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

Abstract

개요

- 논문은 Chain-of-Thought Prompting(CoT Prompting)이 대형 언어 모델의 복잡한 추론 능력 향상 방법을 연구
- CoT Prompting은 프롬프트 내에 일련의 중간 추론 단계를 포함하는 방식이며, 충분히 큰 모델에서 논리적 사고 능력 발현

실험 및 결과

- 세 가지 대형 언어 모델을 대상으로 실험 진행
- CoT Prompting이 산술적, 상식적, 기호적(reasoning) 추론 작업에서 성능 향상
- 특히 PaLM 540B 모델에서 GSM8K 벤치마크(수학 문제)에 대해 SOTA 성능 기록
- 미세 조정된 GPT-3보다도 우수한 결과 달성

---

1. Introduction

개요

- LLM이 최근 NLP 분야에서 혁신을 가져옴
- 모델 크기를 키우면 성능이 향상되지만, 산술적, 상식적, 기호적(reasoning) 추론과 같은 복잡한 작업에서는 단순한 모델 크기 확장이 충분하지 않음

핵심 아이디어

- 자연어 기반 Chain-of-Thought를 생성하면 LLM의 추론 능력이 향상
- 기존 연구:
  - 자연어 기반 중간 단계(reasoning steps)를 학습하는 방식
  - 정형 언어(formal language)를 이용한 neuro-symbolic 방법

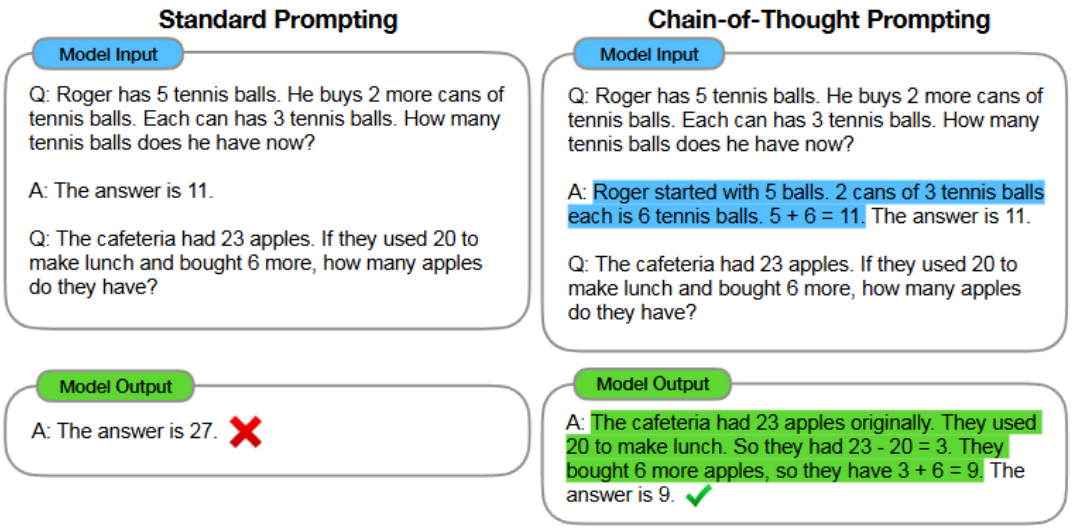
- Few-shot prompting을 이용한 추론 능력 학습

기존 방법의 한계

1. Rationale-augmented training 및 fine-tuning 방식
   - 고품질 중간 추론 데이터 구축 비용이 높음
2. 기존 Few-shot prompting
   - 추론 능력이 필요한 작업에서 성능이 낮고, 모델 크기가 커져도 개선이 미미

