JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 4, pp. 995-1002, Apr. 2025

Check for updates

Context-Aware Decoding과 Layer별 성능 비교를 통한 대규모 언어 모델의 환각현상 완화

류 상 연[1] · 김 균 엽[2] · 강 상 우[3*]
[1]가천대학교 인공지능학과 석사과정
[2]가천대학교 인공지능학과 박사과정
[3]가천대학교 인공지능학과 부교수

Mitigating Hallucinations in LLMs through Context-Aware Decoding and Layer Comparison

Sangyeon Yu[1] · Gyunyeop Kim[2] · Sangwoo Kang[3*]

[1]Master's Student, School of Computing, Gachon University, Seongnam 13120, Korea
[2]PH.D. Student, School of Computing, Gachon University, Seongnam 13120, Korea
[3]Associate Professor, School of Computing, Gachon University, Seongnam 13120, Korea

[요 약]

대규모 언어 모델(LLM)은 뛰어난 성능에도 불구하고 환각적 텍스트를 생성하거나 문맥을 충분히 반영하지 못한다. 이를 해결하기 위해 본 연구에서는 Layer-aware Context-aware Decoding(LACD)을 제안한다. LACD는 모든 레이어에 문맥 지식 충돌 처리 기법을 주입한 뒤, 중간 레이어와 최종 레이어의 토큰 확률을 대조·분석하여 사전 학습 지식과 새로운 문맥 정보 간 균형을 동적으로 조정한다. 구체적으로, Contrast Decoding을 적용해 중간 레이어에서 이미 반영된 문맥 신호를 보존하면서, 최종 레이어가 과도하게 사전 학습 지식에 의존하지 않도록 제어한다. HotPotQA 데이터셋 실험 결과, LACD는 기존의 단순 문맥 보강 기법 대비 최대 1.9% 성능 향상을 달성하였으며, DoLa 및 CAD 등 기존 기법과 유사하거나 더 나은 결과를 보여 대규모 언어 모델의 환각 현상을 효과적으로 완화함을 확인하였다.

[Abstract]

Large language models often generate nonfactual or "hallucinated" text and fail to effectively incorporate new context. To address this issue, we propose Layer-aware Context-aware Decoding (LACD), which injects a context-knowledge conflict handling mechanism into every layer, then compares token probabilities from intermediate and final layers to dynamically balance pretrained knowledge with newly provided context. Specifically, it employs Contrast Decoding to preserve the context signals captured in intermediate layers while preventing the final layer from overreliance on pretrained knowledge. Experimental results on HotPotQA show that LACD improves performance by up to 1.9% over simple context augmentation methods and performs comparably or better than existing techniques such as Decoding by Contrasting Layers and Context-Aware Decoding, effectively mitigating hallucinations.

색인어 : 자연어 처리, 대형 언어모델, 환각 완화, 추론 방식, 레이어 분석

Keyword : Natural Language Processing, Large Language Models, Hallucination, Decoding Strategy, Layer Analysis

http://dx.doi.org/10.9728/dcs.2025.26.4.995

Received 07 March 2025; **Revised** 01 April 2025
Accepted 18 April 2025

***Corresponding Author; Sangwoo Kang**

Tel: +82-31-750-8669
E-mail: swkang@gachon.ac.kr

Ⅰ. 서 론

대규모 언어 모델(LLM; Large Language Models)은 다양한 자연어 처리(NLP; Natural Language Processing) 작업에서 뛰어난 성능을 보이며, 일관적이고 유창한 텍스트를 생성할 수 있는 능력을 가지고 있다. 그러나 이러한 모델은 환각(Hallucination), 즉 사전 학습된 지식이나 주어진 문맥과 맞지 않는 비사실적인 내용을 생성하는 한계를 가지고 있다. 이러한 문제는 특히 의료, 법률과 같이 신뢰성이 중요한 분야에서 심각한 위험을 초래할 수 있다. 예를 들어, 의료 분야에서는 잘못된 진단 정보를 제공하거나, 법률 분야에서는 부정확한 법적 조언을 생성하여 실질적인 피해를 야기할 수 있다.

환각 문제를 해결하기 위한 기존 접근법은 크게 프롬프트 기반 방법과 디코딩 전략으로 나뉜다[1]. 프롬프트 기반 방법은 모델이 문맥 정보를 더 잘 반영할 수 있도록 입력 데이터를 구조화하거나 외부 정보를 추가하는 방식이다. 대표적인 예로 단계적 추론을 유도하는 Chain of Thought (CoT)[2]와 다중 경로 추론을 활용하는 Tree of Thought (ToT)[3]가 있다. 하지만 최근 연구에 따르면 이러한 방법들은 모델의 학습된 지식을 직접적으로 조정하지 않기 때문에 복잡한 추론이 요구되는 상황에서는 여전히 환각 문제가 발생한다는 한계를 가진다고 보고된다.

