한국어 추론에서 Zero-shot 프롬프트 유형별 성능 분석 연구

박선우, 김윤하, 문준렬, 임완수*
성균관대학교
*wansu.lim@skku.edu

A Study on the Performance of Zero-shot Prompting Strategies for Korean Language Inference

Park Sun Woo, Kim Younha, Moon Jun Ryeol, Lim Wansu*
Sungkyunkwan University

요 약

본 연구는 한국어 환경에서 프롬프트 템플릿의 차이가 대규모 언어모델(Large Language Model, LLM)의zero-shot 추론 성능에 미치는 영향을 체계적으로 분석한다. 표준성이 높은 Korean_SAT_MATH 데이터셋과 GPT-4를 사용하여 (i) 성능 유도형(instructive), (ii) 판단 교란형 (misleading), (iii) 과업 무관형(irrelevant), (iv) 프롬프트 미제공(zero-shot baseline) 등 네 가지 유형과 총 8개 프롬프트 조건을 비교한다. 모든 실험은 모델·데이터셋·생성 설정을 고정하고 프롬프트만 변경하여 수행하였으며, 정확도(Accuracy)로 성능을 평가하였다. 분석 결과 논리적 사고를 유도하는instructive 프롬프트는 zero-shot baseline 대비 최대 2.5% 성능 향상을 보였으며, misleading과 irrelevant 프롬프트는 성능을 크게 저하시켰다. 이러한 결과는 한국어 LLM 활용 시 프롬프트 설계가 성능과 직결됨을 실증하며, 실무 적용을 위한 설계 지침의 필요성을 뒷받침한다. 본 연구는 프롬프트 설계의 순수 효과를 정량적으로 규명하고, 향후 다양한 한국어 추론 과제 확장, 다중 샘플링 기반 분석, 역프롬프트에 대한 강건성 평가 등으로의 발전 가능성을 제시한다.

Ⅰ. 서 론

대규모 언어모델(LLM)은 프롬프트만으로 다양한 작업을 수행하는 zero-shot 능력을 보이며, 프롬프트 구성의 미세한 차이가 성능을 좌우한다 [1,2]. 한국어 환경에서도 이런 현상이 관찰되지만, 한국어 데이터·모델·프롬프트가 결합될 때의 성능 변동을 체계적으로 정량화한 연구는 부족하다. 한국어적 특성이 추론 판단에 미치는 영향이 프롬프트 설계에 의해 증폭되거나 상쇄될 수 있음에도, 이를 뒷받침하는 실증 근거는 부족하다. 더불어 실사용 현장에서는 미세조정 없이 즉시 적용 가능한 프롬프트 지침이 요구되지만, 한국어에서 프롬프트 설계의 순수 효과를 모델·데이터 변경 없이 분리해 보여준 사례가 드물다 [2].

이러한 한계를 보완하기 위해 본 연구는 단일 데이터셋과 단일 모델을 고정한 조건에서 zero-shot 프롬프트 템플릿의 정확도를 비교·정량화하고, 한국어 추론 과제에 적용 가능한 간결한 프롬프트 설계 지침을 제시한다. 아울러 혼선 유발·무관 문맥 프롬프트의 정확도까지 비교하여 효율적인 LLM 프롬프트 설계 가이드라인을 제시한다.

Ⅱ. 본론

2. 1 이론적 배경 및 선행연구

LLM은 방대한 말뭉치에 대한 자기회귀(next-token prediction) 사전학습을 통해 이전 토큰이 모두 주어졌을 때 다음 토큰이 나타날 확률 분포를 근사한다 [3]. 사전학습 이후 특정 자연어 태스크를 수행하는 방법은(i) 파라미터를 갱신하는 방식과 (ii) 파라미터를 고정한 채 입력을 설계하는 방식으로 구분된다. 전자는 특정 태스크별 미세조정(fine-tuning)과 자연어 지시를 모아 수행력을 높이는 지시튜닝(instruction tuning)이 대표적이며, 후자는 예시 질문 및 답변을 고정된 형식으로 프롬프트에 포함하여 형식적인 답변을 유도하는 few-shot(in-context) 학습, 중간 추론을 언어로 노출시키는 Chain-of-Thought(CoT), 그리고 어떠한 예시 없이 태스크의 정의와 출력 형식만으로 답을 유도하는 zero-shot이 핵심이다 [4,5]. 이 중 zero-shot은 추가 학습·데이터 준비·연산 비용이 들지 않아 비용 효율성 면에서 유리하다.