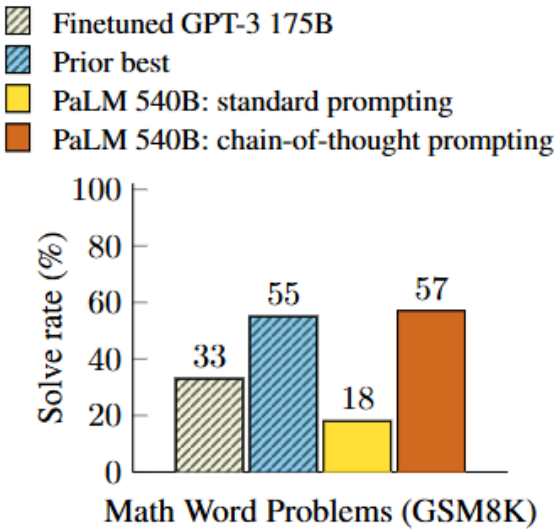
Chain-of-Thought 접근법

- Few-shot prompting을 활용하여 "입력 → 중간 추론 과정(Chain of Thought) → 최종 출력" 구조의 예제를 제공
- CoT를 사용하면 모델이 단순한 정답 예측이 아니라 일련의 논리적 사고 과정을 생성하면서 정답을 도출하도록 유도 가능
- CoT Prompting이 일반적인 Few-shot prompting보다 논리적 추론을 더 잘 수행하는 예시 ⇒



실험 및 결과

- 산술적, 상식적, 기호적(reasoning) 추론 벤치마크에서 CoT Prompting이 표준 Few-shot prompting보다 월등히 높은 성능을 보임
- 특히, PaLM 540B 모델이 GSM8K 벤치마크에서 SOTA 성능을 기록



- Few-shot prompting 방식만으로도 복잡한 추론이 가능하며, 별도의 학습 데이터 없이 다양한 작업 수행 가능

---

2. Chain-of-Thought Prompting

개념 정의

- 인간이 복잡한 문제를 해결할 때 여러 단계를 거쳐 논리적으로 사고하는 과정을 모델에 적용하는 방식
- Chain-of-Thought Prompting은 입력 → Chain-of-Thought → 최종 출력 형태의 프롬프트를 제공하여 모델이 중간 추론 단계를 생성하도록 유도

CoT Prompting 원리

- 충분히 큰 언어 모델은 Chain-of-Thought 예제를 few-shot prompting으로 학습하면 자연스럽게 추론 능력 발현
- 수학 문제, 상식 추론, 기호 조작 등 다양한 문제에서 효과적

CoT Prompting 장점

1. 복잡한 문제 해결 가능
    - 다단계 추론이 필요한 문제에서 높은 성능 발휘
2. 모델의 추론 과정 해석 가능
    - 중간 사고 과정을 제공하여 모델의 응답을 검토 및 디버깅 가능
3. Few-shot prompting만으로 강력한 성능 확보

- 별도의 fine-tuning 없이 프롬프트 예제만으로도 모델의 추론 능력 향상

4. 다양한 분야에 적용 가능
    - 수학, 상식 추론, 기호적 추론 등 여러 작업에 효과적

---

3. Arithmetic Reasoning

3.1 Experimental Setup

실험 목표

- Chain-of-Thought Prompting이 산술적 추론(arithmetic reasoning) 문제 해결에 미치는 영향을 평가
- 특히 Math Word Problems에서 CoT Prompting의 효과를 분석

벤치마크

총 5가지 수학 문제 벤치마크에서 평가:

1. GSM8K
    - 고품질 수학 단어 문제 모음
    - 복잡한 연산 및 다단계 추론 필요
2. SVAMP
    - 구조가 다양한 수학 문제 포함
3. ASDiv
    - 다양한 유형의 수학 문제 포함
4. AQuA
    - 대수학(word algebra problems) 관련 문제
5. MAWPS
    - 기초적인 수학 단어 문제 모음

실험 비교 방법

1. 기존 Few-shot Prompting
    - 일반적인 Few-shot prompting 방식 사용
    - 프롬프트에 단순히 문제-정답(input-output) 예제를 제공

2. Chain-of-Thought Prompting
    - 프롬프트에 "문제 → 중간 추론 과정(Chain-of-Thought) → 정답" 형식의 예제를 제공
3. Baseline: Fine-tuned models
    - 특정 벤치마크에서 GPT-3 fine-tuned과 비교

