거대 언어 모델을 활용한 한국어 제로샷 관계 추출 비교 연구

김진성[1,○,†], 김경민[2,†], 박기남[2,*], 임희석[1,2,*]
고려대학교 컴퓨터학과[1], Human-inspired AI 연구소[2]
jin62304@korea.ac.kr, totoro4007@gmail.com, {spknn,limhseok}@korea.ac.kr,

A Comparative Study on Korean Zero-shot Relation Extraction using a Large Language Model

Jinsung Kim[1,○,†], Gyeongmin Kim[2,†], Kinam Park[2,*], Heuiseok Lim[1,2,*]
Department of Computer Science and Engineering, Korea University[1], Human-inspired AI Research[2]

요 약

관계 추출 태스크는 주어진 텍스트로부터 두 개체 간의 적절한 관계를 추론하는 작업이며, 지식 베이스 구축 및 질의응답과 같은 응용 태스크의 기반이 된다. 최근 자연어처리 분야 전반에서 생성형 거대 언어모델의 내재 지식을 활용하여 뛰어난 성능을 성취하면서, 대표적인 정보 추출 태스크인 관계 추출에서 역시 이를 적극적으로 활용 가능한 방안에 대한 탐구가 필요하다. 특히, 실 세계의 추론 환경과의 유사성에서 기인하는 저자원 특히, 제로샷 환경에서의 관계 추출 연구의 중요성에 기반하여, 효과적인 프롬프팅 기법의 적용이 유의미함을 많은 기존 연구에서 증명해왔다. 따라서, 본 연구는 한국어 관계 추출 분야에서 거대 언어모델에 다각적인 프롬프팅 기법을 활용하여 제로샷 환경에서의 추론에 관한 비교 연구를 진행함으로써, 추후 한국어 관계 추출을 위한 최적의 거대 언어모델 프롬프팅 기법 심화 연구의 기반을 제공하고자 한다. 특히, 상식 추론 등의 도전적인 타 태스크에서 큰 성능 개선을 보인 사고의 연쇄(Chain-of-Thought) 및 자가 개선(Self-Refine)을 포함한 세 가지 프롬프팅 기법을 한국어 관계 추출에 도입하여 양적/질적으로 비교 분석을 제공한다. 실험 결과에 따르면, 사고의 연쇄 및 자가 개선 기법 보다 일반적인 태스크 지시 등이 포함된 프롬프팅이 정량적으로 가장 좋은 제로샷 성능을 보인다. 그러나, 이는 두 방법의 한계를 지적하는 것이 아닌, 한국어 관계 추출 태스크에의 최적화의 필요성을 암시한다고 해석 가능하며, 추후 이러한 방법론들을 발전시키는 여러 실험적 연구에 의해 개선될 것으로 판단된다.

주제어: 한국어 관계 추출, 거대 언어모델, 제로샷, 프롬프트, chain-of-thought, self-refine

1. 서론

관계 추출 연구는 문장, 문서 혹은 대화 등을 포함하는 비정형 데이터로부터 의미론적 관계를 추출하는 것을 목적으로 한다. 관계 추출은 구조화된 관계적 정보를 추출할 수 있으므로 정보 추출 및 지식 베이스 구축 분야에서 중요한 역할을 한다 [1, 2]. 기존의 관계 추출 태스크에서의 많은 연구들은 BERT [3], RoBERTa [4] 등의 사전학습 언어모델을 미세 조정하여 뛰어난 성능을 달성해왔다 [5, 6, 7]. 또한, 미세 조정의 문제 중 하나인 사전학습 단계에서의 학습 방법과 미세 조정에 사용되는 방법 간의 괴리 문제를 해결하는 프롬프트 기반 학습 연구들이 저자원 환경에서 좋은 성능을 성취해왔다. 이러한 연구들은 사전학습 언어모델을 빈칸 추론 태스크를 직접적으로 수행하는 예측자로서 채택함으로써, 모델의 내재적 지식을 다운스트림 태스크에서 효과적으로 이용한다 [8, 9, 10].

