프롬프트 제약을 통한 LLM의 할루시네이션 완화 분석 연구

김윤하, 임완수*
*성균관대학교

younhaxyz@skku.edu, *wansu.lim@skku.edu

Hallucination Mitigation in LLMs via Prompt Constraints

Kim Younha, Lim Wansu*
*Sungkyunkwan University

요 약

본 연구는 문서 기반 생성형 질의응답 환경에서 “모른다” 응답을 명시적으로 허용하는 프롬프트 설계가 대규모 언어 모델(Large Lanugage Model, LLM)의 할루시네이션에 미치는 영향을 분석한다. 이를 위해 범용적으로 활용되는 Qwen2.5-7B-Instruct, Mistral-7B-Instruct, Llama-3.1-8B-Instruct 모델을 대상으로, 자유 응답을 허용하는 plain 프롬프트와 추측을 금지하고 응답 회피를 허용하는 strict 프롬프트를 비교 실험하였다. 실험은 생성형 질의응답 데이터셋인 SQuAD validation 10,000개 샘플을 기반으로 수행되었으며, 모델 출력은 정답, 오답, 그리고 ‘모른다(unknown)’ 응답으로 구분하여 평가하였다. 실험 결과, 모든 모델에서 strict 프롬프트 적용 시 오답률이 감소하는 일관된 경향이 관찰되었으며, 특히 Qwen2.5-7B-Instruct는 오답률 감소 폭이 가장 크게 나타났다. 반면, 오답률 감소와 함께 정답률 역시 일부 감소하는 trade-off가 확인되었고, 모델에 따라 ‘모른다’ 응답으로의 전이가 상이하게 나타났다. 이러한 결과는 프롬프트 기반 제약이 할루시네이션 완화에 효과적일 수 있으나, 그 영향은 모델별 응답 성향에 따라 다르게 작용함을 시사한다. 본 연구는 추가 학습 없이 프롬프트 설계만으로도 LLM의 오류 행동을 제어할 수 있음을 정량적으로 보여주며, 향후 모델 특성에 적응적인 프롬프트 설계 및 신뢰성 평가 기준 확장에 대한 기초적 근거를 제공한다.

Ⅰ. 서 론

대규모 언어 모델(Large Language Model, LLM)은 생성형 질의응답(Question Answering, QA) 태스크에서 높은 성능을 보이고 있으나, 주어진 문맥에 존재하지 않는 내용을 그럴듯하게 생성하는 할루시네이션 문제는 여전히 중요한 한계로 남아 있다. 특히 주어진 문맥을 기반으로 답변해야하는 QA 환경에서는 모델이 확실하지 않은 경우에도 답변을 생성하려는 경향이 오답의 주요 원인이 된다[1].

최근 이러한 문제를 완화하기 위한 방법으로, LLM의 추론을 유도하는 프롬프트 설계가 주목받고 있다[2]. 그러나 해당 접근법이 실제로 모델의 오답률을 얼마나 감소시키는지에 대해서는 동일한 데이터셋과 평가 기준 하에서의 정량적 분석이 충분히 이루어지지 않았다.

본 연구에서는 생성형 QA 환경에서 프롬프트 제약 조건의 유무가 서로 다른 LLM 응답 특성에 미치는 영향을 분석한다. 일반적인 자유 응답 프롬프트와, 문맥에 근거하지 않은 추측을 금지하고 ‘모른다’ 응답을 허용하는 엄격한 프롬프트를 비교하여, 정답·오답·‘모른다’ 응답의 분포 변화를 체계적으로 평가한다. 이를 통해 프롬프트 수준의 제어만으로 할루시네이션 완화 가능성을 분석하는 것을 목표로 한다.

Ⅱ. 본론

2. 1. 이론적 배경 및 선행연구

문서 기반 질의응답을 대상으로 한 선행 연구들은 LLM의 응답 오류를 다양한 관점에서 분석해 왔으며, 그중 문맥과 일치하지 않는 내용을 생성하는 현상이 핵심적인 신뢰성 문제로 다루어져 왔다[3]. 이러한 문제를 해결하기 위해 제안된 접근법들은 적용 수준에 따라 여러 범주로 구분될 수 있다. 대표적으로 외부 지식 검색을 결합하는 retrieval-augmented generation 기반 방법, 디코딩 과정에서 불확실성을 반영하는 기법, 자기 검증이나 반복적 수정과 같은 추론 단계 보완 방식, 그리고 생성 결과를 탐지하거나 수정하는 사후 처리 방법들이 존재한다[4].