실험 평가 방법

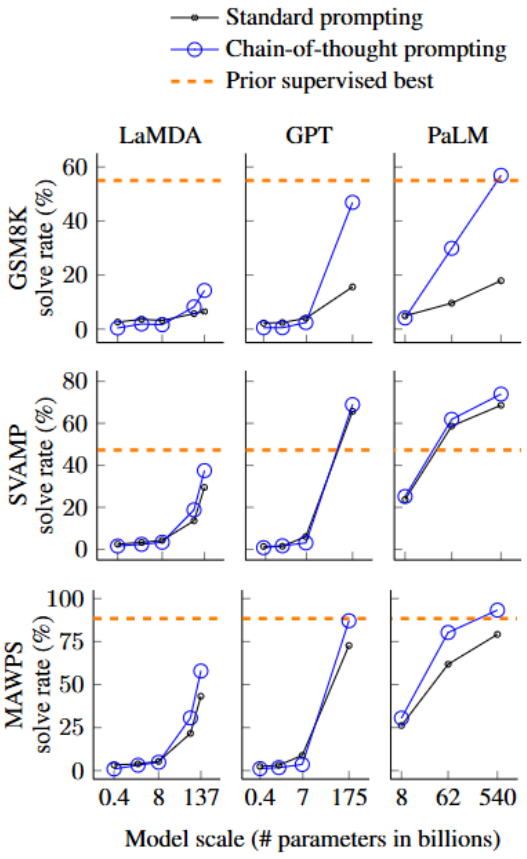
- 모델이 정답을 맞출 확률(정확도) 측정
- Greedy decoding 사용
- LaMDA 모델의 경우 다섯 개의 랜덤 시드로 평균값 측정

3.2 Results

CoT Prompting의 주요 결과

- 산술적 추론에서 기존 Few-shot prompting보다 월등히 높은 성능을 보임
- 특히, PaLM 540B 모델이 GSM8K에서 SOTA 성능 기록
- 모델 크기가 클수록 CoT Prompting의 효과가 더욱 두드러짐

결과 요약



1. Chain-of-Thought Prompting은 대형 모델에서 강력한 성능 향상 제공
    - PaLM 540B에서 GSM8K 벤치마크에서 기존 최상위 모델보다 높은 성능 기록
    - 표준 Few-shot prompting보다 성능이 2배 이상 증가
    - 다른 데이터셋(SVAMP, MAWPS)에서도 일관된 성능 향상
2. 작은 모델에서는 CoT Prompting 효과 제한적
    - 100B 미만의 모델에서는 Chain-of-Thought을 적용해도 성능이 크게 향상되지 않음
    - 작은 모델들은 유창한 답변을 생성하지만 논리적으로 일관되지 않거나 정답을 틀리는 경향
3. 복잡한 문제일수록 CoT Prompting의 이점이 커짐
    - 단순한 문제(One-step 문제)에서는 표준 prompting과 성능 차이가 크지 않음
    - GSM8K처럼 다단계 연산이 필요한 문제에서 가장 큰 성능 향상 확인

3.3 Ablation Study

실험 목적

- CoT Prompting의 성능 향상이 실제로 중간 추론 단계 때문인지 확인
- CoT의 특정 요소를 제거하거나 변형하여 성능 변화를 분석

실험 구성

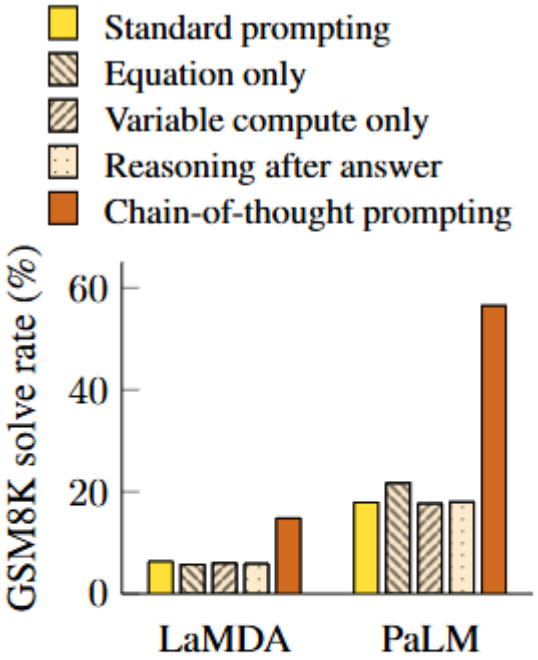


Figure 5: Ablation study for different variations of prompting using LaMDA 137B and PaLM 540B. Results for other datasets are given in Appendix Table 6 and Table 7.