그러나, 이러한 연구들은 관계 추출을 위해 적정한 크기의 사전학습 언어모델만을 고려 대상으로 삼아왔으며, 자연어처리 분야 전반에서 강력한 일반화 성능을 보여주는 GPT-3 [11]를 포함한 생성형 거대 언어모델을 대상으로 한 연구는 최근들어서 활발해지기 시작한 추세이다. 영어권 연구의 경우, 대표적으로 개체명 인식, 관계 추출 등의 태스크를 포함하는 정보 추출 분야에서 거대 언어모델의 생성 능력을 활용하거나 검증하는 연구들이 이미 다루어지고 있다. 예를 들어, [12]은 관계 추출 태스크의 대표적인 벤치마크인 CoNLL [13]에서 GPT-3 등의 거대 언어모델의 저자원 학습 능력, 즉 퓨샷 예제 환경에서의 추론 능력을 검증한다.

이러한 시점에서 한국어 관계 추출 분야에도 이러한 거대한 크기의 언어모델을 활용한 연구에 대한 비교 지표가 필요하며, 거대한 파라미터 수의 언어모델을 효과적으로 추론하도록 유도하는 효과적 프롬프팅 연구에 대한 연구 역시 필수적이다. 특히, 거대 언어모델의 강점인 저자원 환경에서의 추론 능력은 자연어처리 연구 전반에 있어 중요한 과제이며, 관계 추출 연구 역시 새로운 관계의 출현 등의 이유로 인해 퓨샷 및 제로샷 환경 기반 추론 능력을 이끌어 내기 위한 탐구가 중요하다 [14, 15]. 퓨샷 환경의 경우, 프롬프트 기반의 맥락 내 학습(In-context learning)을 통해 거대 언어모델에게 추론 예시를 주어줌으로써 내재적인 지식을 끌어낼 수 있으나, 제로샷 환경에서의 추론의 경우 예제 없이 프롬프팅 기술만을 통해 맥락에 맞는 추론이 일어나야 하므로 더욱 어려움이 존재한다. 그럼에도 불구하고, 현실 세계 환경과의 유사성에 근거하여, 제로샷 환경에서의 양질의 추론 능력을 끌어내기 위한 기법 연구는 매우 유의미하다 [16].

*교신저자(Corresponding author)

따라서, 본 연구는 대표적인 문장 분류 태스크인 관계 추출 태스크에서 거대 언어모델의 생성 기술을 통한 엔드-투-엔드 관계 추출을 위한 방법론에 대한 비교 연구를 진행하되, 제로샷 환경에서의 효과성을 세 가지 방법으로 나누어 검증 및 분석한다; 한국어 관계 추출 태스크 수행을 위한 i) 일반적 프롬프팅 기법, ii) 사고의 연쇄(Chain-of-Thought, CoT) 프롬프팅 기법, iii) 자가 개선(Self-Refine) 프롬프팅 기법 간의 제로샷 환경에서의 한국어 관계 추출 결과에 대한 양적, 질적 비교 분석을 제공한다. 본 연구는 현재 영어권 거대 언어모델 연구에서 활발하게 이루어지고 있는 프롬프팅 연구들 중 효과성이 검증된 방법들에 대한 추론 결과 비교를 제공함으로써, 거대 언어모델을 위한 최적의 프롬프팅 기술의 지속적 발전을 위한 후속 연구의 지표로서 활용되는 것을 지향한다.