이러한 한계를 극복하기 위해 디코딩 전략에 대한 연구가 주목받고 있다. Context-Aware Decoding (CAD)[4]는 문맥이 질의와 함께 주어졌을 때와 주어지지 않았을 때의 출력 확률 차이를 활용하여 문맥 정보를 강조하는 방식을 제안한다. 하지만 CAD는 모델의 사전 학습된 지식을 과도하게 억제하여 생성된 텍스트의 품질이 저하되는 문제를 가진다. [5]는 문맥 정보와 사전 학습된 지식 간의 균형 문제를 해결하기 위해, 관련 문맥과 함께 무관하거나 오답을 유도할 수 있는 문맥을 동시에 활용하는 디코딩 방식을 제안한다. 이 방법은 문맥에 대한 모델의 반응을 비교함으로써 적절한 출력을 유도하려는 접근이지만, 이는 어디까지나 외부 문맥에 대한 반응 차이에 초점을 맞춘 방식일 뿐, 모델 내부에서 지식이 어떻게 표현되고 활용되는지에 대한 내부 동작 메커니즘을 직접적으로 다루지는 않는다. 또 다른 방법인 Decoding by Contrasting Layers (DoLa)[6]는 트랜스포머 레이어 간의 차이를 분석해 상위 레이어에 저장된 사실적 지식을 부각시키는 방식을 제안하지만, 외부 지식을 활용하지 않고 모델 내부 지식에만 의존하며 추론 단계에서만 작동한다는 제약이 있다. 특히, 두 방법 모두 문맥 정보와 사전 학습된 지식 간의 균형을 효과적으로 조절하지 못하는 공통된 한계를 가지고 있다.

이러한 한계점을 해결하기 위해, 본 논문에서 우리는 문맥 정보를 강조하면서 모델 내부 레이어의 정보를 효과적으로 활용하는 Layer-Aware Contextual Decoding (LACD) 방식을 제안한다. LACD는 문맥 정보와 사전 학습된 지식 간의 균형을 동적으로 조정하며, 레이어 수준에서 토큰 분포를 분석해 문맥의 중요도를 반영하도록 설계되었다. 또한 레이어 간 상호 작용을 고려함으로써, 모델 내부 지식과 문맥 정보가 조화롭게 융합되도록 한다. 이를 통해 모델은 기존 사전 학습 지식에 과도하게 의존하지 않으면서도, 새롭게 주어진 문맥 정보를 더욱 정교하게 반영할 수 있다. 결국 LACD는 입력된 문맥 정보를 명확히 반영하면서도, 모델이 더욱 사실적이고 신뢰할 수 있는 텍스트를 생성하도록 돕는다.

Ⅱ. 관련 연구

2-1 Hallucination

LLM의 발전에 따라 Hallucination 문제는 중요한 연구 주제로 부상하고 있다. Hallucination은 LLM이 사실적 근거 없이 생성한 정보로, 실제로 존재하지 않거나 주어진 문맥과 어긋나는 내용을 포함하는 현상을 의미한다. 이는 요약(Summarization), 질의응답(QA), 법률 및 의료 데이터 처리 등 높은 신뢰성이 요구되는 애플리케이션에서 모델의 적용을 심각하게 제한하는 요인[1]이다.

Hallucination의 원인은 크게 세 가지로 구분된다[7]. 첫째, 모델의 사전 지식과 입력 문맥의 충돌이다. LLM은 대량의 웹 데이터를 통해 학습된 사전 지식이 강력하기 때문에, 새로운 정보와 충돌할 경우 비현실적인 응답을 생성하기 쉽다. 둘째, 입력 문맥의 부족이다. 충분한 정보가 제공되지 않으면 모델은 학습된 패턴을 바탕으로 불완전한 데이터를 보완하려고 시도하는데, 이 과정에서 Hallucination이 발생할 수 있다. 셋째, 디코딩 전략의 한계이다. LLM은 확률적 토큰 샘플링을 통해 텍스트를 생성하므로, 특정 조건에서는 비사실적인 정보를 선택할 위험이 커진다.

이러한 문제에 대응하기 위해 다양한 연구가 시도되었으며, 대표적으로 CAD와 DoLa 등이 제안되었다. 이들 방법은 모델이 이미 보유하고 있는 사전학습 능력을 최대한 활용하여, 별도의 파인튜닝 없이도 Hallucination 문제를 완화하는 방향으로 설계되었다.

