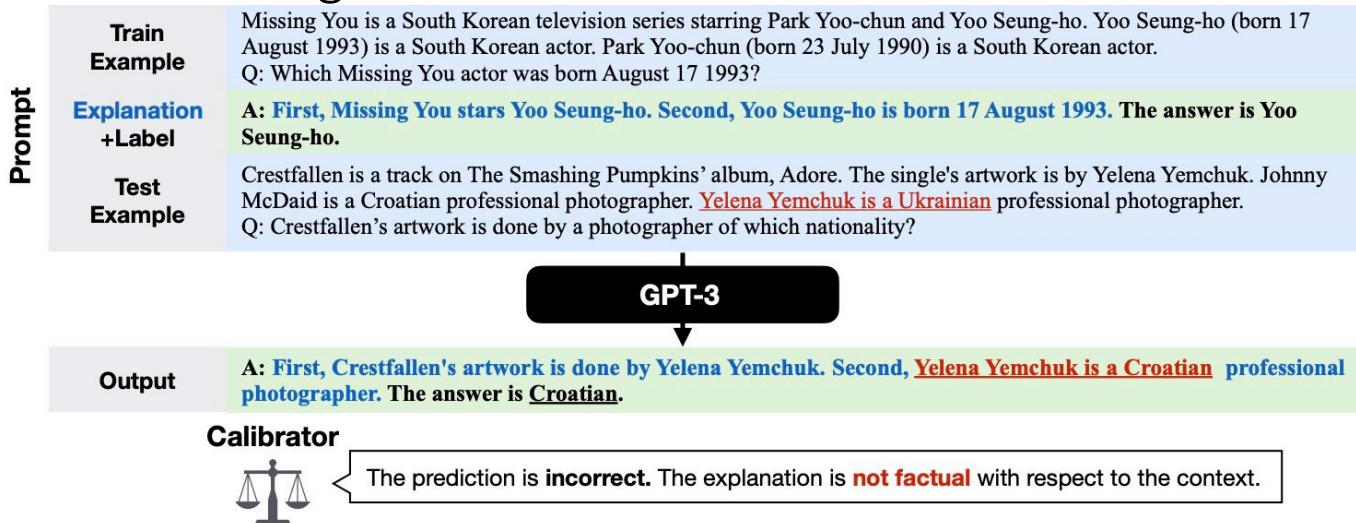
## Automatic CoT



Model	Arithmetic							Commonsense		Symbolic	
	MultiArith	GSM8K	AddSub	AQuA	SingleEq	SVAMP	CSQA	Strategy	Letter	Coin	
Zero-Shot	22.7	12.5	77.0	22.4	78.7	58.8	72.6	54.3	0.2	53.8	
Zero-Shot-CoT	78.7	40.7	74.7	33.5	78.7	63.7	64.6	54.8	57.6	91.4	
Few-Shot	33.8	15.6	83.3	24.8	82.7	65.7	<b>79.5</b>	<b>65.9</b>	0.2	57.2	
Manual-CoT	91.7	46.9	81.3	35.8	86.6	68.9	73.5	65.4	59.0	97.2	
Random-Q-CoT	$87.1 \pm 1.8$	$40.4 \pm 0.4$	$82.7 \pm 1.3$	$31.5 \pm 1.1$	$81.5 \pm 0.3$	$66.7 \pm 1.8$	$71.9 \pm 0.2$	$58.0 \pm 0.1$	$58.2 \pm 0.3$	$95.9 \pm 0.1$	
Auto-CoT	$92.0 \pm 1.7$	$47.9 \pm 3.7$	$84.8 \pm 2.9$	$36.5 \pm 2.2$	$87.0 \pm 1.2$	$69.5 \pm 2.2$	$74.4 \pm 2.5$	$65.4 \pm 0.1$	$59.7 \pm 3.2$	$99.9 \pm 0.1$	

Automatic Chain of Thought Prompting in Large Language Models

# CoT for Text Reasoning



## Background: Rationales in Crowdsourcing IR Judgments

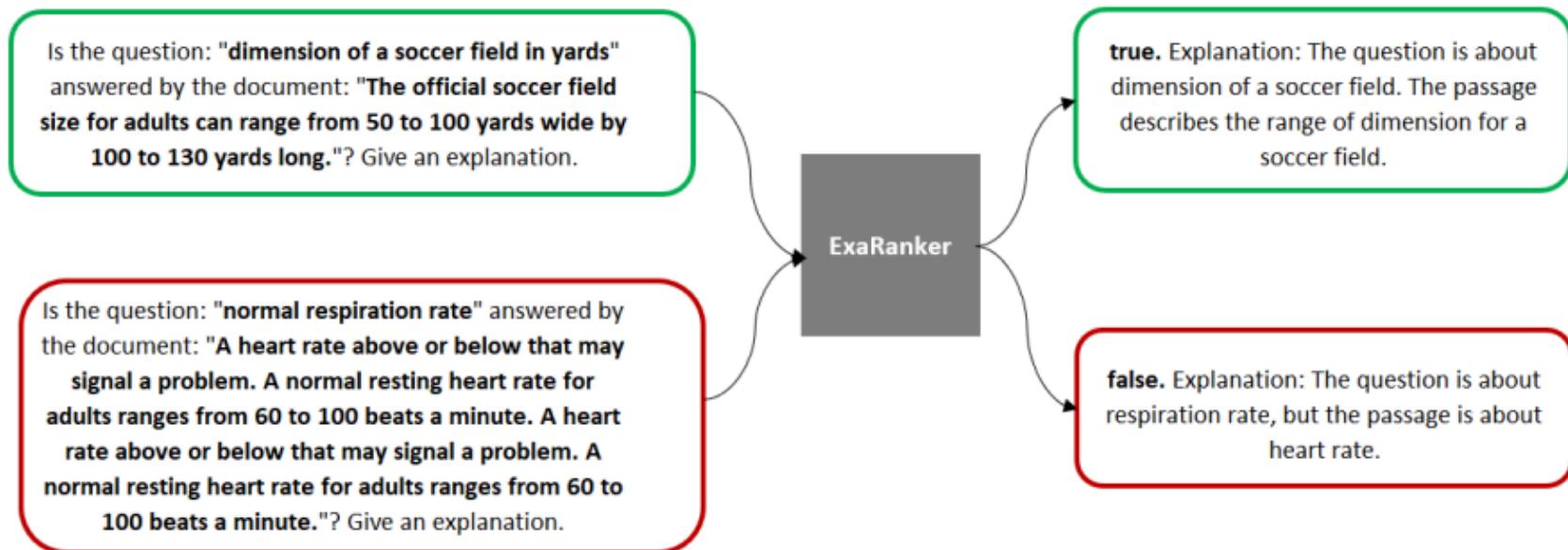
- Adding rationales improves crowdworker consistency and data quality