2. 배경 및 관련 연구

2.1 거대 언어 모델과 프롬프팅

텍스트에 대한 확률 분포를 추정하는 언어 모델은 더 큰 데이터와 파라미터 수를 기반으로, 수 억의 파라미터 수로부터 수천억의 파라미터 수를 가진 모델 [11]에 이르기까지 그 학습 규모를 지속적으로 키워나가고 있다. 막대한 파라미터 수를 기반으로 사전 학습된 거대 언어모델은 대부분의 자연어처리 하류 태스크에서 큰 상승 폭을 가지고 성능 개선을 보여왔다. 이를 기반으로 이전까지 만연하던 미세 조정 기반의 패러다임을 맥락 내 학습 기법을 통해 저자원 환경에서 강력한 성능을 보이는 프롬프팅 기반의 패러다임으로 전환시켜왔다 [17]. 즉, 프롬프트라고 일컫는 일종의 텍스트 템플릿 등을 통해 명시적인 제약 및 지시를 제공함으로써 퓨샷 및 제로샷 환경에서 거대 언어모델의 추론 능력을 이끌어내는 연구들이 활성화되며 지속적으로 큰 중요성을 가지고 있다.

2.2 사고의 연쇄 (Chain-of-Thought)

대표적인 퓨샷 환경에서의 프롬프팅 기법인 사고의 연쇄(Chain-of-Thought, 이하 CoT) [18] 방법론의 경우, 주어진 태스크에 대한 거대 언어모델의 최종 응답이 생성 되기까지의 중간 추론 과정을 명시 혹은 모델이 스스로 명시하도록 지시함으로써, 모델의 내재적 추론 능력을 이끌어내고자 한다. 일부 양질의 추론 과정 예제를 단계별 답변으로서 제공하고 이를 통해 모델로 하여금 산술 추론(Arithmetic reasoning), 상식 추론(Commonsense reasoning) 및 상징적 추론(Symbolic reasoning) 태스크를 해결하도록 유도한다. 특히, 이 간단한 원리를 통해 PaLM [19]과 같은 거대 언어모델의 벤치마크에서의 추론 성능을 대폭 향상하는 모습을 보인다.

하지만, 위 방법론 역시 예제에 접근이 불가한 제로샷 환경에서의 한계를 지니며, 이를 위해 [16]은 제로샷 환경에서의 CoT 기법을 통해 두 단계에 걸쳐 모델 추론을 진행한다. 구체적으로는, "Let's think step-by-step." 지시를 통해 모델로 하여금 응답의 추론 과정에 대한 설명을 생성하고, 이후 이를 다시 모델에게 명시함으로써 최종 추론 결과를 이끌어 낸다. 본 연구 역시 이러한 제로샷 환경에서의 CoT 기법의 활용이 한국어 관계 추출 태스크에서 가지는 효과성에 대해 비교하고자 한다.

표 1. Vanilla 프롬프트 템플릿 예시

Task Instruction

```
You should predict the relation in the given input
between subject and object.
Choose the most appropriate relation among the given
"relations."
Do not generate a relation that does not exist among
given candidates.
Make inferences only within the given input without
external knowledge.
```

Relations (29 relations)

```
"org:founded": the date when the specified
organization was founded,
"org:member_of": organizations to which the
specified organization belongs,
"org:product": products or merchandise produced by
the specified organization,

###
...
###

"per:date_of_birth": the date when the specified
person was born,
"per:product": products or artworks produced by the
specified person
```

Input

```
- input: 공개된 영상은 한국 경제의 심장부에 서 있는
채이헌 허재 이혜준을 조명하며 시작했다.
- subject: 허재 (person)
- object: 한국 (location)
--------
- relation:
```

2.3 자가 개선 (Self-Refine)

자가 개선(Self-Refine) 프롬프팅 기법은 자가 피드백(Self-feedback)을 통해 거대 언어모델의 추론 능력을 향상하고자 한 연구이다 [20]. 이는 이전에 모델이 생성한 응답에 대하여 스스로 피드백을 하게 하고, 이를 바탕으로 다시 개선된 응답을

표 2. CoT 프롬프트 템플릿 예시

Task Instruction

Vanilla와 동일

Relations (29 relations)

Vanilla와 동일

Input

```
- input: 공개된 영상은 한국 경제의 심장부에 서 있는
채이헌 허재 이혜준을 조명하며 시작했다.
- subject: 허재 (person)
- object: 한국 (location)
--------
Let's think step-by-step briefly, and make a
prediction at the end with "Thus, the predicted
relation is:
```

