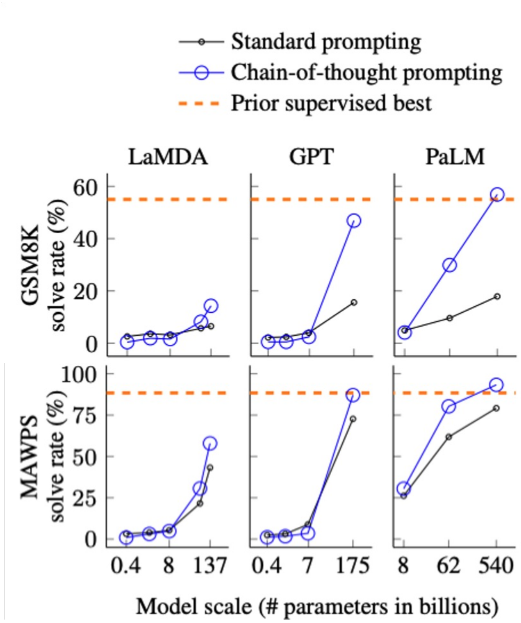
WikiWhy: Answering and Explaining Cause-and-Effect Questions

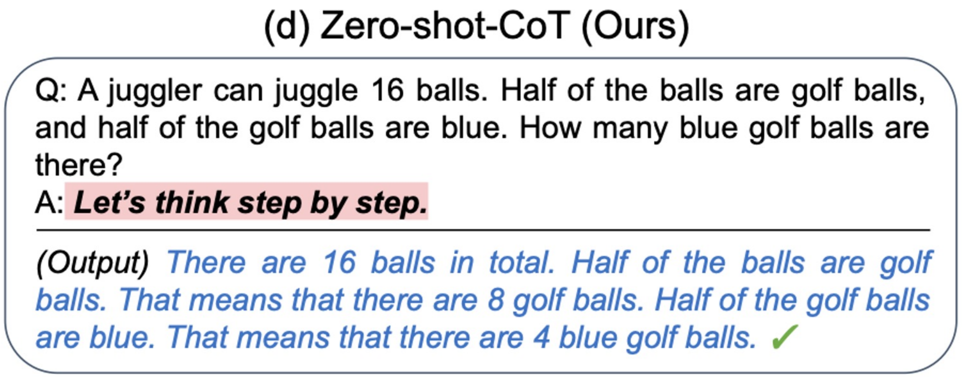
Matthew Ho, Aditya Sharma, Justin Chang,
Michael Saxon, Sharon Levy, Yujie Lu, William Yang Wang

ICLR 2023

UC SANTA BARBARA

Motivation





(d) Zero-shot-CoT (Ours)

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

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

Left: Wei et al. (2022)
Right: Kojima et al. (2022)

Motivation

LLMs are **black boxes**

- Understanding/Interpreting Predictions is difficult
- Making assessing their more involved capabilities is difficult

Input
?
Output

Motivation

QA is the intuitive approach to probe reasoning

- Most QA datasets contain factoid questions
  - Who, What, When, Where
- These questions can be solved with degenerate strategies