<b>Query (Alice)</b>	dogs for adoption		
<b>Narrative (Alice)</b>	I want to find information on adopting a dog. This includes names and locations of rescue organizations or vehicles (e.g. classifieds) as well as documents with info on qualifications, fees (if any), what to expect, resources, etc. Organizations may be rescue organizations, pounds, shelters, etc. but not breeders or pet shops, unless the pet shop runs adoption fairs. A site providing general information on dog adoption is also relevant.		
	<b>Document 1</b>	<b>Document 2</b>	<b>Document 3</b>
<b>Worker 1 Judgment (Tom)</b>	Probably Not Relevant	Definitely Relevant	Definitely Not Relevant
<b>Worker 1 Rationale (Tom)</b>	<i>Rooterville Sanctuary. For adoption: pets, pig, pigs, piggy, piggies, pork.</i>	<i>View our rescue dogs - visit our organization or contact us directly to see what is available.</i>	<i>The dogs listed here all require a new home. These dogs all deserve that second chance and you may be that special person to give it to them. View Rescue Dogs adoption fees. Contact us for more info.</i>
<b>Worker 2 Judgment</b>	Probably Not Relevant	Probably Relevant	Definitely Relevant
<b>Worker 2 Reasoning</b>	I agree that this organization is probably not likely to be one where Alice will find the animal she is looking for, since they seem to focus on pigs, although they mention dogs	It is a site that lists dog rescue organizations, which is what Alice is searching for. But it is an Australian website. I suspect Alice was looking for an organization in the US.	Tom provided a lot of information that shows why this page should be useful for Alice.
<b>Gold Standard</b>	Probably Not Relevant	Probably Relevant	Definitely Relevant

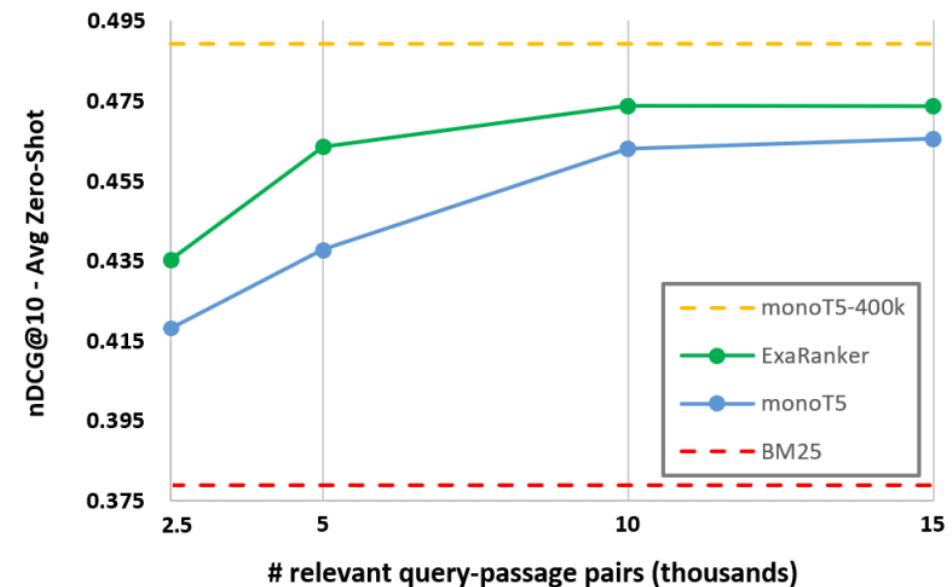
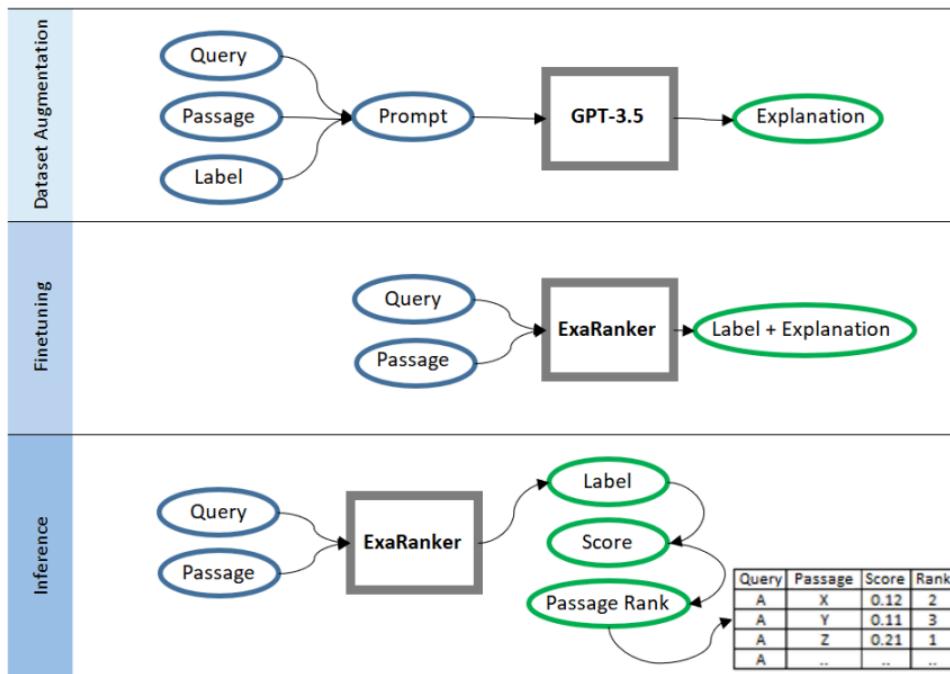
Table 4: Examples of the Two-Stage Task with worker responses for three different documents.