2-2 CAD

대규모 언어 모델에서 문맥(context)과 사전 학습 지식(prior knowledge) 간의 충돌로 인해 Hallucination이 발생하는 문제를 해결하기 위한 또 다른 접근으로, 최근 CAD가 제안되었다

CAD는 모델이 “문맥이 제공된 상태”와 “문맥이 없는 상태”에서 예측한 확률 분포를 대조함으로써, 문맥 정보가 모델의 사전 지식을 덮어쓰도록 하는 디코딩 전략이다. 구체적으로는 PMI(pointwise mutual information) 개념을 활용하여,

문맥이 주어졌을 때 확률 값이 급격히 상승하는 토큰에 가중치를 부여하는 방식으로 확률 분포를 재조정한다.

이러한 접근은 추가 파인튜닝 없이도 사전 학습된 모델이 새로 주어진 문맥을 더 적극적으로 반영하도록 유도하며, Summarization과 QA 등에서 나타나는 Hallucination을 완화하는 데 효과적임이 보고되었다. 특히 지식 충돌 상황—예컨대, 최신 정보가 모델 내부에 저장된 오래된 지식과 어긋날 때—에서도 CAD는 정확한 문맥 기반 응답을 산출하도록 돕는다. 실제로 CNN-DM[8], XSUM[9] 같은 요약 데이터셋과 NQ-Swap[10], MemoTrap[11] 같은 지식 충돌 태스크에서 ROUGE-L, 정확도 및 사실성 평가 지표가 크게 개선되었음을 보고하였다. 이는 CAD가 모델 내부 파라미터를 재학습하지 않고도, 단순히 디코딩 과정에서 문맥 정보의 반영 비중을 강화함으로써 환각을 줄이고 사실성을 높였다.

2-3 Enhancing Contextual Understanding in Large Language Models

"문맥이 제공된 상태"와 "문맥이 없는 상태"에서 예측한 확률 분포를 CAD와 달리 대규모 언어 모델이 문맥과 사전 지식을 균형 있게 활용하도록, 관련 문맥과 함께 무관하거나 오답을 유도할 수 있는 문맥까지 동시에 고려하는 대조 디코딩 기법을 제안한다. 구체적으로, 질의만 주어졌을 때와 관련 문맥, 무관 문맥이 각각 추가되었을 때 산출되는 확률 값을 상호 대조하여, "관련 문맥에서는 높고 무관 문맥에서는 낮은 토큰"에 가중치를 부여함으로써 잘못된 답변으로 치우치는 것을 방지한다. 이는 추가 파인튜닝 없이도 추론 단계에서만 적용 가능하다는 점에서 실용적이지만, 세 가지 logits을 계산해야 하므로 디코딩 시간이 늘어난다는 한계를 가진다. 그럼에도, 여러 QA 및 지식 충돌 태스크에서 모델이 최신·정확한 문맥 정보를 우선시하도록 유도해 환각을 완화하고 정확도를 개선하였다는 장점이 있다.

2-4 DoLa

대규모 언어 모델 내부 레이어 간의 logit 분포 차이를 활용해 Hallucination을 완화하려는 시도로, DoLa가 제안되었다. 이 기법은 기존 Contrastive Decoding이 서로 다른 두 모델 간의 확률 분포 차이를 활용하는 것과 달리, 단일 모델 내의 중간 레이어(early exit)와 최종 레이어에서 산출된 토큰 분포를 비교한다. 이때 Jensen-Shannon Divergence (JSD)[12]를 사용하여 "가장 크게 변화하는" 레이어를 동적으로 찾아내고, 그 분포 차이를 반영해 사실적인 토큰을 우선적으로 선택하도록 유도한다.

이러한 접근은 외부 지식이나 추가 파인튜닝 없이 모델에 이미 내재된 사실적 지식을 강조함으로써, 모델이 생성 과정에서 사실성과 신뢰도를 높이도록 지원한다. 실제로 LLaMA나[13] GPT-4[14] 등 대규모 언어 모델에 적용했을 때, 정확한 지식 기반 응답을 산출하고 환각을 줄이는 데 효과적임이 보고되었다. DoLa에서는 TruthfulQA[15], FACTOR [16], StrategyQA[17], GSM8K[18] 등 다양한 평가에서 모델의 사실적 추론 성능이 유의미하게 개선됨을 보였으며, 추가 연산이나 학습 없이도 모델이 지닌 내부 정보를 더욱 정교하게 활용할 수 있음을 보였다.