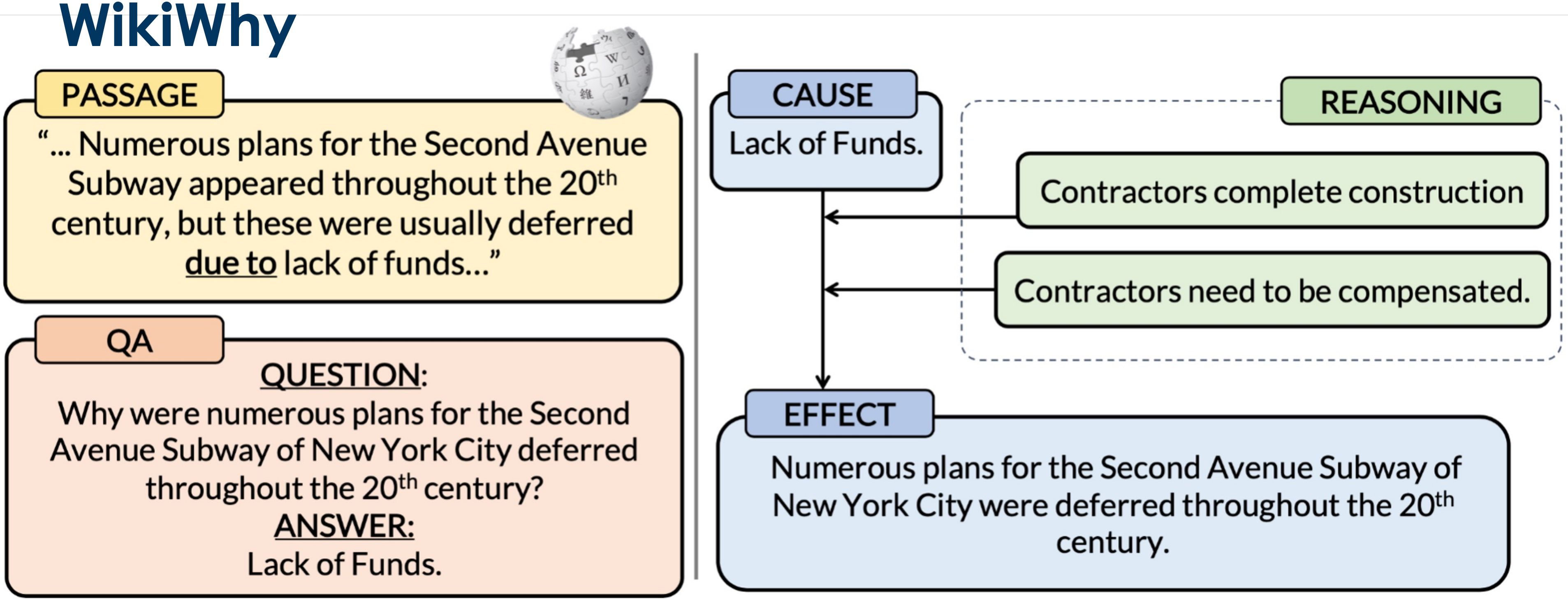
Input
?
Output

How can we distinguish memorization from reasoning?

Explanation as a Task

- Following pedagogy, we can ask **“but why though?”**
  - Requires knowledge and application of underlying mechanics
- Explanation as a task confers
  1. Interpretability
  2. Diagnosis + Debugging
  3. **Insight into understanding**
- Especially important with poorly understood techniques such as CoT

WikiWhy



PASSAGE

"... Numerous plans for the Second Avenue Subway appeared throughout the 20th century, but these were usually deferred due to lack of funds..."

QA

QUESTION:

Why were numerous plans for the Second Avenue Subway of New York City deferred throughout the 20th century?

ANSWER:

Lack of Funds.

CAUSE

Lack of Funds.

