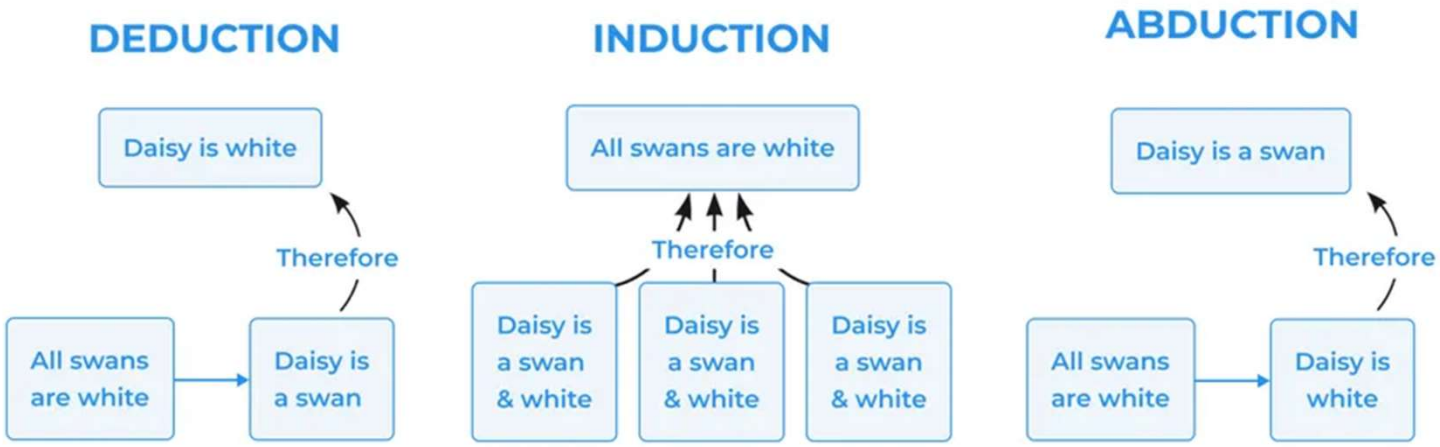
Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners

2024.02.14

서승배

Reasoning



- 소개할 논문에서는 위와 같이 자연어로 표현된 연역, 귀납, 귀추 세 종류의 reasoning task를 다룸

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Dual Process Theory

시스템 1 : 직관적인 빠른 사고

- 고속으로 자동적으로 행하여 멈출 수 없음
- 생각해야하는 노력이 거의 불필요
- 인상을 바로 느끼거나 연상이 가능
- 편견이 있음

시스템 2 : 논리적에 느린 사고

- 시스템 1에서 답이 없을 때 행함
- 생각하는데 주의력이 필요
- 논리적 / 통계적 사고 가능
- 최종 결정권은 시스템 2가 가짐

- Dual Process Theory에 따르면, 인간의 판단과 의사결정은 직관적인 system 1과 논리적인 system 2 사이의 상호작용으로 결정된다.
- System 1은 뇌의 limbic system, system 2는 뇌의 frontal lobe와 관련되어 있다고 판단되어, 두 시스템은 주로 다루는 뇌의 부위가 다르다.
- 인간은 system 1으로 즉각적인 판단을 하며, 논리적인 사고가 필요한 문제라고 판단되면, system 2로 스위칭할 수 있다.

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Dual Process Theory

	시스템1(intuitive)	시스템2(deliberate)
인지적 스타일	휴리스틱	알고리듬적
인지적 인식 여부	낮음	높음
의식적 통제	낮음	높음
자동성	높음	낮음
속도	빠름	느림
신뢰성	낮음	높음
노력	낮음	높음
감정적 균형	높음	낮음

- Dual Process Theory에 따르면, 인간의 판단과 의사결정은 직관적인 system 1과 논리적인 system 2 사이의 상호작용으로 결정된다.
- System 1은 뇌의 limbic system, system 2는 뇌의 frontal lobe와 관련되어 있다고 판단되어, 두 시스템은 주로 다루는 뇌의 부위가 다르다.
- 인간은 system 1으로 즉각적인 판단을 하며, 논리적인 사고가 필요한 문제라고 판단되면, system 2로 스위칭할 수 있다.
- 현재의 LLM은 어떤 system에 더 가깝다고 할 수 있을까?