Ⅲ. 제안 방법

3-1 배경

대규모 언어 모델은 임베딩 레이어, N개의 트랜스포머[19] 레이어, 그리고 최종 예측을 위한 어파인 레이어로 구성된다. 입력된 토큰 시퀀스 $x_1, x_2, \ldots x_{t-1}$는 먼저 임베딩 레이어를 통해 벡터 시퀀스로 변환된다. 이 벡터들은 트랜스포머 레이어를 거치며 점진적으로 의미적이고 문맥적인 정보를 통합하게 된다.

i번째 트랜스포머 레이어의 출력은 h_t^i로 표현되며, 최종 N번째 레이어의 출력 h_t^N은 모델이 학습한 지식을 가장 많이 반영한다. 이를 기반으로 어파인 레이어는 다음 토큰 x_t에 대한 확률 분포를 다음과 같이 예측한다.

$$p(x_t|x_{<t}) = softmax(W^T h_t^N) \quad (1)$$

여기서 W는 은닉 벡터 h_t^N를 어휘 차원으로 변환하기 위한 가중치 행렬로, 보통 $W \in R^{d \times V}$형태를 갖는다. 따라서 $W^T h_t^N$ 연산을 통해 차원 R^V의 logit 벡터를 얻고, 이를 소프트맥스에 통과시켜 각 토큰에 대한 확률 분포를 산출한다.

[20]에 따르면, 트랜스포머 레이어는 하위 레이어에서 주로 기초적인 문법적 패턴을 학습하고, 상위 레이어로 갈수록 추상적이고 의미론적인 정보를 처리하며 정보를 점진적으로 통합하는 경향이 있다. 특히, 상위 레이어는 사전에 학습된 지식을 강하게 반영하여 풍부한 의미 정보를 제공하지만, 동시에 새로운 문맥 정보를 충분히 활용하지 못할 위험성도 지적된다.

3-2 중간 레이어 Logit 검출(Early Exit)

그림 1A는 LLaMA-7B 모델에서 중간 레이어의 은닉 벡터 h_t^i를 사용하여

$$p(x_t|x_{<t}) = softmax(W^T h_t^i) \quad (2)$$

형태의 다음 토큰 확률 분포를 추출하는 과정을 보여준다. 이는 사전 학습 지식이 과도하게 반영되어 새로운 문맥이 무시되는 문제를 완화하기 위해, 모델 내부 중간 지점에서 이미

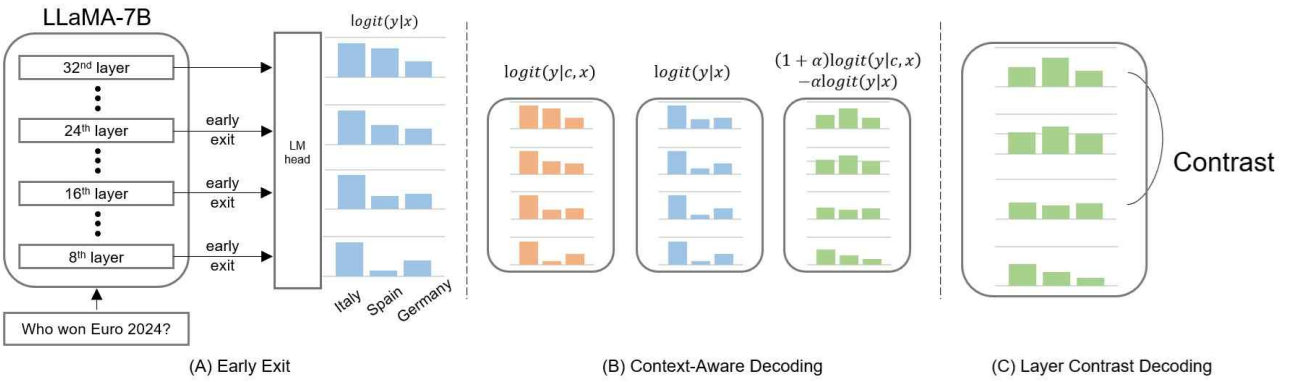

그림 1. 제안 방법 전체 구조도

Fig. 1. Overall architecture of the proposed approach

새로 주어진 질문이나 문맥 정보가 어느 정도 반영되는 시점을 포착하고, 최종 레이어로 갈수록 발생할 수 있는 Hallucination이나 편향을 제어하려는 것이다

이후, 선택된 중간 레이어(i번째)에서 얻은 확률 분포와 최종 레이어(N번째) 분포 간 Contrast Decoding을 수행함으로써, 상위 레이어에서 급격히 확률이 높아진 토큰이 실제 문맥에 부합하는지 검증한다. 이렇게 함으로써, 과도한 사전 지식이 개입된 잘못된 예측을 억제하고, 새로 주어진 정보를 더 충실히 반영한 토큰을 최종적으로 결정할 수 있다. 결과적으로, 불필요한 Hallucination을 일으키는 빈도를 낮추면서, 정확도와 문맥 적합성을 함께 향상시키는 효과를 기대할 수 있다.