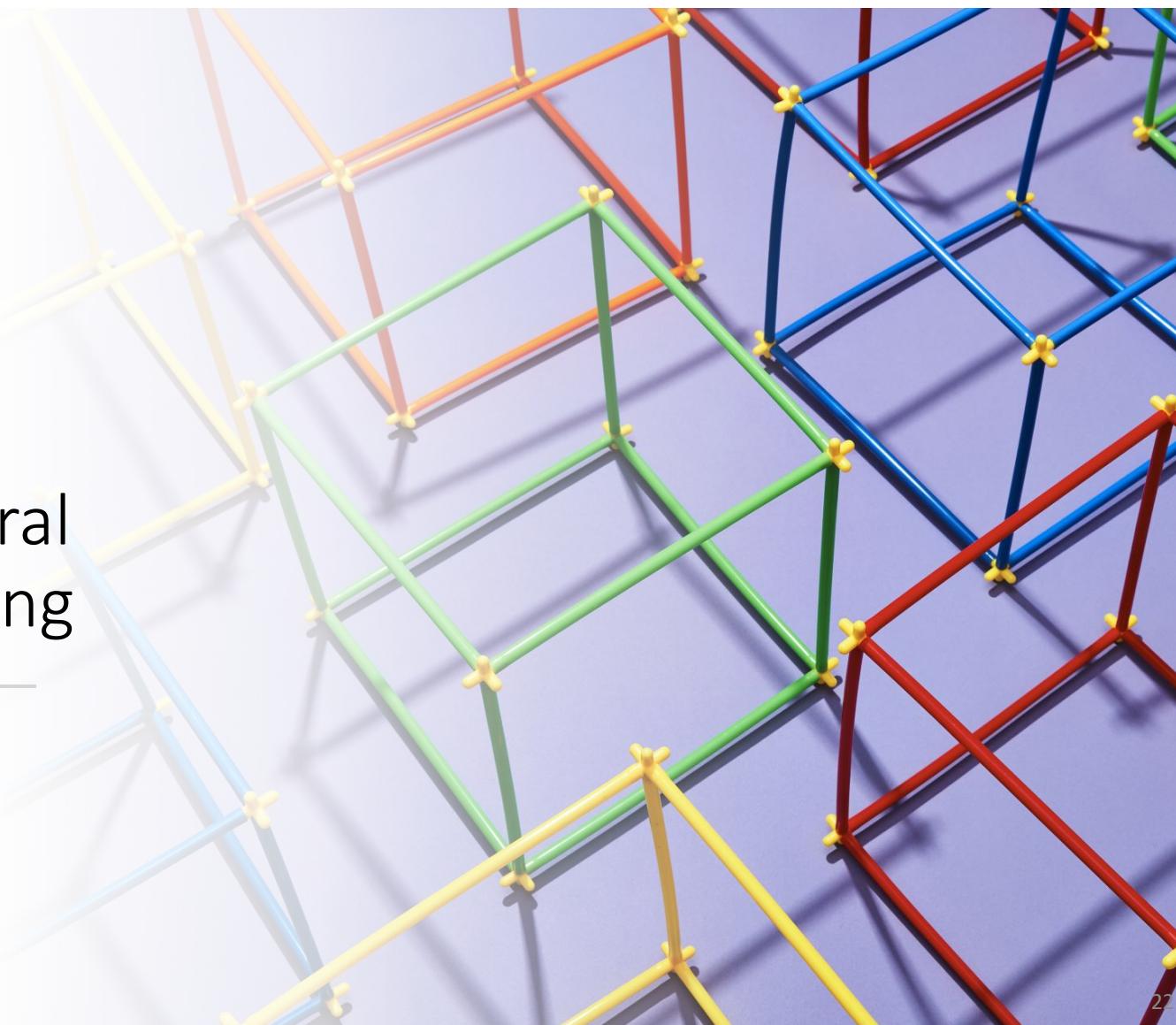
## Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments

# ExaRanker: Explanation-Augmented Neural Ranker



# ExaRanker: Explanation-Augmented Neural Ranker





## Part III: From CoT to Natural Language Reasoning

---

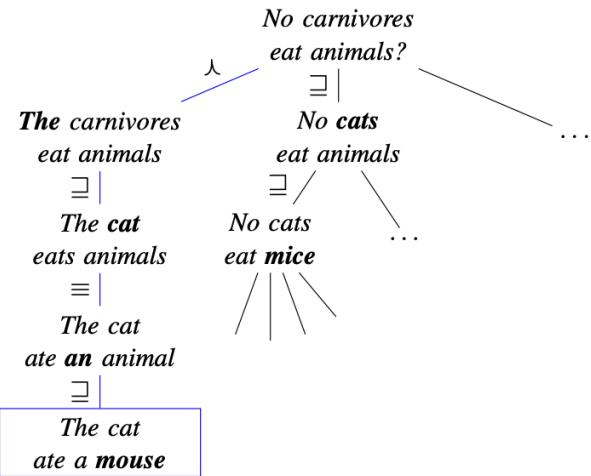
ECIR 2023 Tutorial: Neuro-Symbolic  
Representations for IR – Part 3.1

## Background: Natural Logic

- Task: Does a piece of text follow from or contradict another?  
[Dagan 2005, MacCartney & Manning 2009]
- Logical inference with first order predicate calculus
- Natural logic works well with facts like those in a KB

NatOP: Name	Definition
$\uplus$ : Alternation	$x \cap y = \emptyset \wedge x \cup y = U$
$\smile$ : Cover	$x \cap y \neq \emptyset \wedge x \cup y = U$
$\equiv$ : Equivalence	$x = y$
$\sqsubseteq$ : Forward Entailment	$x \subset y$
$\perp$ : Negation	$x \cap y = \emptyset \wedge x \cup y = U$
$\sqsupseteq$ : Reverse Entailment	$x \supset y$
$\#$ : Independence	All other cases

Table 1: Natural logic relations (NatOps) and their set theoretic definitions.