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Abstract

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners

Xiaojuan Tang[1,3]*, Zilong Zheng[3]*, Jiaqi Li[3],
Fanxu Meng[1,3], Song-Chun Zhu[1,2,3], Yitao Liang[1,3]†, Muhan Zhang[1,3]†

[1]Peking University, [2]Tsinghua University
[3] National Key Laboratory of General Artificial Intelligence, BIGAI

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Introduction

논문 설명

- 배경과 문제점
  - 최근 CoT 등등 프롬프팅을 기반으로 한 LLM의 추론 능력이 주목받고 있음.
  - LLM이 정말로 인간과 비슷한 방식으로 논리적 추론을 수행하는지 아직 알 수 없으며, 복잡한 논리적 추론 문제에 대해 여전히 한계를 보이고 있음.
- 목표
  - LLM이 인간과 비슷하게 system 2 방식의 논리적 사고가 가능한지, 혹은 system 1처럼 휴리스틱적인 방식에 추론을 의존하는지 알아보고자 함
  - 구체적으로는, LLM에서 language token들 사이의 학습된 semantics가 추론에 중요한 역할을 한다고 가정하고, 이를 증명하기 위한 실험을 설계하였음
- 방법론
  - 귀납, 연역, 귀추 세 가지 추론 방식을 포함하며 multi-hop reasoning을 요구하는 추론 데이터셋에서 semantics를 제거하지 않은 것과 제거한 것 사이의 추론 정확도를 비교하였다.
- 결과
  - Semantics가 LLM의 추론에 중요한 역할을 한다는 점을 알았으며, semantics가 commonsense와 일치할 때보다(ex. 하늘은 파랗다) semantics가 commonsense와 일치하지 않을 때(ex. 하늘은 갈색이다) 추론의 성능이 크게 하락하는 것을 발견하였다.
  - 실험 결과는 LLM이 추론 능력에 있어 system 2보다는 system 1에 가깝다는 것을 암시하며, 휴리스틱적이며 신뢰성이 낮은 추론을 수행한다는 점을 알 수 있었다.

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Method: Decoupling Semantics from In-Context Reasoning

Task Definitions

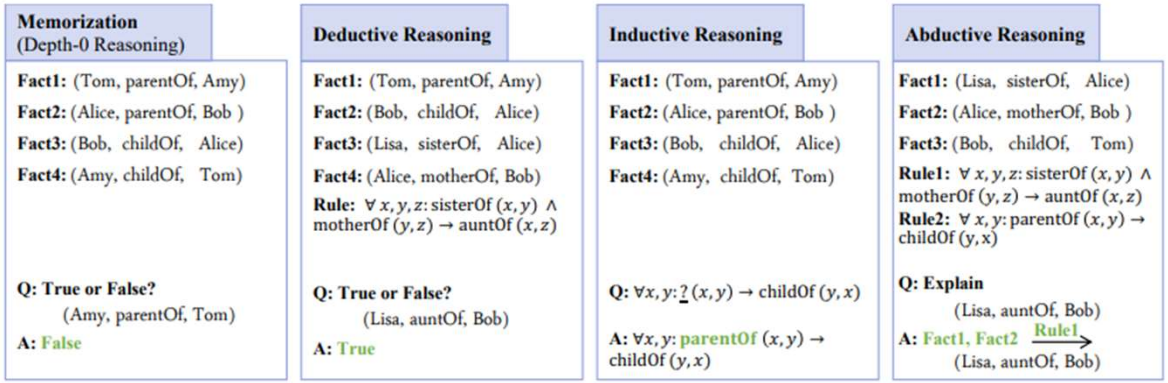


Figure 1: Task Definitions. **Memorization**: retrieving the predicted fact from in-context knowledge. **Deductive**: predicting the correctness of the predicted fact given rules and facts. **Inductive**: generating a rule based on multiple facts with similar patterns. **Abductive**: explaining the predicted fact based on given rules and facts.