3-3 문맥과 사전 학습 지식 간의 충돌

그림 1B는 문맥과 사전 학습 지식 간 Contrast Decoding 과정을 도식화한 것이다. 언어 모델은 주어진 문맥 c와 입력 쿼리 x를 기반으로 응답 y를 생성한다. 이를 수학적으로 나타내면 아래와 같다.

$$y_t \sim p_\theta(y_t|c,x,y_{<t}) \propto \exp(logit_\theta(y_t|c,x,y_{<t})) \tag{3}$$

그러나 문맥 c가 모델의 사전 학습된 데이터 분포를 벗어나거나, 사전 지식과 상충할 경우 모델은 문맥 정보를 효과적으로 반영하지 못하고 사전 학습된 지식에 지나치게 의존할 수 있다. 예를 들어, "가장 최근 유로를 우승한 나라?"라는 질문에 대해 "2024년 스페인은 유럽 축구 선수권 대회를 우승하였다"는 문맥이 제공되었음에도, 모델이 사전 학습된 오래된 지식에 의존하여 "이탈리아"라는 잘못된 응답을 생성할 수 있다.

이 문제를 해결하기 위해, 우리는 문맥 정보를 반영한 확률 분포 $p_\theta(y_t|c,x,y_{<t})$와 사전 학습 지식에 의존한 확률 분포 $p_\theta(y_t|x,y_{<t})$를 대조적으로 조합하는 방식을 제안한다. layer i번째 early exit에서 p^i_{adj}는 다음과 같이 정의할 수 있다.

$$p^i_{adj}(y_t|c,x,y_{<t}) \propto (p_\theta(y_t|c,x,y_{<t}))^{1+\alpha_{ctx}} (p_\theta(y_t|x,y_{<t}))^{-\alpha_{ctx}} \tag{4}$$

여기서 α_{ctx}는 문맥 정보의 중요성을 조절하는 하이퍼파라미터로, 문맥 c가 포함된 확률과 그렇지 않은 확률 간의 비율을 조정한다. 이 대조적 조정을 통해 사전 학습된 지식과 추가된 문맥 정보 간의 간극을 해소하고자 하였다.

3-4 레이어간 Contrast Decoding

그림 1C는 최종 레이어와 하위 레이어 간 Contrast Decoding 과정을 도식화한 것이다.

$$p_{contrast} = (1+\alpha_{layer})p^N_{adj} - \alpha_{layer}p^i_{adj} \tag{5}$$

형태로 이는 최종 레이어(N) 분포 p^N_{adj}에 중간 레이어(i) 분포 p^i_{adj}를 일정 비율 α_{layer}로 상쇄하여, 사전 학습 지식에 치우친 토큰을 억제하고, 중간 레이어에서 이미 반영된 새로운 문맥 정보를 살리기 위함이다. 마지막으로, $p_{contrast}$에 대해

$$x = \arg\max p_{contrast}, \tag{6}$$

방식으로 다음 토큰을 결정한다. 이 과정을 통해, 중간 레이어가 반영한 새로운 문맥 정보를 충분히 살리면서, 최종 레이어에서 사전 학습 지식이 지나치게 강해져 발생할 수 있는 Hallucination이나 편향을 완화할 수 있다

Ⅳ. 실 험

본 논문에서는 제안한 방법을 평가하기 위해 HotPot QA (distractor)[21] 검증 세트를 활용한다. HotPotQA는 각 질문에 대한 문맥 정보가 명확히 포함되어 있어, 모델이 문맥을

얼마나 잘 반영하는지 평가할 수 있는 적합한 환경을 제공한다. 이러한 특성 덕분에, 본 연구에서는 모델이 새로운 문맥 정보를 얼마나 잘 처리하는지, 그리고 문맥과 사전 학습 지식 간의 균형을 어떻게 조정하는지를 평가하기 위해 HotPotQA를 선택하였다. 베이스라인(baseline) 모델로는 사전에 대화·질의응답 등 광범위한 텍스트 데이터를 학습한 huggyllama/llama-7b를 사용하며, 추가 파인튜닝 없이 디코딩 과정만 변경하여 모델 성능 변화를 관찰한다. 모델의 평가 지표로는 Exact Match(EM), BLEU[22], ROUGE-L[23]을 사용하여, 모델이 생성한 답변의 정확도와 문장 품질을 종합적으로 평가한다.

4-1 프롬프트 구성 및 Few-Shot 예시

본 연구에서는 사전 학습된 모델을 추가 파인튜닝하지 않고, Few-shot 프롬프트만으로 질의응답 태스크를 수행한다. 이를 위해 사전에 준비된 질문-답변 예시 6개를 데모 텍스트 형태로 구성하였으며, 예시 질문은 “Is Mars called the Red Planet?” 등 비교적 단순한 일반 상식에 관한 것이고, 답변은 “yes.”, “Mount Everest.” 등으로 명확하게 설정하였다.