NaturalLI: Natural Logic Inference for Common Sense Reasoning

# Natural Language Inference (NLI) Test Collections

## Stanford NLI Corpus

A man inspects the uniform of a figure in some East Asian country.	<b>contradiction</b> C C C C C	The man is sleeping
An older and younger man smiling.	<b>neutral</b> N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	<b>contradiction</b> C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	<b>entailment</b> E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	<b>neutral</b> N N E C N	A happy woman in a fairy costume holds an umbrella.

A large annotated corpus for learning natural language inference

## e-SNLI Corpus

Premise: An adult dressed in black **holds a stick**.

Hypothesis: An adult is walking away, **empty-handed**.

Label: contradiction

Explanation: Holds a stick implies using hands so it is not empty-handed.

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.

Hypothesis: A young **mother** is playing with her **daughter** in a swing.

Label: neutral

Explanation: Child does not imply daughter and woman does not imply mother.

e-SNLI: Natural Language Inference with Natural Language Explanations

## EntailmentBank

**Question:** How might eruptions affect plants?

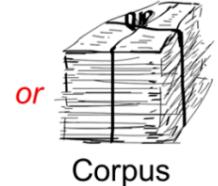
**Answer:** They can cause plants to die

### Hypothesis

**H (hypot):** Eruptions can cause plants to die

### Text

sent1: eruptions emit lava.  
sent2: eruptions produce ash clouds.  
sent3: plants have green leaves.  
sent4: producers will die without sunlight  
sent5: ash blocks sunlight.



### Entailment Tree

**H (hypot):** Eruptions can cause plants to die

int1: Eruptions block sunlight.

sent4: producers will die without sunlight.

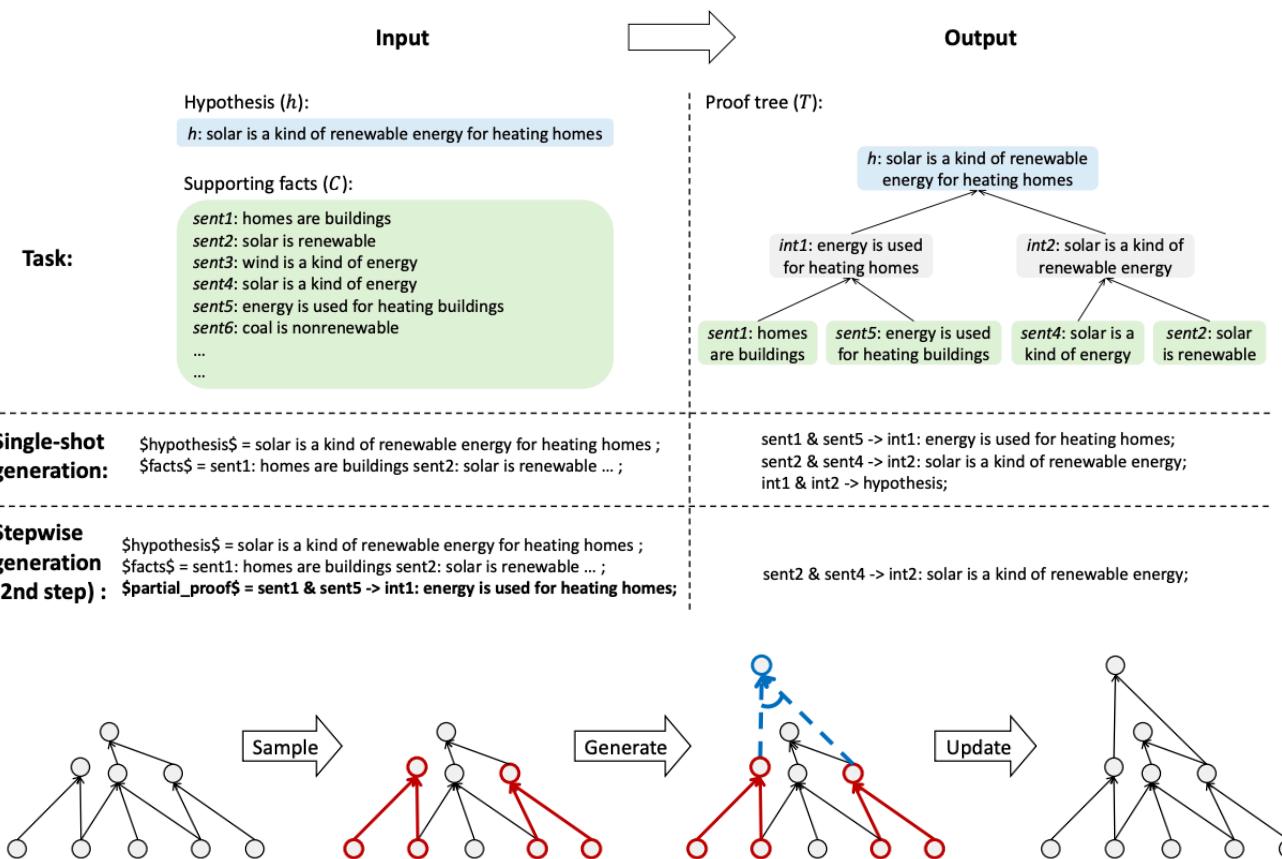
sent2: eruptions produce ash clouds.

sent5: ash blocks sunlight.