1. Equation Only Prompting
    - 모델이 수학 문제를 풀 때 중간 추론 없이 수식만 출력하도록 유도
    - 결과: GSM8K에서 거의 성능 향상 없음
    - 결론: 자연어로 된 문제를 수식으로 바로 변환하는 것은 어렵기 때문에, CoT의 자연어 추론 과정이 필요
2. Variable Compute Only Prompting
    - Chain-of-Thought Prompting이 유용한 이유가 추론 과정 자체가 아니라 추가적인 토큰을 소비하면서 더 많은 계산을 하기 때문인지 테스트

- 결과: Baseline과 유사한 성능
- 결론: 단순히 더 많은 연산을 수행하는 것이 CoT 효과의 원인이 아님

3. Chain of Thought After Answer
   - Chain-of-Thought이 정답을 유도하는 것이 아니라 단순한 설명에 불과한지 검증
   - 실험 방식:
     - 모델이 먼저 정답을 출력한 후, CoT 과정을 추가
   - 결과: Baseline과 유사한 성능
   - 결론: CoT는 단순한 설명이 아니라, 실제 추론 과정에서 중요한 역할을 함

3.4 Robustness of Chain-of-Thought

실험 목적

- CoT이 프롬프트 예제 및 문장 스타일 변화 등에 대해 얼마나 강인한지 평가
- 다른 사람이 작성한 CoT 예제를 사용해도 성능이 유지되는지 확인

프롬프트 예제 변화

실험 방식

- GSM8K 및 MAWPS 데이터셋을 사용하여 다양한 CoT 예제 변형 실험
- Annotator A, B, C가 독립적으로 작성한 CoT 예제 사용
- CoT 스타일을 단순화한 예제(짧고 간결한 설명) 사용

결과

- Annotator가 다르더라도 CoT Prompting은 여전히 높은 성능 유지
- 간결한 CoT 스타일(짧은 문장)도 성능에 큰 영향 없음
- 기존 Few-shot prompting보다 모든 변형된 CoT에서 성능이 월등히 높음

CoT 예제 출처 변화

실험 방식

- 기존 수작업으로 작성한 CoT 예제 대신, GSM8K 훈련 세트에서 무작위로 선택한 8개의 예제를 사용
- 예제의 출처와 관계없이 CoT가 효과적인지 확인

결과

- 무작위로 선택한 GSM8K 예제 사용 시에도 CoT Prompting이 여전히 높은 성능 유지
- 표준 Few-shot prompting보다 일관된 성능 향상 확인