재추론 하도록 하는 과정을 반복적으로 수행하게 함으로써 정제된 최종 응답을 생성하게 한다. 이를 통해 가령, 작은 사이즈의 사전학습 언어모델 혹은 분류기 등의 별도의 양질화 모듈 및 계층이 없이도 단일 거대 언어모델이 그러한 역할 역시 수행하도록 하여 답변의 품질을 향상시킨다. ChatGPT [1] 등을 포함한 거대 언어모델에 자가 개선 기법을 적용하여 반의어 생성, 대화 응답 생성, 감정 반전 등의 다양한 하류 태스크를 해결하며, 베이스 모델 대비 상당한 성능 향상을 보인다.

3. 방법론 및 실험

3.1 방법론

일반 프롬프트 제공 표 1은 관계 추출 태스크 수행을 위한 태스크 지시 및 관계 라벨 정보를 포함한 일반 프롬프트(이하 Vanilla 프롬프트)를 나타낸다. 관계 라벨은 KLUE 벤치마크 베이스라인의 평가 방법과의 동일한 평가를 위해 "no_relation" 관계를 제외한 29개의 관계 내에서 모델로 하여금 추론을 수행하도록 하며, 관계 라벨명과 함께 각 라벨이 어떤 관계를 함의하는지에 대한 부가적인 설명을 함께 제공한다.

또한, 모델의 입력으로 문장과 함께 주어(Subject) 및 목적어(Object) 쌍이 제공되는데, 이 때 각 개체에 해당하는 개체 유형(Entity type)을 함께 제공한다. 개체 유형의 경우, 데이터셋 내에서 축약어의 형태로 부여되어 있는 것을 기반으로 다음과 같이 더 설명성 있는 형태로 치환하여 제공된다; {'PER': 'person', 'ORG': 'organization', 'DAT': 'date and time', 'LOC': 'location, 'POH': 'other proper nouns', 'NOH: 'other numerals'}

[1] https://openai.com/chatgpt

표 3. Self-Refine 프롬프트 템플릿 예시

Phase 1

Task Instruction

Vanilla와 동일

Relations (29 relations)

Vanilla와 동일

Input

```
- input: 공개된 영상은 한국 경제의 심장부에 서 있는
채이헌 허재 이혜준을 조명하며 시작했다.
- subject: 허재 (person)
- object: 한국 (location)
--------
- relation:
```

Phase 2

Refining Prompt

```
(System prompt)
You aim to refine your previous prediction if any
incorrectly inferred result exists.
--------
(User prompt)
Your previous answer for {Input example}: {Previous
answer}.
Is there any relation other than the one you already
chose that seems more appropriate?
please answer in the following format.
- Yes/No:
- Feedback:
- relation:
```

사고의 연쇄 프롬프트 제공 표 2는 CoT 방법론의 적용을 위한 프롬프트의 추가 제공을 보여주며("`Let's think step-by-step ...`" 이하 부분), 상단의 태스크 지시 등의 구성은 Vanilla 프롬프트 구성과 동일하다.

자가 개선 프롬프트 제공 표 3은 Self-Refine 방법의 프롬프트 구성을 나타낸다. 상단의 *Phase 1* 부분에서 Vanilla 프롬프팅 방법과 동일하게 언어모델의 최초 관계 추론 결과를 생성한다. 이렇게 생성한 응답을 *Phase 2*의 {Previous answer} 슬롯에, 대상 문장 및 (주어, 목적어) 쌍을 다시 한번 {Input Example}에 채워넣은 후, 새로 정의된 Refining Prompt 즉, 새로운 태스크 제시 등과 함께 제공한다. 이를 통해 모델로 하여금 자신의

표 4. KLUE 검증 데이터셋에서의 세 가지 방법론(Vanilla, CoT, Self-Refine)의 제로샷 관계 추출 성능

방법	Vanilla	CoT	Self-Refine
Micro F1	51.67	40.67	45.00
	50.33	40.00	42.33
	55.67	38.67	44.67
평균	**52.56**	39.78	44.00

기존 응답에 대한 피드백 및 이를 기반으로 한 관계 재추론을 하도록 유도한다. 해당 방법론을 제안한 본래의 연구에서는 경우에 따라 최대 4번의 반복적 출력 개선을 하도록 하였으나, 본 연구에서는 API 비용으로 인해 한번의 개선 과정을 거친 후의 출력을 최종 결과로 사용한다.