Explaining Answers with Entailment Trees

# Generating proofs: NLProofS

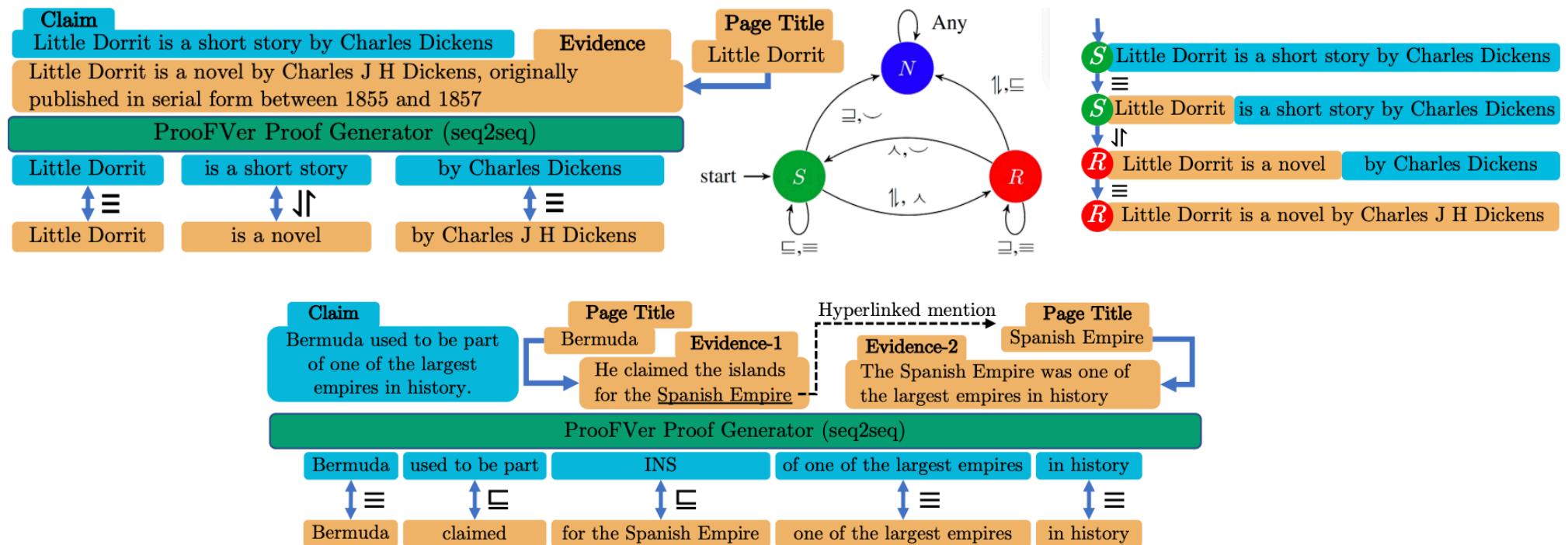
- Task definition
  - Derive hypothesis from a subset of C through reasoning steps creating a proof tree.
- Algorithm
  - (1) a stepwise prover for generating candidate proof steps (T5 model)
  - (2) a verifier for scoring the validity of proofs (RoBERTa fine-tuned)
  - (3) an algorithm for searching for proofs that have high aggregated proof scores



Generating Natural Language Proofs with Verifier-Guided Search, EMNLP 2022

## Application: Fact Verification

- Task: Given a claim, retrieve sentences to prove or disprove it



ProoFVer: Natural Logic Theorem Proving for Fact Verification, TACL 2022

# Multi-Document Fact Verification

## The AdMIRaL pipeline

**Claim:** The 66th Primetime Emmy Awards was hosted by an Iraqi comedian born in 1973.

### Evidence Documents:

#### 66th Primetime Emmy Awards

The 66th Primetime Emmy Awards honored the best in U.S. prime time television programming from June 1, 2013 until May 31, 2014, as chosen by the Academy of Television Arts & Sciences. Comedian and Late Night host **Seth Meyers hosted the ceremony** for the first time.

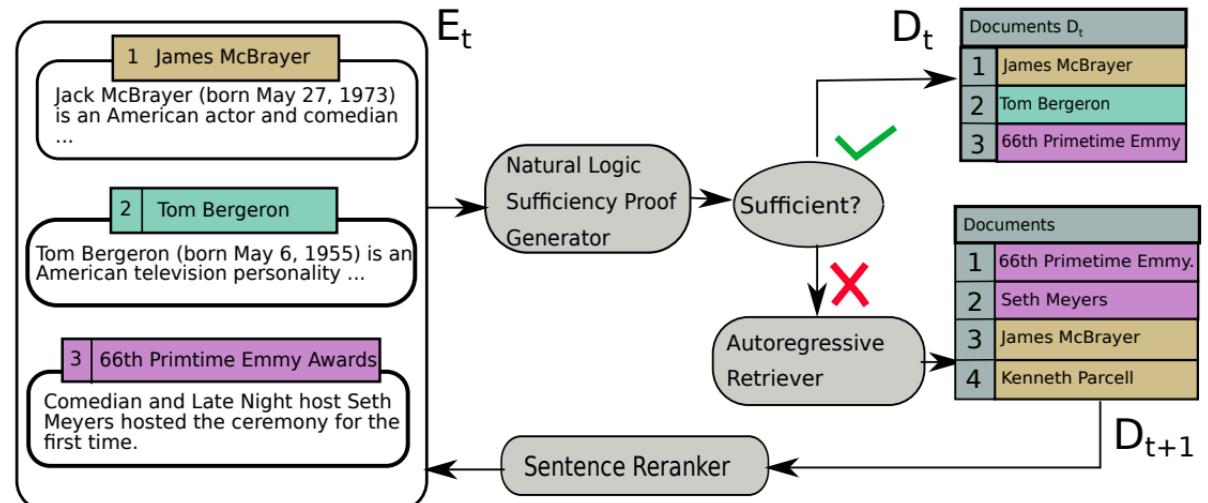
#### Seth Meyers

Seth Adam Meyers (**born December 28, 1973**) is an **American comedian**, writer, producer, political commentator, actor, media critic, and television host. He hosts Late Night with Seth Meyers, a late-night talk show on NBC. Prior to that, he was a cast member and head writer for NBC's Saturday Night Live (2001–2014).

### Verdict: Refuted

Figure 1: A FEVER example where multiple documents are required for verification (relevant evidence in red).