REASONING

Contractors complete construction

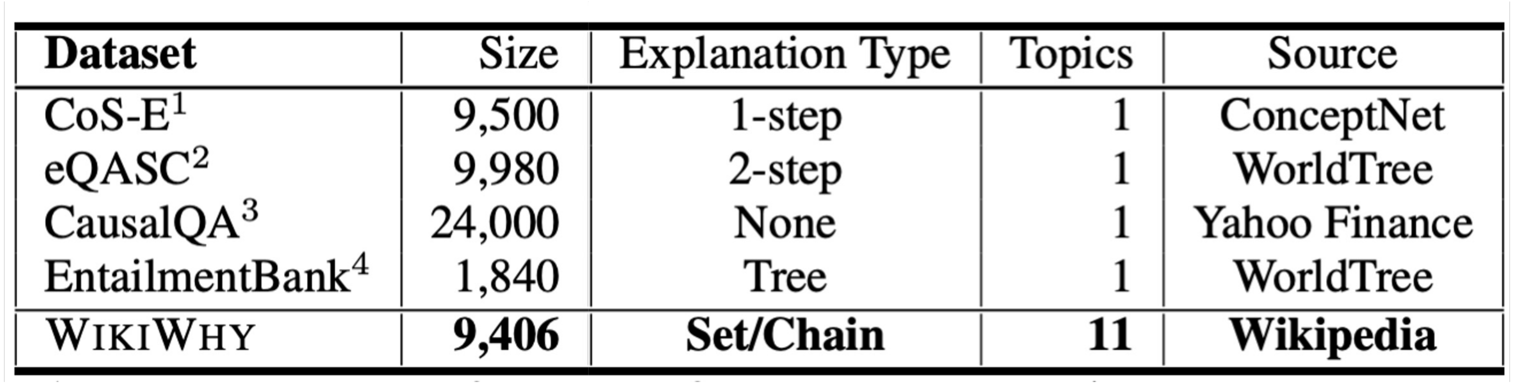
Contractors need to be compensated.

EFFECT

Numerous plans for the Second Avenue Subway of New York City were deferred throughout the 20th century.

WikiWhy contains over 9,000 "why" question-answer-rationale triples.

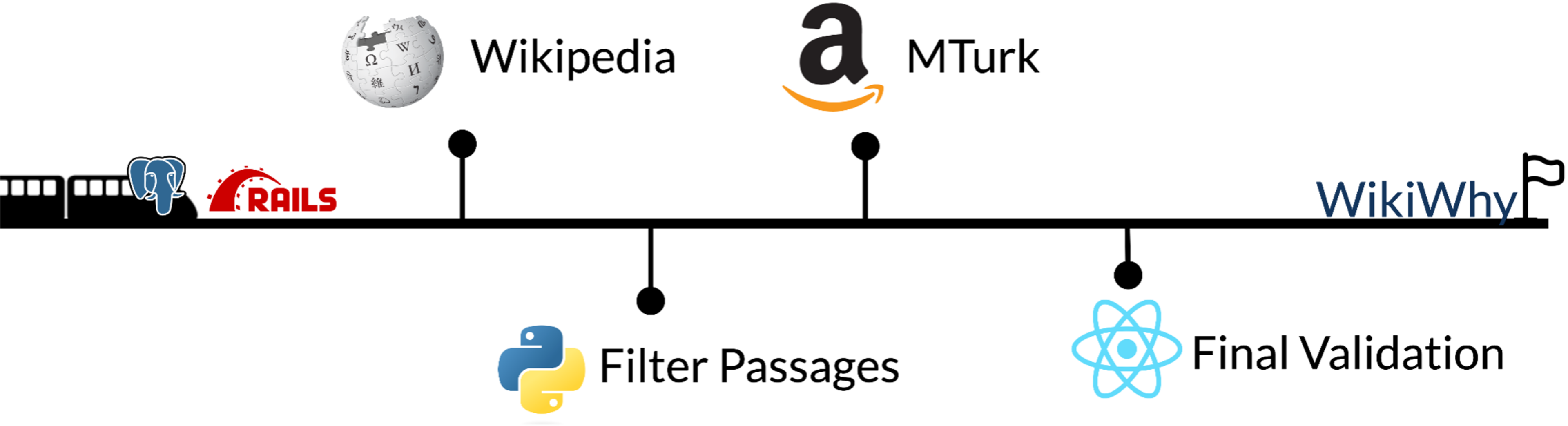
Related Benchmarks



Dataset	Size	Explanation Type	Topics	Source
CoS-E[1]	9,500	1-step	1	ConceptNet
eQASC[2]	9,980	2-step	1	WorldTree
CausalQA[3]	24,000	None	1	Yahoo Finance
EntailmentBank[4]	1,840	Tree	1	WorldTree
WIKIWHY	**9,406**	**Set/Chain**	**11**	**Wikipedia**

WikiWhy presents detailed explanations across a variety of domains.