Prompt Content

```
Interpret each question literally, and as a
question about the real world;
And you can get information from Supporting
information.
...
Q: Is Mars called the Red Planet?
A: yes.
Q: What is the tallest mountain in the world?
A: Mount Everest.
...
Supporting Information : Rodri is the mvp player
of Euro 2024
Q: Who won Euro 2024?
A: Spain
```

그림 2. 프롬프트 구성 예시
Fig. 2. Example of prompt configuration

표 1에서는 이러한 프롬프트 구성과 Few-shot 예시를 구체적으로 보여준다. 특히, 문맥이 필요한 경우 “Supporting information: …” 형식으로 새로운 정보를 첨부하여 모델이 이를 활용하도록 유도할 수 있다. 반면, 문맥 미포함 유형은 준비된 Few-shot 예시와 질문만 연결해 단순 QA 형식의 프롬프트를 구성하며, 이 경우 모델은 기존 사전 학습 지식과 예시를 바탕으로 답변을 추론한다. 이렇게 두 가지 유형을 마련함으로써, 모델이 추가 문맥이 있을 때와 없을 때 각각 어떤 출력을 내는지 비교할 수 있다.

표 1. Decoding 전략에 대한 실험 결과
Table 1. Experimental results for decoding strategies

Model	Exact Match (EM)	BLEU	ROUGE-L
Baseline (w/o context)	2.2282	1.2161	4.3839
Baseline (w. Context)	38.839	17.117	53.367
CAD	39.082	18.175	56.214
DoLa	38.933	18.128	56.247
LACD	40.729	18.592	57.114

4-2 Decoding 전략에 대한 실험 결과

본 실험은 제안 방법론의 성능을 검증하기 위해 여러 가지 디코딩 전략과의 EM, BLEU, ROUGE-L 성능을 비교실험한다. 문맥 미포함 Baseline은 EM 2.22%, BLEU 1.2, ROUGE-L 4.3로 매우 낮은 성능을 보였고, 이는 사전 학습된 지식만으로는 QA 태스크에 충분하지 않음을 나타낸다. 반면, 문맥 포함 Baseline은 EM 38.84%, BLEU 17.1, ROUGE-L 53.3로 크게 상승하여, Supporting Facts가 모델의 정확도와 문장 품질 향상에 중요한 역할을 함을 확인했다. 또한 CAD와 DoLa는 문맥 포함 Baseline 대비 소폭 성능 개선을 달성해, 새 문맥 정보와 내부 지식을 조정함으로써 Hallucination을 완화할 수 있음을 보여준다. 특히, 제안한 LACD는 문맥 포함 Baseline 대비 EM 약 1.9%, BLEU 약 1.4, ROUGE-L 약 3.8점 높은 결과를 기록하며, 사전 학습 지식과 추가 문맥 간 균형을 효율적으로 제어하는 전략이 환각 억제와 답변 품질 개선에 유리함을 확인하였다.

4-3 α_{ctx}에 대한 실험 결과

본 실험은 주어진 문맥과 문맥이 없을 때의 α_{ctx}값에 따른 모델 성능 변화를 측정한 결과이다. 표 3에서 볼 수 있듯이, α_{ctx}값이 0.3일 때 EM 40.729, BLEU 18.592, 두 지표에서 최고 성능을 보였다. 이는 문맥 정보에 약 130%의 가중치를 부여하고 사전 학습된 지식에 약 70%의 가중치를 부여할 때 모델이 최적의 성능을 발휘함을 의미한다. α_{ctx}값이 0.3보다 낮을 경우 문맥 정보의 활용이 충분하지 않아 성능이 저하되었으며, 0.3보다 높을 경우에는 문맥 정보에 과도하게 의존하여 사전 학습된 지식의 활용이 제한되면서 성능이 감소하는 경향을 보였다. 이러한 결과는 문맥 정보와 사전 학습 지식 간의 균형이 모델의 정확한 응답 생성 및 환각(hallucination) 완화에 중요한 역할을 한다는 사실을 뒷받침한다.