**Claim:** The 66th Primetime Emmy Awards was hosted by an Iraqi comedian born in 1973.



Natural Logic-guided Autoregressive Multi-hop Document Retrieval for Fact Verification, EMNLP 2022

# Reasoning and NLI Summary

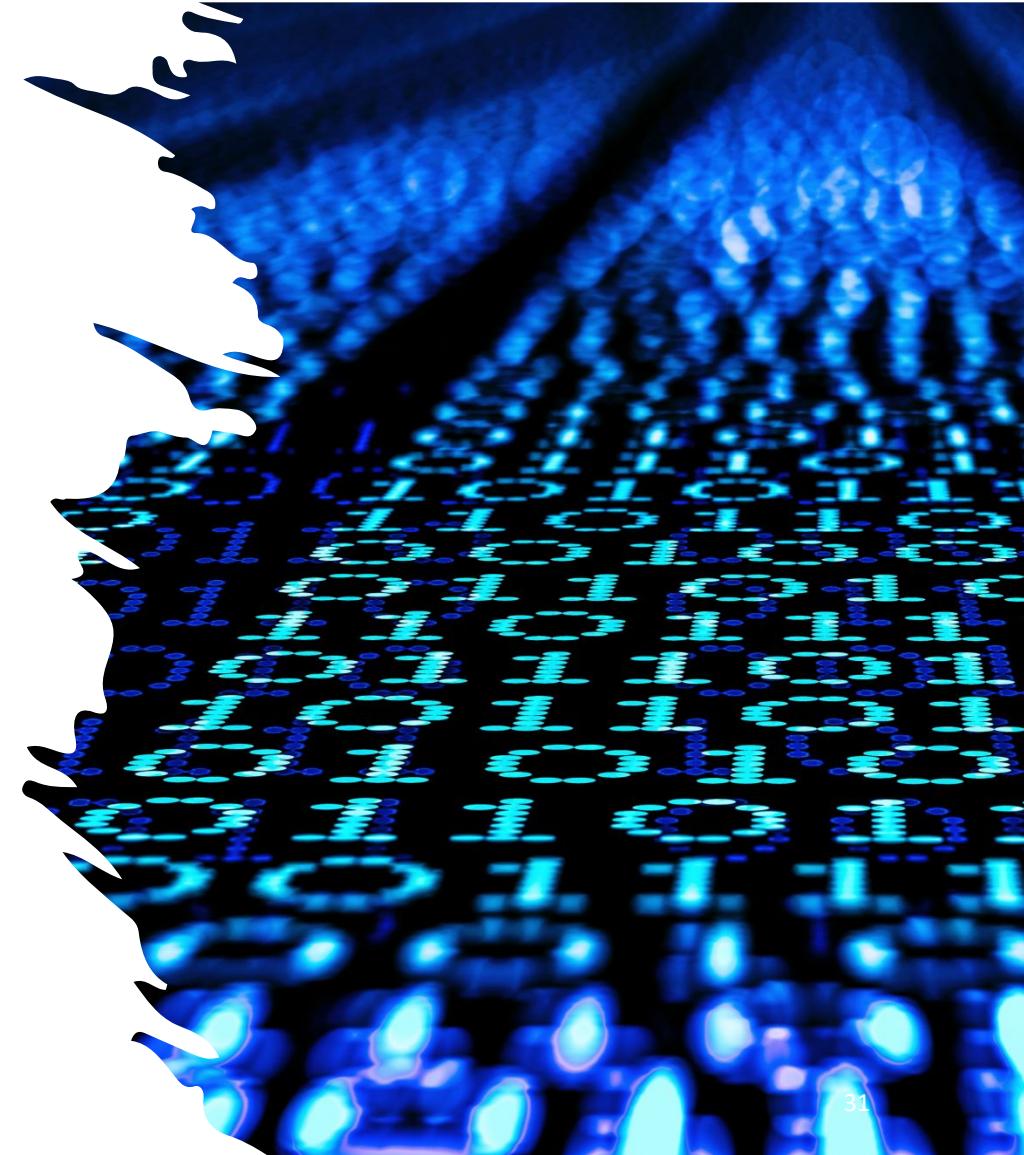
---

- NLI with neural models is open research area
- Applications to fake news detection including multi-hop reasoning across retrieved documents
- Bringing together CoT reasoning with NLI proof approaches remains an open research area



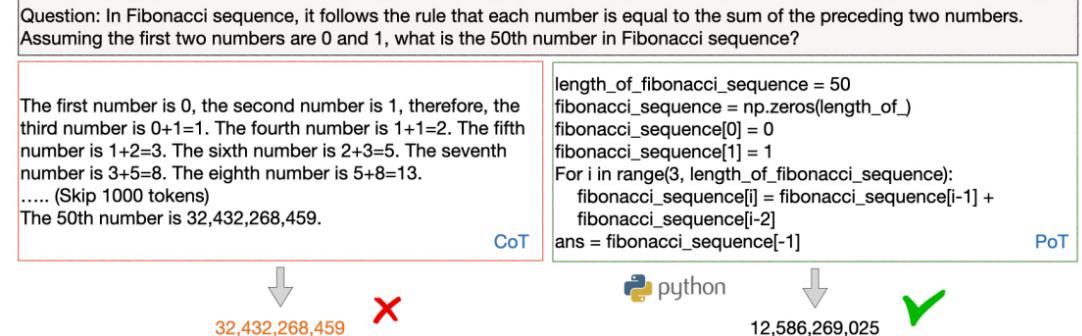
# Part IV: CoT to Tool Use and Program Synthesis