또한 최근에는 모델 구조나 추가 학습을 변경하지 않고, 프롬프트나 시스템 규칙을 통해 모델의 응답 행동을 직접 제어하는 방식이 효율적인 대안으로 주목받고 있다[5]. 이러한 접근은 계산 비용이 낮고 다양한 모델에 쉽게 적용할 수 있다는 장점으로 인해, 실용적인 할루시네이션 완화 전략으로 활용되고 있다.

그러나 기존 연구들에서는 서로 다른 데이터셋, 평가 기준, 또는 모델 설정을 사용하는 경우가 많아, 프롬프트 제약 자체의 효과를 정량적으로 비교하기에는 한계가 있다. 특히 문서 기반 생성형 QA 환경에서, 자유 응답 프롬프트와 ‘추측 금지 및 모른다 응답 허용’ 프롬프트를 동일한 데이터와 지표 하에서 직접 비교한 분석은 상대적으로 부족하다. 본 연구는 이러한 공백을 보완하기 위해, 프롬프트 수준의 제어가 LLM의 오답 및 할루시네이션 행동에 미치는 영향을 체계적으로 분석한다.

2. 2. 실험 설정

본 연구에서는 범용적으로 활용되는 LLM인 Qwen2.5-7B-Instruct, Mistral-7B-Instruct, Llama-3.1-8B-Instruct 세 가지 LLM을 대상으로 실험을 수행한다. 평가 데이터셋으로는 영어 문맥 기반 생성형 QA 데이터셋인 SQuAD validation split을 사용하며, 총 10.6K 샘플 중 10,000개를 무작위로 선택하여 실험에 활용한다.

프롬프트 설정은 주어진 문맥과 질문만 제공하는 plain 프롬프트와, 문맥 외 추측을 금지하고 정보가 부족할 경우 “모른다” 응답을 허용하는 strict 프롬프트의 두 가지로 구성한다. 각 프롬프트 조건에서 동일한 데이터와 모델 설정을 사용하여 비교 실험을 수행한다.

모델 출력은 정답, 오답, 그리고 ‘모른다(unknown)’ 응답으로 구분한다. 이때 unknown 응답은 모른다는 의미를 가진 총 16가지의 구문을 지정하여 LLM 답변에 구문 포함 여부로 판단한다. 오답 비율 변화에 따른 할루시네이션 감소율을 지표로 사용하였으며, 각 조건별 정답·오답·unknown 응답 비율를 함께 분석한다.

2. 3. 실험 결과 및 분석

본 연구는 문서 기반 생성형 QA 환경에서 프롬프트 제약 조건이 LLM의 응답 특성에 미치는 영향을 정량적으로 분석한다. 표 1은 추측을 금지하고 “모른다”는 응답을 허용하는 strict 프롬프트가 오답률에서 유의미한 영향을 미친다는 점을 보인다. 아래는 프롬프트 설계에 따른 모델의 오류 감소 양상을 비교 분석한다.

실험 결과, 모든 모델에서 plain 프롬프트 대비 strict 프롬프트를 적용했을 때 오답률이 일관되게 감소하는 경향을 보인다. 특히 Qwen2.5-7B-Instruct의 경우 오답률이 12.20%에서 4.36%로 7.84%p 감소하여 가장 큰 폭의 개선을 보였으며, 동시에 unknown 응답 비율이 16.40%로 세 모델 중 가장 높게 나타났다. 이는 strict 프롬프트가 Qwen2.5-7B-Instruct 모델로 하여금 불확실한 상황에서 추측 대신 응답을 회피하도록 강하게 유도했음을 의미한다.

반면, 오답률 감소와 함께 정답률 역시 모든 모델에서 하락하는 trade-off가 관찰되었다. Mistral-7B-Instruct는 정답률 감소 폭이 1.35%p로 가장 작았으나, 오답률 감소 역시 2.30%p로 비교적 제한적인 수준에 그쳤다. Llama-3.1-8B-Instruct의 경우 strict 프롬프트 적용 시 unknown 응답 비율은 증가하였으나, 오답률 감소 폭은 0.67%p로 상대적으로 미미하였다. 이러한 결과는 strict 프롬프트의 효과가 모델별로 상이하게 나타나며, 일부 모델에서는 오류 감소보다는 응답 회피로 효과가 전이됨을 시사한다.

종합하면, ‘모른다’는 응답을 허용하는 프롬프트 제약은 할루시네이션을 효과적으로 감소시킬 수 있으나, 그 효과의 크기와 정답률 저하 정도는 모델 특성에 따라 다르게 나타난다. 이는 프롬프트 기반 할루시네이션 완화 기법이 범용적으로 동일한 효과를 보장하지 않으며, 모델별 응답 성향을 함께 고려한 분석이 필요함을 시사한다.