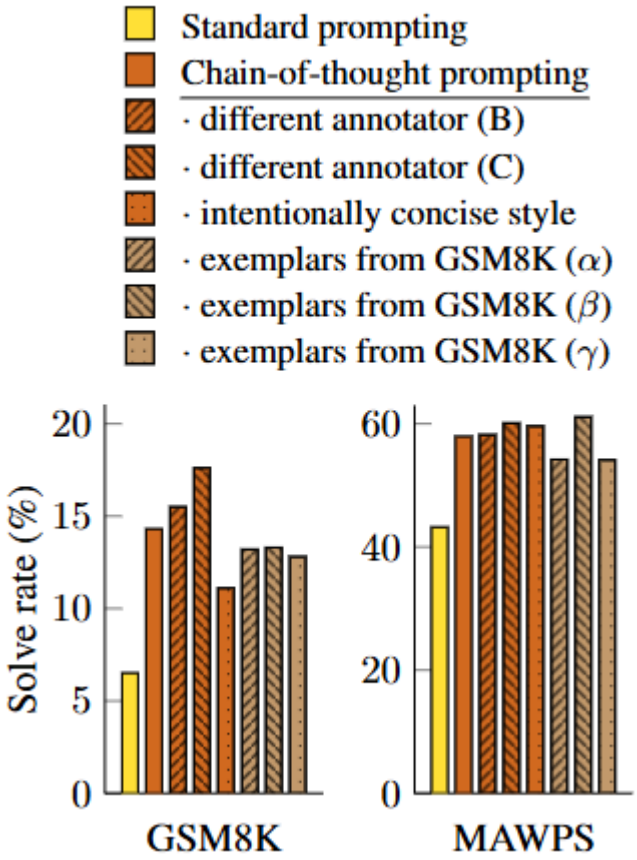
CoT 예제 순서 및 개수 변화

실험 방식

- CoT 예제의 순서를 무작위로 섞거나, 예제 개수를 변경하여 모델 성능 비교
- 예제 개수를 줄이면 성능이 감소하는지 확인

결과

- 예제 순서 변화에 대해 성능 변동이 거의 없음
- 예제 개수가 줄어들면 성능 저하 발생, 하지만 여전히 표준 Few-shot prompting보다 우수



---

4. Commonsense Reasoning

실험 개요

- 목표: Chain-of-Thought Prompting(CoT)이 상식적 추론(Commonsense Reasoning) 문제 해결 능력을 향상시키는지 평가
- 특징: 상식 추론은 명확한 정답이 없거나 여러 단계의 논리적 추론이 필요한 경우가 많음
- CoT가 복잡한 개념적 관계를 더 잘 처리할 수 있는지 확인

벤치마크

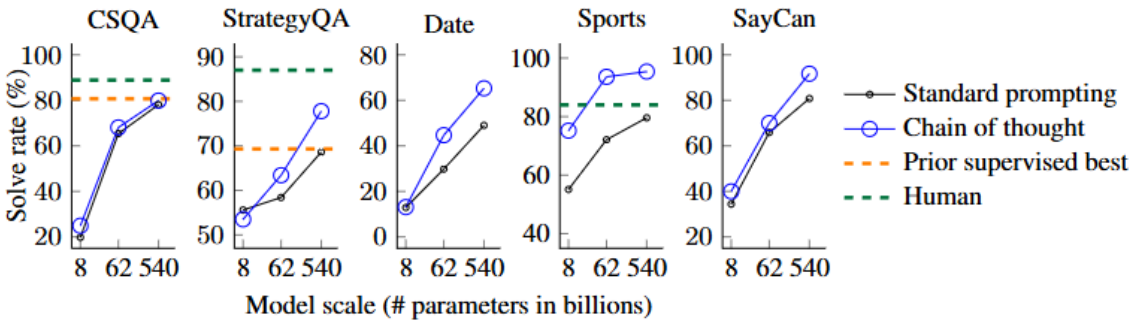
다양한 상식 추론 유형을 포함한 5가지 벤치마크에서 실험 진행

1. CommonsenseQA (CSQA)
    - 일반 상식 문제를 포함한 QA 데이터셋
2. StrategyQA
    - 단순한 상식이 아니라 논리적 추론이 필요한 문제
3. BIG-bench - Date Understanding
    - 날짜와 시간 개념을 이해하고 처리하는 문제
4. BIG-bench - Sports Understanding
    - 스포츠 관련 문장의 타당성을 평가하는 문제
5. SayCan
    - 로봇이 자연어 명령을 이해하고 실행하는 능력을 평가

실험 방식

- Few-shot prompting 방식 적용
    - 각 벤치마크에 대해 Chain-of-Thought 포함한 예제 제공
    - BIG-bench 데이터는 평가 데이터에서 10개의 예제만 사용
    - SayCan 데이터는 훈련 세트에서 6개 예제 선택