Data Collection Pipeline



Wikipedia

MTurk

Filter Passages

Final Validation

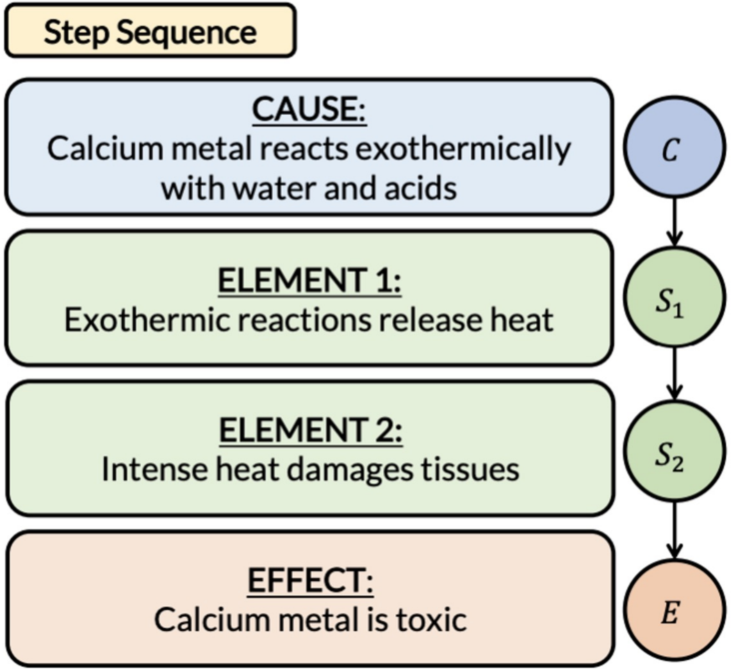
WikiWhy

Task(s) Formulation and Benefits

1. Vanilla QA
   - Susceptible to pattern matching

1. **Explaining *between* Cause and Effect**
   - Reasoning Task: Requires **Application**
   - In-Between the Lines: De-emphasizes **Memorization**
   - Generation over Extraction: Reduces **Artifact Exploitation**
   - Fixed Cause: Simplifies **Evaluation**