- Memorization과 reasoning task로 실험하였으며, 데이터셋에서 semantics를 제외한 세팅과 semantics를 보존한 세팅 사이의 accuracy를 비교함.
- 그림에서는 논리식으로 실험한 것처럼 나와 있는데, natural language와 논리식에 대해서 둘 다 실험을 하였다.

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Method: Decoupling Semantics from In-Context Reasoning

Evaluation Datasets

Deductive Reasoning

Fact1: (Tom, parentOf, Amy)
Fact2: (Bob, childOf, Alice)
Fact3: (Lisa, sisterOf, Alice)
Fact4: (Alice, motherOf, Bob)
Rule: $\forall x, y, z$: sisterOf $(x, y) \wedge$ motherOf $(y, z) \rightarrow$ auntOf (x, z)

Q: True or False?
(Lisa, auntOf, Bob)
A: True

Inductive Reasoning

Fact1: (Tom, parentOf, Amy)
Fact2: (Alice, parentOf, Bob)
Fact3: (Bob, childOf, Alice)
Fact4: (Amy, childOf, Tom)

Q: $\forall x, y$: ? $(x, y) \rightarrow$ childOf (y, x)

A: $\forall x, y$: parentOf $(x, y) \rightarrow$ childOf (y, x)

Abductive Reasoning

Fact1: (Lisa, sisterOf, Alice)
Fact2: (Alice, motherOf, Bob)
Fact3: (Bob, childOf, Tom)
Rule1: $\forall x, y, z$: sisterOf $(x, y) \wedge$ motherOf $(y, z) \rightarrow$ auntOf (x, z)
Rule2: $\forall x, y$: parentOf $(x, y) \rightarrow$ childOf (y, x)

Q: Explain
(Lisa, auntOf, Bob)
A: Fact1, Fact2 $\xrightarrow{\text{Rule1}}$
(Lisa, auntOf, Bob)

Given a set of rules and facts, you have to reason whether a statement is true or false. Here are some facts and rules:

The bear likes the dog.
The cow is round.
The cow likes the bear.
The cow needs the bear.
The dog needs the squirrel.
The dog sees the cow.
The squirrel needs the dog.
If someone is round then they like the squirrel.
If the bear is round and the bear likes the squirrel then the squirrel needs the bear.
If the cow needs the dog then the cow is cold.

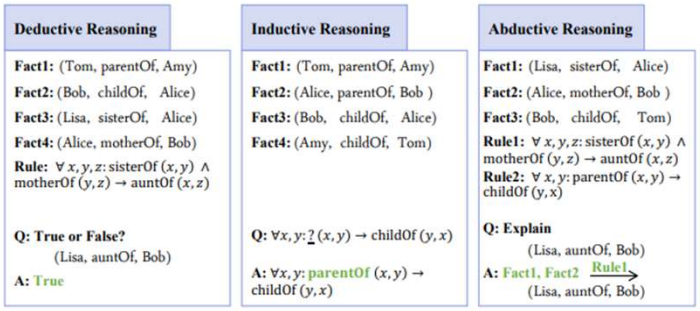
Does it imply that the statement "The cow likes the squirrel." is True?

- Symbolic Tree dataset: 인물 간의 관계에 대한 facts와 rules로부터 새롭게 주어지는 facts에 대한 추론을 수행해야 한다.
- ProofWriter: 비슷하게 facts와 rules로부터 새롭게 주어지는 facts에 대한 참/거짓/알수없음을 추론해야 하는데, 알수없음은 연구에서 제외하였음.

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Method: Decoupling Semantics from In-Context Reasoning

Decoupling Semantics



Deductive Reasoning

Fact1: (Tom, parentOf, Amy)
Fact2: (Bob, childOf, Alice)
Fact3: (Lisa, sisterOf, Alice)
Fact4: (Alice, motherOf, Bob)
Rule: $\forall x, y, z$: sisterOf $(x, y) \wedge$ motherOf $(y, z) \rightarrow$ auntOf (x, z)

Q: True or False?
(Lisa, auntOf, Bob)
A: True

Inductive Reasoning

Fact1: (Tom, parentOf, Amy)
Fact2: (Alice, parentOf, Bob)
Fact3: (Bob, childOf, Alice)
Fact4: (Amy, childOf, Tom)