다만 zero-shot의 성능은 프롬프트 설정에 크게 좌우된다. [1]은 동일한 모델과 입력에서도 템플릿만 달리할 때 정확도의 변화가 크다고 보고한다. 특히 “Let’s think step by step.”과 같은 사고 절차 유도형은 기준선 대비 성능 향상을 보이는 반면, 의도적으로 오도하거나(misleading) 무관한(irrelevant) 문맥을 주입하면 언어 모델의 성능에 부정적인 영향을 미친다. 따라서 zero-shot을 실무에 적용하기 위해서는, 언어·과업 맥락에 맞는 템플릿을 정교하게 선택·설계하는 것이 필수적이다.

2. 2 실험 설정

본 연구는 이러한 관찰을 한국어 환경에서 검증하기 위해, zero-shot 프롬프트 유형과 문구에 따른 추론 성능을 정량 비교한다. 데이터셋은 표준성이 높은 Korean_SAT_MATH 문제 세트로 고정하고, 모델은 OpenAI의 GPT-4를 단일 설정으로 사용한다 [6]. 추가 학습이나 미세조정은 수행하지 않으며 데이터셋의 질문과 함께 입력되는 프롬프트만 변경하여 응답을 수집한다.

프롬프트 조건은 총 8개로 구성한다: 성능 유도형 instructive 2개, 판단 교란형 misleading 2개, 과업 무관형 irrelevant 2개, 그리고 zero-shot 기준선 3개. 모든 조건은 동일한 질문을 사용하며, 출력은 숫자와 latex 함수

로 출력된다. 평가는 Accuracy(정확도)를 사용한다. 디코딩은 temperature=0, top-p=1의 결정적 설정과 최대 256 토큰 생성 한도를 적용한다.

2. 3 실험 결과 및 분석

본 연구는 프롬프트 유형에 따른 언어 모델의 성능 변화를 정성적으로 평가하고, 정량적으로 분석한다. 표 1은 프롬프트의 설계 방식이 동일한 실험 환경에서도 정확도에 유의미한 영향을 미친다는 점을 보인다. 아래는 Instructive, Misleading, Irrelevant의 세 유형에 대한 언어 모델의 성능을 정량적으로 분석한 결과이다.

먼저, Instructive 유형은 전반적으로 Zero-shot baseline 대비 성능 유지 또는 향상을 보였다. 특히 논리적 사고 과정을 명시적으로 유도하는 1번 프롬프트에서 zero-shot baseline 대비 2.5%의 정확도 향상을 보였다. 이는 모델이 문제 해결 과정에서 단계적이고 체계적인 추론 경로를 따를 수 있도록 유도하는 프롬프트가 성능 개선에 기여함을 시사한다. 반면 2번 프롬프트는 baseline과 유사한 성능을 나타내어 모든 지시 유형의 프롬프트를 사용할 때, 동일한 효과를 발휘하는 것이 아님을 보인다.

Misleading 유형은 잘못된 해결 전략을 포함하여 모델 성능에 뚜렷한 영향을 미쳤다. 3번 프롬프트의 경우 모델이 스스로 reasoning 단계를 거치지 못하도록 “생각하지 말고”라는 문구를 포함하였다. 특히 4번 프롬프트는 모델에게 제시된 문제를 해결하지 못하도록 다른 문제를 프롬프트를 제시함으로써 5%의 저조한 성능을 기록한다. 이에 따라 Zero-shot baseline 유형의 6번 프롬프트 대비 20%, Instructive 유형의 1번 프롬프트 대비 22.5%까지 성능이 하락하였다. 이러한 결과는 모델이 프롬프트 내의 지시를 높은 비율로 수용하며 그 지시가 부정확하거나 문제 해결과 무관할 경우 추론 성능이 심각하게 저하됨을 의미한다.