Explanation Topologies



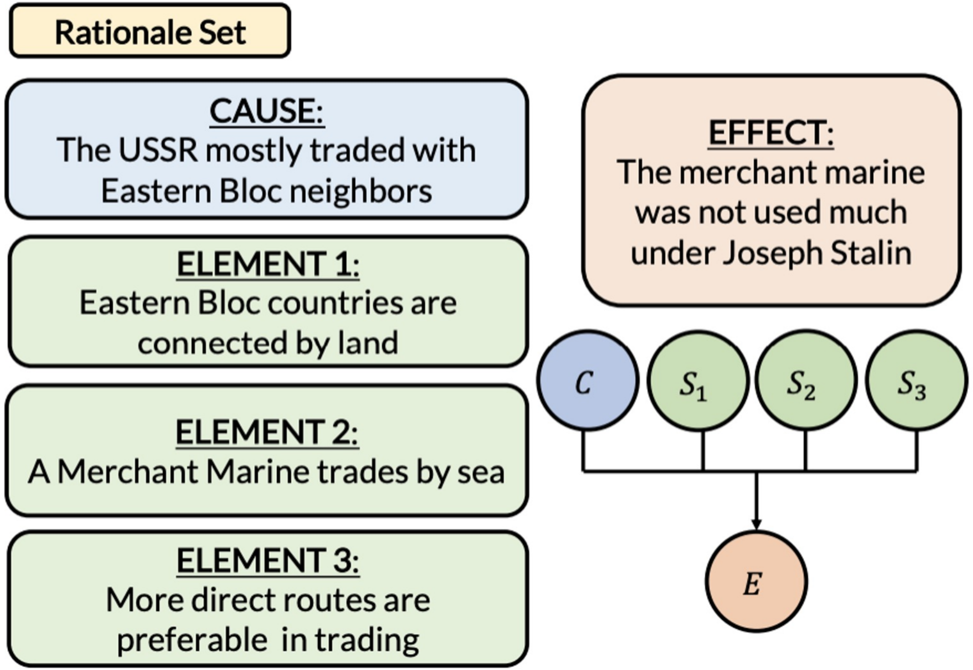
Step Sequence

CAUSE:
Calcium metal reacts exothermically with water and acids

ELEMENT 1:
Exothermic reactions release heat

ELEMENT 2:
Intense heat damages tissues

EFFECT:
Calcium metal is toxic



Rationale Set

CAUSE:
The USSR mostly traded with Eastern Bloc neighbors

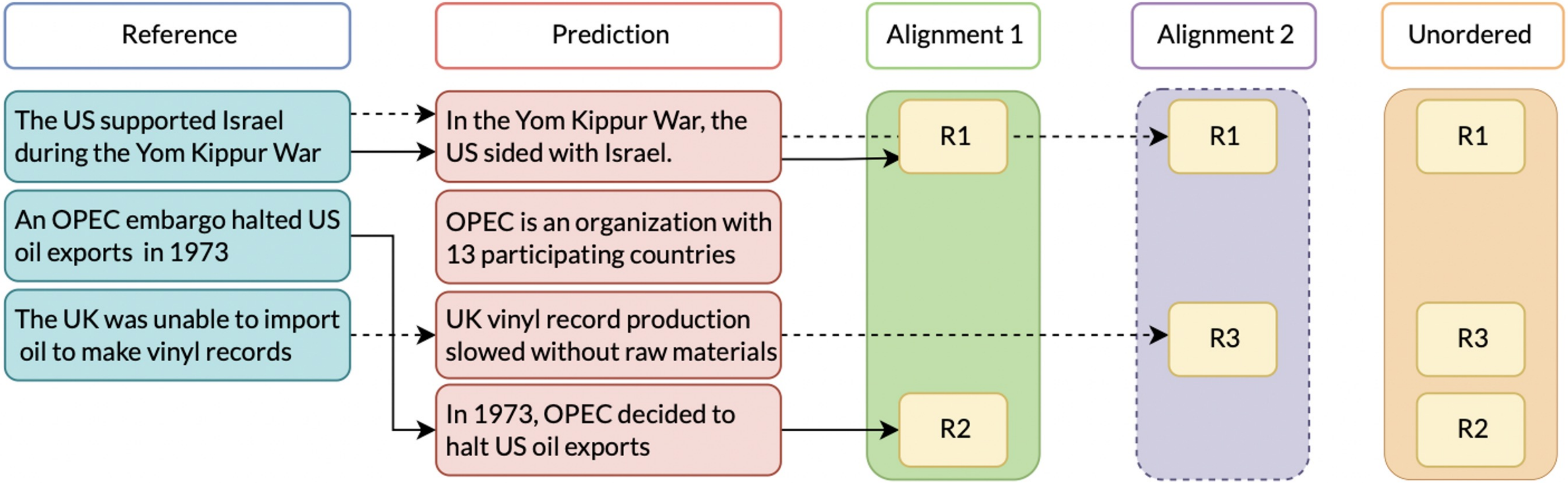
ELEMENT 1:
Eastern Bloc countries are connected by land