표 1에서 확인한 바와 같이, 세 모델 중 Qwen 계열은 프롬프트 제약에 따른 응답 변화가 가장 뚜렷하게 나타났다. 이에 따라 표 2에서는 Qwen 계열 LLM들을 대상으로, 동일한 plain 및 strict 프롬프트 조건 하에서 동일 계열 모델 간 응답 특성을 추가로 비교 분석한다.

표 2의 결과를 보면, Qwen 계열의 모든 모델에서도 plain 프롬프트 대비 strict 프롬프트 적용 시 오답률이 감소하고 unknown 응답 비율이 증가하는 일관된 경향이 관찰된다. 특히 Qwen2-7B-Instruct 모델은 strict 프롬프트를 적용했을 때 정답률이 4.57%p 향상된 유일한 사례로 나타나며, 88.83%의 정답률로 Qwen 계열 모델 중 가장 높은 성능임을 보인다. 이는 프롬프트 제약이 응답 회피 증가 없이도 성능 개선으로 이어질 수 있음을 보여준다. 모든 Qwen 계열 모델들은 Mistral 및 Llama 계열 모델과 비교했을 때, plain에서 strict로 전환될 때 오답률 감소 폭이 상대적으로 크게 나타난다. 이 중 오답률 감소가 가장 큰 모델은 여전히 Qwen2.5-7B-Instruct로 나타났고, Qwen2.5-7B 또한 6.19%p 감소로 높은 감소율을 나타냈다. 이러한 결과는 설계한 strict 프롬프트 제약이 Qwen 계열 모델에서 보다 효과적으로 작동함을 보여준다.

Ⅲ. 결론

본 연구는 문맥 기반 생성형 QA 환경에서 동일한 데이터셋과 모델 설정을 유지한 채, 프롬프트 제약 조건만을 변화시켜 zero-shot 환경에서의 응답 특성 변화를 정량적으로 분석하였다. 문맥과 질문만 제공한 프롬프트와 ‘모른다’ 응답을 허용하는 프롬프트를 비교한 결과, 모든 모델에서 오답률 감소가 관찰되었으나, 정답률 감소와 응답 회피 증가라는 trade-off가 모델별로 상이하게 나타남을 확인하였다.

이러한 결과는 프롬프트 기반 제약이 추가 학습 없이도 할루시네이션 완화에 기여할 수 있음을 보여주며, 그 효과가 모델의 응답 성향에 따라 달라질 수 있음을 시사한다. 향후 연구로는 다양한 문서 유형으로의 확장과 ‘모른다’ 응답의 적절성 분석을 통해 프롬프트 설계 지침의 범용성을 강화할 계획이다.

모델	프롬프트	정답(%)	오답(%)	unknown응답(%)
Qwen2.5-7B-Instruct	plain	87.26	12.2	0.54
	strict	79.24	**4.36**	16.4
Mistral-7B-Instruct	plain	89.46	10.37	0.17
	strict	88.11	8.07	3.82
Llama-3.1-8B-Instruct	plain	70.51	29.19	0.3
	strict	63.93	28.52	7.55

표 1 프롬프트 유형에 따른 Qwen2.5, Mistral, Llama3.1 모델 추론 결과

모델	프롬프트	정답(%)	오답(%)	unknown응답(%)
Qwen2-7B	plain	88.16	11.69	0.15
	strict	86.37	8.89	4.74
Qwen2-7B-Instruct	plain	84.26	15.53	0.21
	strict	**88.83**	8.0	3.17
Qwen2.5-7B	plain	84.29	15.63	0.08
	strict	84.54	9.44	6.02
Qwen2.5-7B-Instruct	plain	87.26	12.2	0.54
	strict	79.24	**4.36**	16.4

표 2 프롬프트 유형에 따른 Qwen 계열 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] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., et al., “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” ACM Transactions on Information Systems, vol. 43, no. 2, pp. 1–55, 2025.

[2] Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., and Wang, L. L., “Know your limits: A survey of abstention in large language models,” Transactions of the Association for Computational Linguistics, vol. 13, pp. 529–556, 2025.

[3] Sadat, M., Zhou, Z., Lange, L., Araki, J., Gundroo, A., Wang, B., et al., “DelucionQA: Detecting hallucinations in domain-specific question answering,” Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 822–835, 2023.

[4] Tonmoy, S. M. T. I., Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., et al., “A comprehensive survey of hallucination mitigation techniques in large language models,” arXiv preprint arXiv:2401.01313, 2024.

[5] Dhuliawala, Shehzaad, et al. "Chain-of-verification reduces hallucination in large language models." Findings of the association for computational linguistics: ACL 2024. 2024.