표 2. α_{ctx}에 대한 실험 결과
Table 2. Experimental results for α_{ctx}

α_{ctx}	Exact Match (EM)	BLEU	ROUGE-L
0.1	40.068	17.833	56.801
0.2	40.473	18.209	57.002
0.3	40.729	18.592	57.114
0.4	40.486	18.581	57.144
0.5	39.082	18.14	56.2

4-4 비교 레이어에 따른 비교 실험

본 실험은 디코딩 과정에서 비교 레이어를 달리 설정했을 때, 모델 성능이 어떻게 달라지는지 확인하기 위해 진행되었다. 표 4은 레이어 12, 16, 20, 24, 28에서 디코딩을 시작했을 때 측정한 EM, BLEU, ROUGE-L 점수를 요약한 결과이다. 반면, 상위 레이어로 갈수록 성능이 현저히 떨어졌는데, 이는 너무 상위 레이어에서 디코딩을 진행하면 사전 학습 지식과 추가 문맥이 균형을 이루지 못해 Hallucination이 증가하거나 일관된 응답을 생성하기 어려워지는 것으로 해석된다. 특히 본 연구에서는 LLaMA 7B 모델을 기준으로 학습을 진행하였으며, 모델의 구조적 특성상 중간 레이어에 해당하는 16번째 레이어에서 가장 높은 성능이 나타난 것으로 판단된다. 중간 레이어는 사전 학습된 언어 지식과 문맥 정보가 최적의 균형을 이루는 지점으로, 이러한 구조적 특성이 성능 향상에 기여한 것으로 해석할 수 있다. 결과적으로, 비교 시작 레이어를 적절히 선택하는 것이 문맥 활용과 사전 학습 지식의 균형을 유지하면서 모델 성능을 최대화하는 데 중요하다는 점이 확인되었다.

표 3. 비교 레이어에 대한 실험 결과
Table 3. Experimental results for the comparing layer

Layer	Exact Match (EM)	BLEU	ROUGE-L
12	40.702	18.584	57.129
16	40.729	18.592	57.114
20	38.788	9.895	38.792
24	7.6705	5.2754	24.92
28	2.988	3.866	18.395

4-5 α_{layer}에 대한 실험 결과

본 실험은 레이어 선택 과정에서 α_{layer} 값의 변화에 따른 모델 성능을 평가한 결과이다. 표 4에 나타난 바와 같이, α_{layer} 값이 0.5일 때 EM 점수가 40.729로 가장 높게 측정되었다. α_{layer}가 0.5일 때 최종 레이어와 중간 레이어 간의 정보가 최적으로 조합되어 모델 성능을 극대화할 수 있었다. α_{layer} 값이 0.5보다 낮을 경우 최종 레이어의 영향력이 상대적으로 약화되어 성능이 다소 저하되었으며, 0.5보다 높을 경우에는 최종 레이어의 가중치가 지나치게 높아져 균형이 깨지면서 성능이 소폭 감소하는 경향을 보였다. 전반적으로 0.3에서 0.7까지의 범위에서 성능 변화가 크지 않은 것으로 보아 제안된 방법이 α_{layer} 값에 대해 비교적 안정적인 성능을 유지함을 확인 할 수 있었다.

표 4. α_{layer}에 대한 실험 결과
Table 4. Experimental results for α_{layer}

α_{layer}	Exact Match (EM)	BLEU	ROUGE-L
0.3	40.689	18.584	57.124
0.4	40.675	18.578	57.112
0.5	40.729	18.592	57.114
0.6	40.675	18.582	18.582
0.7	40.702	18,586	57,118

V. 결 론

본 논문은 문맥 정보와 레이어 간의 Contrast Decoding을 통한 새로운 디코딩 방법인 LACD을 제안한다. 이는 기존의 CAD, DoLa, 그리고 Baseline 모델과 비교하여, 문맥 정보와 사전 학습된 지식 간의 충돌을 모델 내부의 레이어 선택을 통하여 효과적으로 해결하고자 한다. 실험 결과, LACD는 EM 약 1.9%, BLEU 약 1.4, ROUGE-L 약 3.8점만큼 성능을 향상하여 모델의 정확성과 신뢰도를 크게 높일 수 있음을 확인하였다.

추후에는 모델 내에서 레이어를 선택하는 방식을 한층 더 정교화하기 위해, 동적인 방법을 활용하여 문맥 정보 집중도를 분석하는 방법을 탐색할 예정이다. 이를 통해, 문맥 정보를 보다 효과적으로 반영하면서 사전 학습된 지식이 과도하게 반영되는 문제를 완화할 수 있는 최적의 전략을 마련하고자 한다.

감사의 글

이 성과는 2025년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2022R1A2C1005316).