실험 결과



1. Chain-of-Thought Prompting은 모든 상식 추론 문제에서 성능 향상

- CSQA: 기존 Few-shot prompting 대비 성능 향상은 적지만, 일부 문제에서 개선
- StrategyQA: 기존 SOTA(69.4%)를 초과하여 75.6% 기록
- Sports Understanding: 인간 수행 능력(84%)보다 높은 95.4% 달성

2. 모델 크기가 클수록 CoT Prompting의 효과 증가

- PaLM 540B 모델에서 가장 큰 성능 향상 확인
- 기존 Few-shot prompting 방식보다 정확도가 높아짐
- 특히 다단계 추론이 필요한 문제에서 성능 차이가 극명

3. 특정 데이터셋에서의 성능 차이

- CSQA: Chain-of-Thought Prompting이 기존 방법보다 성능 향상을 크게 이끌어내지 못함
- StrategyQA, Sports Understanding: CoT 적용 시 기존 방법보다 큰 성능 향상

---

5. Symbolic Reasoning

실험 개요

- 목표: CoT이 기호적 추론(Symbolic Reasoning) 문제 해결에 미치는 영향을 평가
- 특징:
  - 기호적 추론은 단순한 계산이 아니라 논리적 조작 및 패턴 인식을 필요로 함
  - CoT가 단순한 언어 모델의 통계적 학습을 넘어 규칙을 일반화할 수 있는지 확인

벤치마크

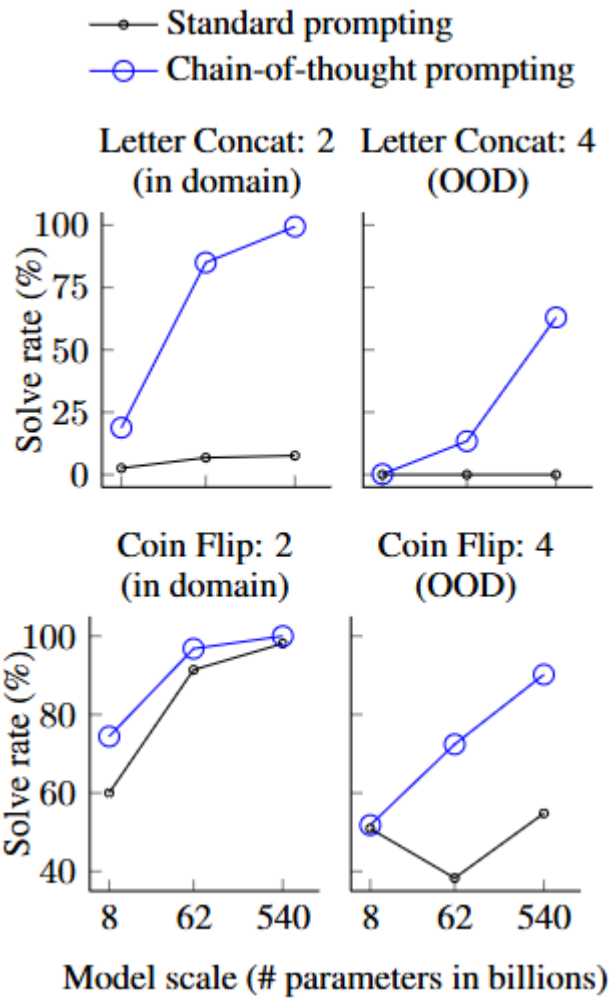
기호적 추론을 포함한 2가지 주요 작업에서 실험 진행

1. Last Letter Concatenation
    - 주어진 단어들의 마지막 글자를 연결하는 문제
2. Coin Flip (State Tracking)
    - 여러 명이 동전의 상태를 변경할 때, 최종 상태를 추론하는 문제

실험 방식

- Few-shot prompting 방식 적용
    - 8개의 Chain-of-Thought 예제 포함
    - 각 기호적 문제에 대해 별도의 CoT 예제 작성
- 평가 방법:
    - 정답률(Accuracy)