ECIR 2023 Tutorial: Neuro-Symbolic  
Representations for IR – Part 3.1



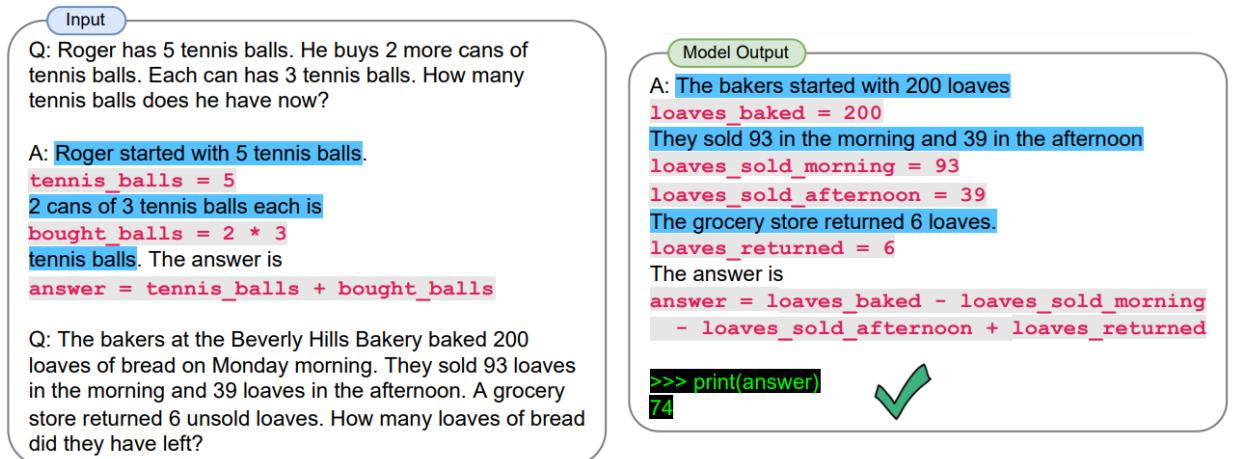
## Beyond text: Code-based Hybrid Output

- Program of Thought (PoT)
  - CoT with Python executable output to solve the task.

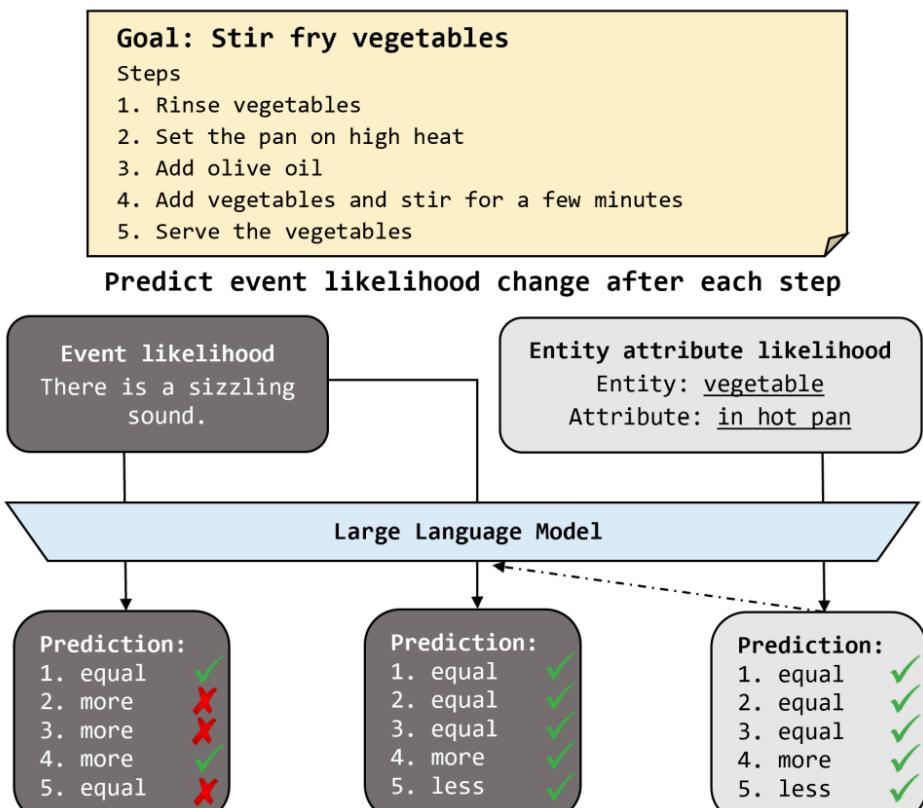


Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks

- Program-aided LLMs
  - Text and code output that is interpreted



# QA and Procedural Reasoning



Goal: Wash sneakers  
 Context: I remove shoelaces. I rinse.  
 Question: What is the likelihood that my feet get wet by wearing the sneakers?  
 Follow up: Are the sneakers wet?  
 Intermediate answer: Yes  
 Follow up: Will my feet get wet by wearing wet sneakers?  
 Intermediate answer: Yes  
 Answer: likely

```

class Wash_Sneakers():
    # Init
    # Remove shoelaces
    # Rinse
    def init(self, event0, subevent0):
        self.event0 = event0 # My feet get
        wet by wearing the sneakers.
        self.event0.subevent = subevent0 #
        The sneakers are wet
    def remove_shoelaces(self):
        self.event0.subevent.change =
            "equally likely" # The sneakers
            are wet
        self.event0.change = "equally likely"
            "# My feet get wet by wearing
            the sneakers."
    def rinse(self):
        self.event0.subevent.change =
            "more likely" # The sneakers are
            wet
        self.event0.change = "more likely" #
            My feet get wet by wearing the
            sneakers.
  