Q: $\forall x, y$: ? $(x, y) \rightarrow$ childOf (y, x)

A: $\forall x, y$: parentOf $(x, y) \rightarrow$ childOf (y, x)

Abductive Reasoning

Fact1: (Lisa, sisterOf, Alice)
Fact2: (Alice, motherOf, Bob)
Fact3: (Bob, childOf, Tom)
Rule1: $\forall x, y, z$: sisterOf $(x, y) \wedge$ motherOf $(y, z) \rightarrow$ auntOf (x, z)
Rule2: $\forall x, y$: parentOf $(x, y) \rightarrow$ childOf (y, x)

Q: Explain
(Lisa, auntOf, Bob)
A: Fact1, Fact2 $\xrightarrow{\text{Rule1}}$
(Lisa, auntOf, Bob)

Given a set of rules and facts, you have to reason whether a statement is true or false. Here are some facts and rules:

The bear likes the dog.
The cow is round.
The cow likes the bear.
The cow needs the bear.
The dog needs the squirrel.
The dog sees the cow.
The squirrel needs the dog.
If someone is round then they like the squirrel.
If the bear is round and the bear likes the squirrel then the squirrel needs the bear.
If the cow needs the dog then the cow is cold.

Does it imply that the statement "The cow likes the squirrel." is True?

Given a set of rules and facts, you have to reason whether a statement is true or false.
Here are some facts and rules:

The e4 likes the e5.
The e14 is e2.
The e14 likes the e4.
The e14 needs the e4.
The e5 needs the e26.
The e5 sees the e14.
The e26 needs the e5.
If someone is e2 then they like the e26.
If the e4 is e2 and the e4 likes the e26 then the e26 needs the e4.
If the e14 needs the e5 then the e14 is e1.

Does it imply that the statement "The e14 likes the e26." is True?

- Symbolic Tree dataset
  - Semantics를 dataset으로부터 분리하기 위하여, relation name(ex. parent)을 r1으로 지칭하는 등 의미를 판별할 수 없는 형태로 바꾸었음.
  - Entity name을 e1으로 지칭하는 등 entity의 표현도 바꾸어 봤는데, 이 부분은 그다지 결과에 변화가 없었다고 함

- ProofWriter
  - Subject와 object를 entity ID로 대체하였음.
  - Anne is kind -> e1 is e2

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Experiments: Input Forms

- 편의상 semantics를 제외한 버전을 symbols, semantics를 보존한 버전을 semantics setting이라고 한다. (misleading할 수 있는데, symbols setting은 논리식 형식을 말하는 것이 아니라, 방법론에서 설명했던 것처럼 relation이나 entity를 의미를 판별할 수 없게 교체한 것)

		Zero-Shot	Zero-Shot-CoT
Symbols	logic	52.6	56.1
	natural language	49.0	51.1
Semantics	logic	61.4	61.9
	natural language	69.3	64.3

- ChatGPT로 주었을 때 각각의 setting이 자연어 형태와 논리식 형태 중 어느 것이 더 적합한지 실험.
- Symbols setting에서는 논리식 형태로($\forall$, ~, $\wedge$, $\vee$) 주었을 때 더 좋은 결과가 나왔고, semantics setting에서는 자연어 형태로 주었을 때 더 좋은 결과가 나옴
- 아마도 LLM이 학습할 때 e1 $\vee$ e2 아니면 Patrick or Anne 식으로 학습하지, e1 or e2 같은 형태로 학습을 하진 않았을 것 같다.

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Experiments: Semantics Matter in LLMs' reasoning

Table 2: The reasoning results of Symbolic Tree. Results are in %.