실험 결과



1. Chain-of-Thought Prompting은 기호적 추론 문제에서도 성능 향상

- PaLM 540B 모델에서 거의 100%의 solve rate 기록
- Last Letter Concatenation 및 Coin Flip 문제에서 일반적인 Few-shot prompting 보다 높은 성능 확인
- 특히, OOD(새로운 길이의 입력)에서도 일정 부분 일반화 능력 확인

2. Out-of-Domain (OOD) 일반화 성능

- 기존 Few-shot prompting은 OOD 문제에서 성능 저하
- CoT Prompting을 사용하면 OOD 문제에서도 일정 부분 성능 유지

⇒ CoT가 패턴 학습을 넘어 규칙을 일반화하는 데 기여 가능

3. 모델 크기에 따른 차이

- 100B 미만 모델에서는 CoT Prompting을 적용해도 성능 향상이 크지 않음
- PaLM 540B 등 대형 모델에서는 OOD 문제에서도 일정 성능 유지

---

6. Discussion

CoT Prompting의 핵심

- CoT Prompting은 산술, 상식, 기호적 추론에서 기존 Few-shot prompting보다 강력한 성능 향상을 제공
- CoT가 특히 모델 크기가 클수록 더 효과적이며, 단순한 답변 예측이 아닌 논리적 사고 과정을 촉진

CoT의 Scaling 효과

- 일반 Few-shot prompting의 경우 모델 크기가 커져도 성능 향상이 제한적
- CoT를 적용하면 모델 크기가 증가할수록 추론 능력이 비약적으로 향상

⇒ CoT Prompting이 대형 모델에서 "Emergent Abilities(새로운 능력의 발현)"를 유도하는 중요한 요소

추후 연구 과제

1. CoT가 진정한 "추론(reasoning)"을 수행하는가?
    - CoT가 인간처럼 논리적으로 사고하는지, 아니면 단순한 패턴 인식인지 아직 불분명

2. Annotation 비용 문제
    - Few-shot setting에서는 CoT 예제 제작 비용이 낮지만, Fine-tuning을 위해서는 CoT 데이터를 대량으로 생성해야 하는 부담 존재
    - 해결 방안으로 자동화된 CoT 데이터 생성 또는 Zero-shot CoT 방법론 연구 필요
3. CoT의 정확성 문제
    - CoT의 중간 추론 과정이 반드시 정답을 보장하지 않으며, 논리적 오류가 포함될 가능성 존재
    - CoT의 신뢰성을 향상시키는 방법(예: Self-consistency 방법론) 연구 필요
4. 소규모 모델에서 CoT 효과 증대 가능성
    - 현재 CoT는 대형 모델에서만 효과적이며, 소규모 모델에서도 효과를 극대화하는 방법 연구 필요

---

7. Related Work

1. 추론을 위한 중간 과정 생성 연구
    - 기존 연구에서는 모델이 직접 reasoning step을 생성하도록 학습
    - 기존 연구는 Fine-tuning이 필수적이었지만, CoT Prompting은 Few-shot setting에서도 강력한 성능 발휘
2. 기호적 추론 및 수학 문제 해결 연구
    - 본 연구는 순수한 자연어 기반 prompting만으로 높은 성능을 달성
3. Few-shot Learning 및 Prompting 연구
    - 본 연구는 기존 Few-shot prompting의 한계를 극복하고 논리적 추론이 필요한 문제에서도 높은 성능을 발휘하는 CoT 방법을 제안

---

8. Conclusions

연구의 주요 기여

- CoT이 대형 언어 모델에서 강력한 논리적 추론 능력을 이끌어낼 수 있음을 입증
- 산술, 상식, 기호적 추론을 포함한 다양한 문제에서 CoT Prompting이 기존 Few-shot prompting보다 월등한 성능을 보임
- 특히, CoT는 모델 크기가 증가할수록 더욱 강력한 성능 향상을 보이며, Emergent Abilities를 촉진하는 핵심 요소임

향후 연구 방향

- CoT의 신뢰성 향상: 논리적 오류를 줄이는 방법 연구 필요
- Zero-shot 및 소규모 모델에서도 CoT 효과 극대화
- CoT 데이터 자동 생성 및 annotation 비용 절감 기술 연구