3.2 실험

실험을 위한 데이터셋은 KLUE 벤치마크 [2]의 관계 추출 말뭉치가 활용되었으며, 검증 데이터셋에서 300개의 관계 샘플을 무작위 추출하여 추론 작업이 진행되었다. 실험을 위한 대상 거대 언어모델의 경우 ChatGPT(gpt-turbo-3.5-0613)을 채택했다. 총 세 번의 무작위 시드 설정을 통해 산출된 성능 및 이들의 평균 값을 기재하였으며, 성능 평가를 위한 지표는 KLUE 벤치마크의 관계 추출 태스크 평가에 사용된 Micro F1 (%) 점수를 계산하였다.

표 4의 실험 결과에 따르면, 평균 성능 52.56%로 Vanilla 프롬프팅을 적용했을 때 타 방법론 대비 최소 8.56%p의 차이를 보이며 가장 높은 성능을 보였다. 전체적인 성능 기준으로 F1 점수가 높은 프롬프팅 순서는 차례대로 Vanilla → Self-Refine → CoT임을 관측했다.

4. 분석 및 결론

4.1 분석

표 5는 동일한 예제에 대해 세 가지 프롬프팅 방법론을 적용한 결과의 예시를 나타낸다. 해당 예제에서 CoT 방법론의 경우, 한번에 관계를 최종 추론하지 않고 여러 단계의 중간 추론 과정을 생성하는 도중에 잘못된 정보를 생성하게 된다. 즉, 주어이면서 사람(person) 개체 유형을 가진 *새뮤얼 테일러 콜리지*를 4번 추론 과정에서 기관(organization)으로 착각함으로써, 최종 추론까지 호도되는 결과를 보여준다.

[2]https://github.com/KLUE-benchmark/KLUE

표 5. 동일한 예제에 대한 각 방법론의 생성 결과

Input

- input: 윌리엄 워즈워스(William Wordsworth, 1770년 4월 7일-1850년 4월 23일)는 새뮤얼 테일러 콜리지와 함께 쓴 《서정 담시집》으로 영문학에 있어 낭만주의를 개창하는데 기여한 영국의 중요한 낭만주의 시인이다.
- subject: 새뮤얼 테일러 콜리지 (person)
- object: 영국 (location)
- relation: *per:place_of_residence*

Generated Results

* **Vanilla**: “per:place_of_residence”
* **CoT**:
1. 윌리엄 워즈워스(William Wordsworth)는 누구인가요? - 윌리엄 워즈워스는 영국의 중요한 낭만주의 시인입니다.
2. 윌리엄 워즈워스(William Wordsworth)의 서정 담시집은 어떤 역할을 했나요? - 윌리엄 워즈워스의 서정 담시집은 영문학에 있어 낭만주의를 개창하였습니다.
3. 서정 담시집을 함께 쓴 사람은 누구인가요? - 새뮤얼 테일러 콜리지(Samuel Taylor Coleridge)와 함께 썼습니다.
4. 새뮤얼 테일러 콜리지는 어디에 위치한 기관인가요? - 새뮤얼 테일러 콜리지는 영국에 위치한 기관입니다.
Thus, the predicted relation is: “org:place_of_headquarters”
* **Self-Refine**:
- Yes/No: No
- Feedback: The relation “org:place_of_headquarters” is the most appropriate in this context.
- relation: org:place_of_headquarters