Category	Model	Baseline	deduction	induction	abduction
Symbols	ChatGPT	Zero-Shot	52.6	6.10	1.50
		Zero-Shot-CoT	55.7	7.86	4.90
		Few-Shot-CoT	54.8	-	18.2
		Zero-Plus-Few-Shot-CoT	55.7	-	-
	GPT-4	Zero-Shot	68.8	9.28	25.0
		Zero-Shot-CoT	71.1	8.93	31.2
		Few-Shot-CoT	67.6	-	44.2
Semantics	ChatGPT	Zero-Shot	66.1	36.4	2.94
		Zero-Shot-CoT	65.5	32.2	3.40
		Few-Shot-CoT	67.1	-	21.8
		Zero-Plus-Few-Shot-CoT	67.2	-	-
	GPT-4	Zero-Shot	79.2	52.5	27.3
		Zero-Shot-CoT	86.2	53.9	33.4
		Few-Shot-CoT	91.1	-	69.2
Random		-	50.1	3.57	-
Logic-based		-	100	57.1	100

- 앞서 언급한 Symbolic Tree dataset에 대하여 semantics와 symbols 두 가지 환경에서 추론의 accuracy를 테스트.
- 사용 모델은 ChatGPT와 GPT-4
- Semantics 환경이 일관적으로 좋은 결과를 얻었으며, 이는 semantics가 LLM의 reasoning 능력에 중요하다는 저자들의 가설을 뒷받침한다.
- 그러나 Logic-based method와 비교했을 때는, 두 방법 모두 성능이 현격히 떨어진다.

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Experiments: More Fine-grained Analysis about Semantics

Table 4: Semantics, removing rules/facts and counter-commonsense reasoning experiments (ChatGPT and GPT-4). Results are in %.

	deductive (Few-Shot-CoT)		inductive (Zero-Shot-CoT)	
	ChatGPT	GPT-4	ChatGPT	GPT-4
Semantics	71.8	90.0	25.0	53.6
Symbols	53.7	67.6	7.14	21.4
Remove R/F	70.1	90.4	7.14	35.7
Counter-CS	48.9	73.4	7.14	17.8

- 앞서의 실험을 보충하기 위해, 좀더 semantics의 영향을 잘 알아보기 위해 LLM의 안에 저장되어 있을 commonsense knowledge를 이용하기로 함
  - 1. 논리 추론에 필요한 전제를 일부 제외하여 비는 부분은 commonsense를 이용하여 추론을 수행하게 하였음
  - 2. commonsense에 반하는 rule을 in-context로 주었음. 이는 새로운 knowledge에만 의존하여 추론을 진행하도록 LLM에 요구함(ex. 부모이며 남성이면 어머니이다)
- 실험 결과, semantics, deductive setting에서 추론에 필요한 전제를 일부 제거해도 결과가 원래와 비슷한 accuracy를 얻음
  - LLM이 논리적 추론 과정에서 in-context knowledge를 얼마나 효과적으로 사용하는지에 의문을 남김
- Counter-CS setting은 semantics setting보다 성능이 확연히 떨어짐
  - LLM이 의미에 기반하여 추론을 수행하기 때문에, commonsense에 반하는 추론은 좋지 않은 결과가 나온다.

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Experiments: More Fine-grained Analysis about Semantics

Table 3: The deduction results of ProofWriter tasks (ChatGPT). Results are in %.

Category	Baseline	depth-1	depth-2	depth-3	depth-5
Symbols	Zero-Shot	69.1	62.3	59.4	52.8
	Zero-Shot-CoT	56.2	49.4	45.2	38.6
	Few-Shot-CoT	65.8	58.1	-	-
Semantics	Zero-Shot	69.0	63.5	60.3	51.4
	Zero-Shot-CoT	51.5	45.8	40.3	30.9
	Few-Shot-CoT	62.5	56.7	-	-

- 특이하게, ProofWriter는 CoT method가 오히려 성능을 하락시켰는데, 이는 symbolic tree와 다르게, 개별적인 문장들의 의미가 거의 없다는 점 때문일 수도 있다.(ex. Dogs like birds)
- Step-by-step이 이상한 의미의 부작용을 더 크게 만든다.

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

Conclusion and Discussion

- Semantics가 commonsense와 일치할 때, LLM은 괜찮은 추론 성능을 보인다.
- Semantics가 commonsense와 일치하지 않거나 decoupled일 때는 LLM은 in-context new knowledge를 추론에 사용하는 데에 어려움을 겪는다.
- LLM의 추론에서 language token들 사이의 학습된 semantics가 추론에 중요한 역할을 한다.
- Semantics의 영향을 고려하는 새로운 추론 벤치마크가 필요하다.
- External knowledge base를 잘 활용하는 것이 중요할 수 있다.
- LLM이 가지고 있던 지식이 아닌 in-context knowledge를 좀더 추론에 잘 활용하는 방법이 필요하다.