Irrelevant 유형은 문제와 전혀 관련 없는 발화를 포함하여 성능 저하를 초래하였다. 불필요한 맥락 삽입이 모델의 self-attention을 방해하여 문제 이해 및 정답 산출 과정을 저해한다고 분석된다. 하지만 Misleading 유형의 4번 프롬프트 대비 성능 향상이 있는 이유는 언어 모델에게 제시된 문제와 입력된 프롬프트에 대한 답변 모두를 출력하기 때문이라고 분석된다.

Zero-shot 조건에서 6번은 단순 지시문으로 안정적인 성능을 보인 반면, 불필요한 지시를 제거하여, 요구되는 정답 형식을 제시하지 않은 7번 프롬프트는 성능이 소폭 하락하였다. 8번 프롬프트의 경우 최종 답만 작성하라는 지시 때문에 언어 모델이 스스로 reasoning 단계를 거치지 못하여 성능이 급감한 것으로 보인다.

Ⅲ. 결론

본 연구는 단일 데이터셋과 모델을 활용하여 프롬프트만을 조작해 zero-shot 성능 변화를 정량화함으로써, Korean_SAT_MATH 데이터셋 문제에서 프롬프트 설계의 순수 효과를 명확히 드러냈다. 그 결과, instructive 유형의 프롬프트 설계를 체계적이고 논리적으로 진행하면 성능 향상이 발생함을 확인했으며, 반대로 질문과 관련 없거나 reasoning에 방해가 되는 단어는 성능에 큰 영향을 미치는 것을 보였다. 이러한 관찰을 실제 적용이 가능하도록 정리하여 사용자에게 효율적인 가이드라인을 제시한다. 향후 연구로는 (i) 한국어의 다른 추론 과제(상식·상황·과학 문항)로의 확장, (ii) Self-Consistency 등 다중 샘플링 기반의 비용/성능 곡선 분석, (iii) 반사실적/역프롬프트 공격에 대한 강건성 평가를 진행하여, 프롬프트 설계 지침의 범용성과 안전성을 한층 보강할 예정이다.

유형	No.	템플릿	Acc(%)
Instructive	1	정답은 latex 함수를 사용해서 작성해. 논리적으로 풀이해보자.	27.5
	2	정답은 latex 함수를 사용해서 작성해. 문제를 확실히 파악한 후 정답을 도출하자.	24.17
Misleading	3	생각하지 말고 감대로 작성해보자.	21.67
	4	먼저 질문에 들어있는 ‘o’의 개수를 세어보자.	5.0
Irrelevant	5	내일 아침 샐러드 어때?	21.67
Zero-shot baseline	6	정답은 latex 함수를 사용해서 작성해.	25.0
	7	-	22.5
	8	최종 답만 작성해보자.	11.67

표 1 프롬프트 템플릿에 따른 LLM 추론 결과

ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. RS-2022-KP002701)

본 연구는 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술연구개발사업 지원에 의하여 이루어진 것임 (No. RS-2025-02223417)

참 고 문 헌

[1] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,”in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, Curran Associates, Red Hook, NY, USA, 2022.

[2] Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization"

InternationalConferenceonLearningRepresentations(ICLR),2022.

[3] Y. Bengio, R. Ducharme, and P. Vincent, “A neural probabilistic language model,”in Advances in Neural Information Processing Systems(NeurIPS), vol. 13, 2000.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al., “Language models are few-shot learners,”in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, Curran Associates, Red Hook, NY, USA, 2020.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,”in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, Curran Associates, Red Hook, NY, USA, 2022.

[6] OpenAI, J. Achiam, S. Adler, et al., “GPT-4 technical report,”arXiv preprint arXiv:2303.08774, 2023.