참고문헌

[1] S. M. T. I. Tonmoy, S. M. M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, A. Das, “A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language

Models," arXiv:2401.01313, 2024. https://doi.org/10.48550/arXiv.2401.01313

[2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, ... and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv:2201.11903, 2022. https://doi.org/10.48550/arXiv.2201.11903

[3] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," arXiv:2305.10601, 2023. https://doi.org/10.48550/arXiv.2305.10601

[4] W. Shi, X. Han, M. Lewis, Y. Tsvetkov, L. Zettlemoyer, and W. Yih, "Trusting Your Evidence: Hallucinate Less with Context-Aware Decoding," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783-791, 2024. https://doi.org/10.18653/v1/2024.naacl-short.69

[5] Z.Zhao, E. Monti, J. Lehmann, and H. Assem, "Enhancing Contextual Understanding in Large Language Models through Contrastive Decoding," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4225–4237, 2024. https://doi.org/10.18653/v1/2024.naacl-long.237

[6] Y. S. Chuang, Y. Xie, H. Luo, Y. Kim, J. Glass, and P. He, "DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models," arXiv:2309.03883, 2024. https://doi.org/10.48550/arXiv.2309.03883

[7] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, ... and T. Liu, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," arXiv:2311.05232, 2023. https://doi.org/10.48550/arXiv.2311.05232

[8] R. Nallapati, B. Zhou, C. Santos, C. Gulcehre, and B. Xiang, "Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280-290, 2016. https://doi.org/10.18653/v1/K16-1028

[9] S. Narayan, S. B. Cohen, and M. Lapata, "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797-1807, 2018. https://doi.org/10.18653/v1/D18-1206

[10] S. Longpre, K. Perisetla, A. Chen, N. Ramesh, C. DuBois, and S. Singh, "Entity-Based Knowledge Conflicts in Question Answering," arXiv:2109.05052, 2021. https://doi.org/10.48550/arXiv.2109.05052

[11] J. Liu and A. Liu. MemoTrap: A Diagnostic Benchmark for Evaluating Language Models' Susceptibility to Memorization Traps [Internet]. Available: https://github.com/liujch1998/memo-trap

[12] J. Lin, "Divergence Measures Based on the Shannon Entropy," *IEEE Transactions on Information Theory*, Vol. 37, No. 1, pp. 145-151, January 1991. https://doi.org/10.1109/18.61115

[13] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. A. Lachaux, T. Lacroix, ... and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," arXiv:2302.13971, 2023. https://doi.org/10.48550/arXiv.2302.13971

[14] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, ... and S. Jain, "GPT-4 Technical Report," arXiv:2303.08774, 2023. https://doi.org/10.48550/arXiv.2303.08774

[15] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," arXiv:2109.07958, 2021. https://doi.org/10.48550/arXiv.2109.07958

[16] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, ... and C. Leahy, "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," arXiv:2101.00027, 2020. https://doi.org/10.48550/arXiv.2101.00027

[17] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant, "Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies," *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 346-361, April 2021. https://doi.org/10.1162/tacl_a_00370

[18] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, Ł. Kaiser, ... and J. Schulman, "Training Verifiers to Solve Math Word Problems," arXiv:2110.14168, 2021. https://doi.org/10.48550/arXiv.2110.14168

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez , ... and I. Polosukhin, "Attention Is All You Need," arXiv:1706.03762, 2017. https://doi.org/10.48550/arXiv.1706.03762

[20] I. Tenney, D. Das, and E. Pavlick, "BERT Rediscovers the Classical NLP Pipeline," arXiv:1905.05950, 2019. https://doi.org/10.48550/arXiv.1905.05950

[21] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering,"

in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369-2380, 2018. https://doi.org/10.18653/v1/D18-1259

[22] K. Papineni, S. Roukos, T. Ward, T., and W.-J. Zhu, “BLEU: A Method for Automatic Evaluation of Machine Translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311-318, 2002. https://doi.org/10.3115/1073083.1073135

[23] C. Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” *Text Summarization Branches Out*, pp. 74-81, 2004.

류상연(Sangyeon Yu)

2023년 : 가천대학교 (공학사)
2024년 : 가천대학교 대학원 (공학석사)
2024년～현 재: 가천대학교 AI소프트웨어학과 석사과정
※관심분야 : 자연어처리(Natural Language Processing), 대형 언어 모델(Large Language Model), 기계독해(Machine Reading Comprehension) 등

김균엽(Gyunyeop Kim)

2021년 : 가천대학교 (공학사)
2022년 : 가천대학교 대학원 (공학석사)
2024년 : 가천대학교 대학원 (공학박사)
2021년～2023년: 가천대학교 AI소프트웨어학과 석사과정
2023년～현 재: 가천대학교 AI소프트웨어학과 박사과정
※관심분야 : 자연어처리(Natural Language Processing), 대형 언어모델(Large Language Model), 강화학습(Reinforcement Learning), 멀티모달(Multimodal) 등

강상우(Sangwoo Kang)

2012년 : 서강대학교 대학원 (공학박사-컴퓨터공학)
2012년～2016년: 서강대학교 연구교수
2016년～현 재: 가천대학교 인공지능학과 부교수
※관심분야 : 자연어처리(Natural Language Processing), 음성 대화 처리(Speech Dialogue Processing), 정보 검색(Information Retrieval), 기계독해(Machine Reading Comprehension), 기계번역(Machine Translation) 등