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

더 생각해볼 것

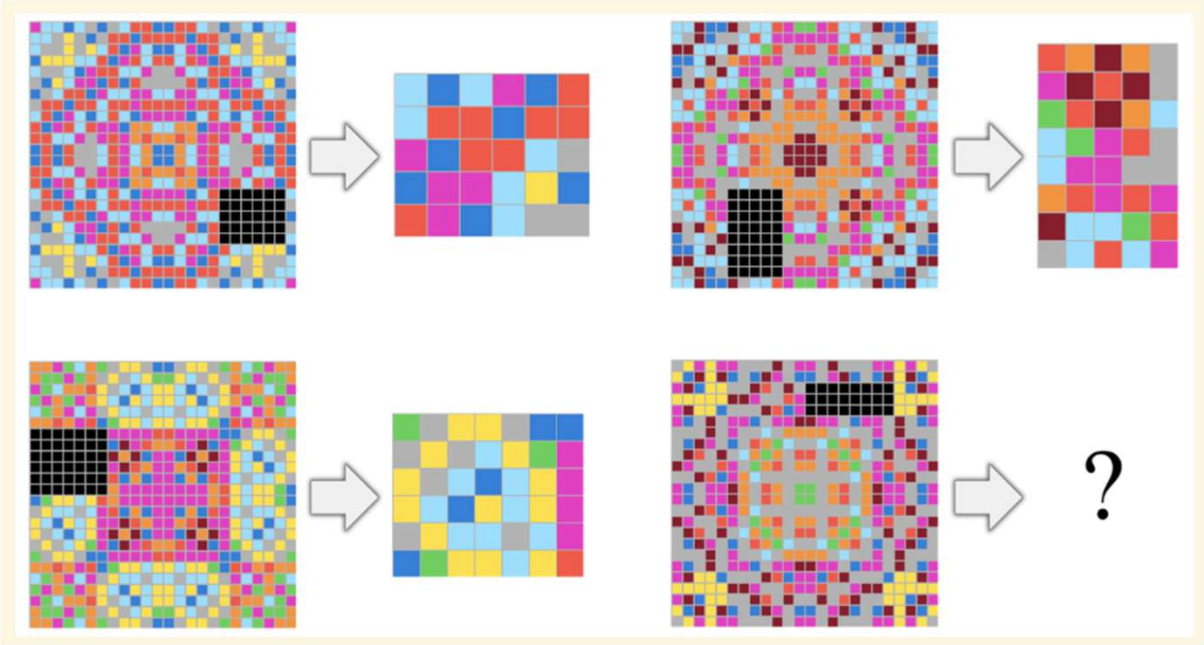
- CoT 등등의 프롬프팅 방식을 이용하여 복잡한 논리적 문제를 간단한 문제로 분해하는 방식이 효과적인 진짜 이유
  - 간단한 문제로 분해하는 것 자체는 인간과 비슷하다고 할 수 있다.
  - 그러나 사실 간단한 문제들은(3+5) 복잡한 문제(241+{5213*2023+(202-23)*412})보다 training data에 존재할 확률이 높다.
  - 만약 정말로 training data에 존재하는 예시들에 의존하는 추론만 가능하다면, training data에 존재하지 않는 예시에 대한 추론은 현재의 LLM만으로 불가능. (Generalization 능력 부족)

- 1D-ARC

- 인간의 눈에는 mirror task가 크게 어려워 보이지는 않지만, mirror task에서만 유독 예측 정확도가 떨어지며 프롬프팅을 동원하여도 잘 안 되는 이유 – 그냥 LLM의 훈련 데이터에 이런 task가 없었던 것 같음.
- 마찬가지로 LLM이 익숙하지 않은 pattern에 대해 generalize하는 능력이 있다기보다는 휴리스틱에 가까워 보임

Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

더 생각해볼 것



Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners (2023.05)

HCC LAB SNU

QnA