ELEMENT 2:
A Merchant Marine trades by sea

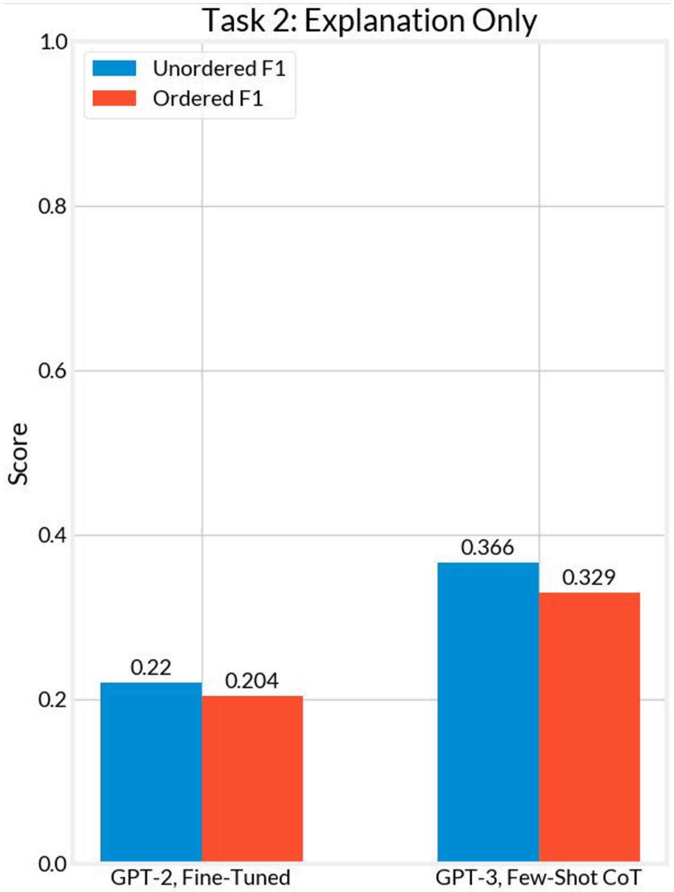
ELEMENT 3:
More direct routes are preferable in trading

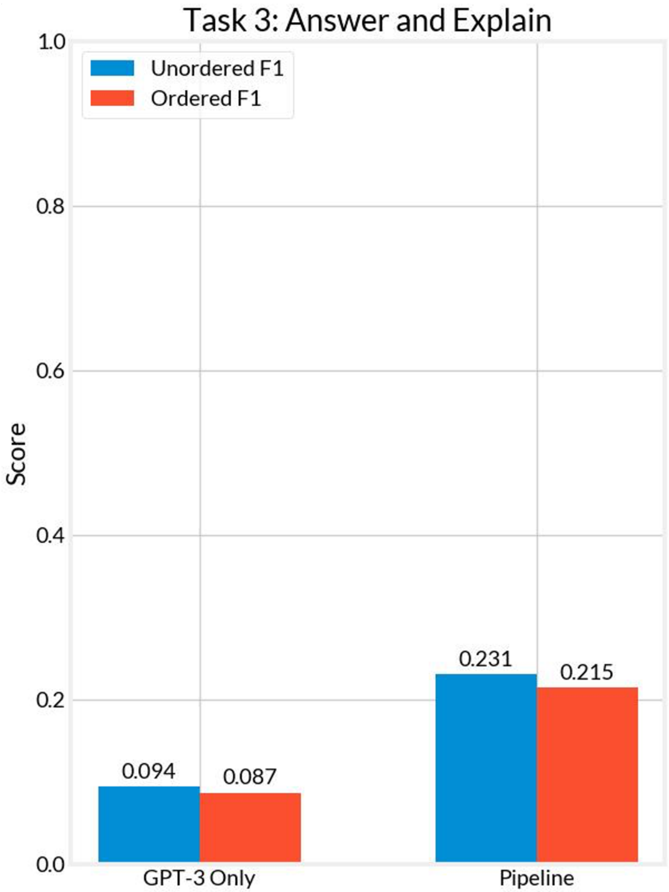
EFFECT:
The merchant marine was not used much under Joseph Stalin

Automatic Evaluation Metric


Reference
Prediction
Alignment 1
Alignment 2
Unordered
The US supported Israel during the Yom Kippur War
An OPEC embargo halted US oil exports in 1973
The UK was unable to import oil to make vinyl records
In the Yom Kippur War, the US sided with Israel.
OPEC is an organization with 13 participating countries
UK vinyl record production slowed without raw materials
In 1973, OPEC decided to halt US oil exports
R1
R2
R1
R3
R1
R3
R2