```

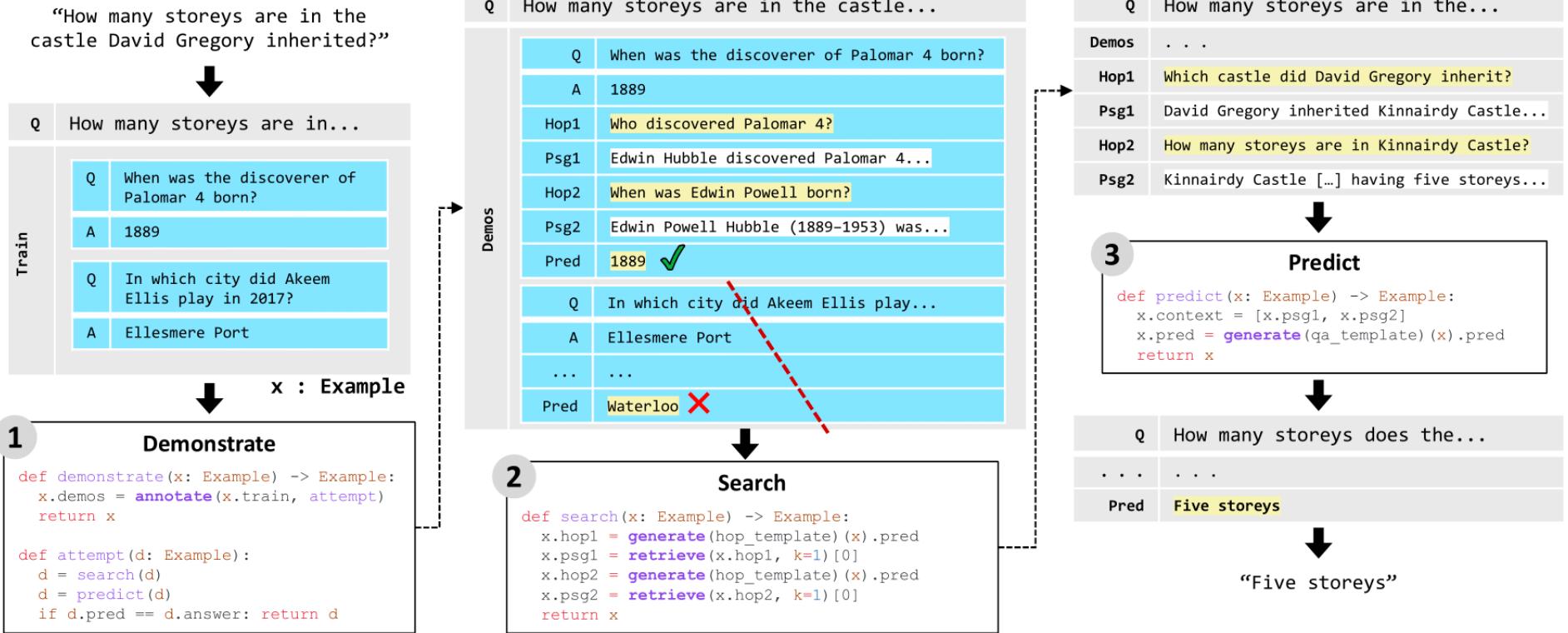
# DSP: Rethinking Search/QA Pipelines

How many storeys are in the castle David Gregory inherited?		
<b>Vanilla LM</b>	LM: Castle Gregory has <u>three</u> storeys.	✗ Hallucinates a fictitious castle
<b>Retrieve-then-Read</b>	RM: "St. Gregory Hotel is a nine-floor boutique hotel in D.C..." LM: St. Gregory Hotel has <u>nine</u> storeys.	✗ Retrieves a different building
<b>Multi-Hop DSP Program</b>	LM: "Which castle did David Gregory inherit?" RM: "David Gregory inherited Kinnairdy Castle in 1664..." LM: "How many storeys does Kinnairdy Castle have?" RM: "Kinnairdy Castle is a tower house, having five storeys..."  LM: <b>Kinnairdy Castle</b> has <u>five</u> storeys.	✓

LM = Language Model  
RM = Retrieval Model

DEMONSTRATE–SEARCH–PREDICT: Composing retrieval and language models for knowledge-intensive NLP

# DSP: Rethinking Search/QA Pipelines



DEMONSTRATE–SEARCH–PREDICT: Composing retrieval and language models for knowledge-intensive NLP

<https://github.com/stanfordnlp/dsp>

## DSP QA Results

- LM = GPT-3.5, RM=ColBERTv2

	Open-SQuAD		HotPotQA		QReCC	
	EM	F1	EM	F1	F1	nF1
<b>Vanilla LM</b>	16.2	25.6	28.3	36.4	29.8	18.4
<b>No-retrieval LM SoTA</b>	20.2 <sup>¶</sup>	–	33.8 <sup>¶</sup>	44.6 <sup>¶</sup>	–	–
<b>Retrieve-then-Read</b>	33.8	46.1	36.9	46.1	31.6	22.2
<b>Self-ask w/ ColBERTv2 Search</b>	9.3	17.2	25.2	33.2	–	–
<b>+ Refined Prompt</b>	9.0	15.7	28.6	37.3	–	–
<b>Retrieval-augmented LM SoTA</b>	34.0 <sup>¶</sup>	–	35.1 <sup>¶</sup>	–	–	–
<b>Task-aware DSP Program</b>	<b>36.6</b>	<b>49.0</b>	<b>51.4</b>	<b>62.9</b>	<b>35.0</b>	<b>25.3</b>

```
def convqa_attempt(d: Example) -> Example:
    d.demos = all_but(demos, d) # all (raw)
    examples that don't intersect with the
    conversation of d

    d = convqa_search(d, k=2)
    if max(precision(d.answer, p) for p in
           d.passages) < .8: return None # skip examples
    where search fails

    d = convqa_predict(d, n=20)
    if max(F1(c.pred, d.answer) for c in
           d.candidates) < .75: return None # skip
    examples where predict fails out of n=20
    attempts

    return d
```

DEMONSTRATE–SEARCH–PREDICT: Composing retrieval and language models for knowledge-intensive NLP

# Conclusion

- LLMs are building blocks to encode knowledge
- LLMs support new explanation-based capabilities (CoT and PoT) that improve their effectiveness for a variety of NLP and search tasks
- Formal NLI symbolic approaches are effective using LLMs and retrieval for challenging multi-document reasoning tasks
- Hybrid pipelines of language and code (“tool use”) is a possible future of complex search and assistants



# New Interactive Knowledge Assistance Track (iKAT) for 2023!

<https://www.trecikat.com/>

- New and diverse types of information needs
- Focuses on **system understanding of user knowledge**
  - Structured task context and PKGs
- Richer types of **interaction**
- Generation: measure **knowledge grounding** of responses
- Discuss with the new organizers
  - Mohammed, Shubham, Leif, Jeff