4.2 결론

본 연구는 최근 자연어처리 분야의 대부분의 추론 태스크에서 강력한 성능을 보이는 거대 언어모델을 활용한 한국어 제로샷 관계 추출 연구에 대해 다루었다. 특히, 거대 언어모델의 생성 능력을 활용하여 한국어 문장으로부터 두 개체 간의 관계를 추론할 때, 세 가지의 프롬프팅 방법론 간의 비교를 통해 실험 및 분석 결과를 제공했다. 관계 추출을 위한 태스크 지시 등이 포함된 일반적인 프롬프팅 방법론과 더불어, 최근 여러 태스크에서 효과성을 증명한 사고의 연쇄(CoT), 자가 개선(Self-Refine) 프롬프팅 방법론을 적용한 추론 결과를 비교 분석했다.

실험 결과에 따르면, 이산 추론 및 대화 응답 생성 등 여러 도전적인 태스크의 벤치마크에서 뛰어난 성취를 보인 CoT 및 Self-Refine 프롬프팅 방법론이 한국어 관계 추출 태스크에 다소 직관적으로 적용되었을 때, Vanilla 프롬프팅 보다 낮은 성취도를 보였다. 이는 한국어 관계 추출에 최적화된 미래의 거대 언어모델 기반 프롬프팅 연구의 심도있는 발전의 필요성을 암시한다고 이해할 수 있다. 즉, 생성 기법을 중심으로 학습된

거대 언어모델의 강력한 일반화 능력 및 내재적 지식을 한국어 관계 추출 태스크에서 온전히 유도하기 위해서는, 프롬프팅 방법론에 대한 보다 심도 있는 미래 연구가 유의미함을 보였다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음(IITP-2022-2018-0-01405). 이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2021R1A6A1A03045425). 이 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2022R1A2C1007616).

참고문헌

[1] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis, "Overview of the tac 2010 knowledge base population track," *Third text analysis conference (TAC 2010)*, Vol. 3, No. 2, pp. 3–3, 2010.

[2] K. Swampillai and M. Stevenson, "Inter-sentential relations in information extraction corpora," *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Jun. 2019. [Online]. Available: https://aclanthology.org/N19-1423

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] Y. Hur, S. Son, M. Shim, J. Lim, and H. Lim, "K-epic: Entity-perceived context representation in korean relation extraction," *Applied Sciences*, Vol. 11, No. 23, p. 11472, 2021.

[6] B. Lee and Y. S. Choi, "Graph based network with contextualized representations of turns in dialogue," *arXiv preprint arXiv:2109.04008*, 2021.

[7] G. Kim, J. Son, J. Kim, H. Lee, and H. Lim, "Enhancing korean named entity recognition with linguistic tokenization strategies," *IEEE Access*, Vol. 9, pp. 151 814–151 823, 2021.

[8] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," *arXiv preprint arXiv:2012.15723*, 2020.

[9] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "Ptr: Prompt tuning with rules for text classification," *arXiv preprint arXiv:2105.11259*, 2021.

[10] X. Chen, N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, and H. Chen, "Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction," *arXiv preprint arXiv:2104.07650*, 2021.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[12] S. Wadhwa, S. Amir, and B. C. Wallace, "Revisiting relation extraction in the era of large language models," *arXiv preprint arXiv:2305.05003*, 2023.

[13] D. Roth and W.-t. Yih, "A linear programming formulation for global inference in natural language tasks," *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*, pp. 1–8, 2004.

[14] X. Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, and M. Sun, "Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4803–4809, 2018.

[15] C.-Y. Chen and C.-T. Li, "Zs-bert: Towards zero-shot relation extraction with attribute representation learning," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3470–3479, 2021.

[16] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, Vol. 35, pp. 22 199–22 213, 2022.

[17] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and

G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Computing Surveys*, Vol. 55, No. 9, pp. 1–35, 2023.

[18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24 824–24 837, 2022.

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

[20] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang *et al.*, “Self-refine: Iterative refinement with self-feedback,” *arXiv preprint arXiv:2303.17651*, 2023.