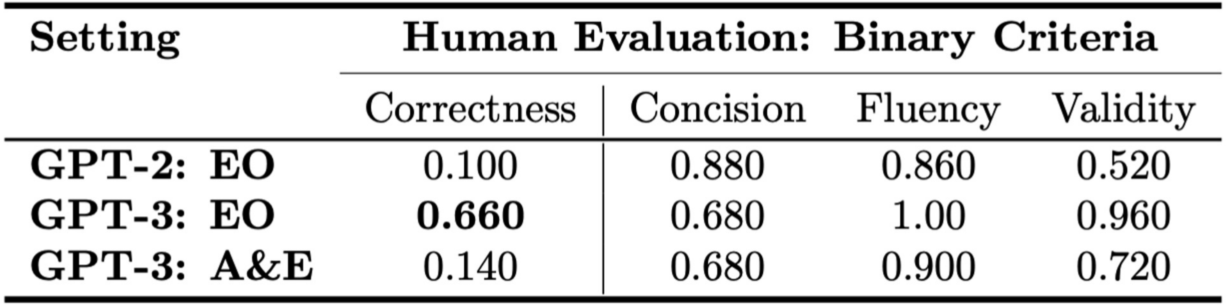
Results



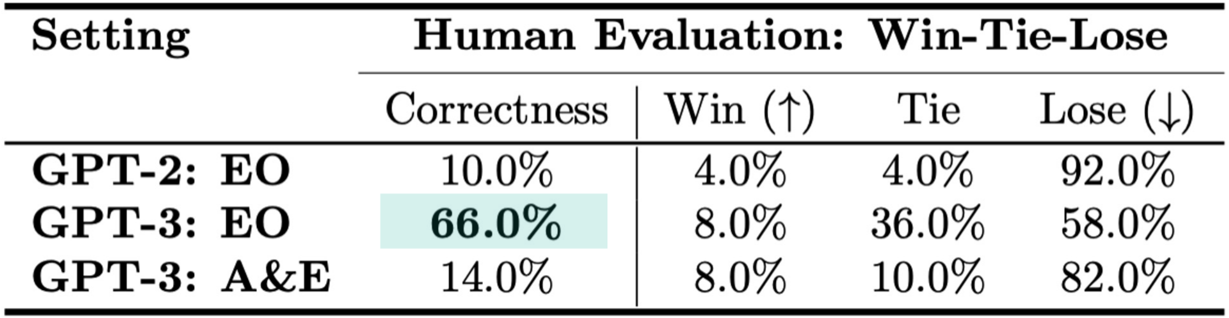


Note: we use `text-davinci-002` for all GPT-3 baselines

Human Evaluation



Setting	**Human Evaluation: Binary Criteria**			
	Correctness	Concision	Fluency	Validity
GPT-2: EO	0.100	0.880	0.860	0.520
GPT-3: EO	**0.660**	0.680	1.00	0.960
GPT-3: A&E	0.140	0.680	0.900	0.720



Setting	**Human Evaluation: Win-Tie-Lose**			
	Correctness	Win (↑)	Tie	Lose (↓)
GPT-2: EO	10.0%	4.0%	4.0%	92.0%
GPT-3: EO	**66.0%**	8.0%	36.0%	58.0%
GPT-3: A&E	14.0%	8.0%	10.0%	82.0%

Future Work

Models

- WT5
- Llama/Alpaca
- ChatGPT
- GPT-4

Prompting Techniques

- Self-Consistency
- Maieutic Prompting
- Subgoal Search
- Least-to-Most

Better Metrics

- Phrasing Robustness
- Logical Validity
- Explanatory Completeness

Thank You!

Poster Session @ 4:30pm-6:30pm

- github.com/matt-seb-ho/WikiWhy
- arxiv.org/abs/2210.12152

UC SANTA BARBARA