LLM의 '중간 분실(Lost-in-the-Middle)' 현상: 위치적 주의 편향과 해결 방안에 대한 연구 보고서

1. 서론: 장문 컨텍스트 시대의 숨겨진 도전 과제

대규모 언어 모델(LLM) 기술은 수백만 토큰에 달하는 방대한 정보를 한 번에 처리하는 '장문 컨텍스트(Long Context)' 시대로 진입하며 빠르게 발전하고 있습니다. 이러한 발전은 모델이 제공된 정보의 위치와 무관하게 모든 컨텍스트를 균일하게 활용할 수 있을 것이라는 기대를 낳았으나, 실제 모델의 작동 방식은 이 이상과 거리가 멉니다. 이 간극은 모델의 신뢰성과 직결되는 중대한 도전 과제를 제시합니다.이 발전의 이면에는 모델이 입력된 컨텍스트를 편향적으로 활용하여 발생하는 **'중간 분실(Lost-in-the-Middle)'** 이라는 심각한 문제가 존재합니다. 이 현상은 LLM이 입력 정보의 시작과 끝부분은 비교적 잘 활용하지만, 중간에 위치한 정보는 간과하거나 무시하여 성능이 급격히 저하되는 경향을 의미합니다. 이 문제는 특히 검색 증강 생성(Retrieval-Augmented Generation, RAG) 시스템의 부상으로 인해 핵심적인 관심사로 떠올랐습니다. RAG 시스템의 신뢰성은 검색된 정보의 순서와 무관하게 모델이 이를 충실히 활용하는 능력에 달려 있기 때문입니다.본 보고서는 '중간 분실' 현상을 심층적으로 분석하고, 그 근본 원인으로 지목되는 모델의 내재적 **'위치적 주의 편향(Positional Attention Bias)'** 을 탐구하는 것을 목표로 합니다. 나아가, 추론 시점에 주의(Attention) 메커니즘을 직접 보정하는 **'중간 발견(Found-in-the-Middle)'** 접근법과 특수 훈련을 통해 모델의 체질을 개선하는 **'PAM QA'** 방식 등 구체적인 해결 방안을 제시하고자 합니다.다음 장에서는 '중간 분실' 현상이 구체적으로 어떤 양상으로 나타나는지, 그리고 이것이 모델 성능에 미치는 영향을 자세히 살펴보겠습니다.

2. '중간 분실' 현상의 정의와 U자형 성능 곡선

'중간 분실' 현상은 LLM이 장문의 컨텍스트를 처리할 때, 정보의 실제 중요도와 무관하게 그 위치에 따라 정보 활용 능력이 달라지는 문제를 지칭합니다. 이는 단순히 긴 글을 처리하는 능력을 넘어, 모델의 정보 처리 메커니즘에 대한 근본적인 질문을 제기한다는 점에서 전략적으로 매우 중요합니다.이 현상은 성능 그래프에서 명확한 **U**자형 곡선**(U-shaped Performance Curve)** 으로 나타납니다. 즉, 정답을 찾는 데 필요한 핵심 정보가 입력 컨텍스트의 맨 처음 에 위치할 때(초두 효과, Primacy Bias)와 맨 끝 에 위치할 때(최신 효과, Recency Bias) 모델의 정확도가 가장 높게 나타납니다. 반면, 동일한 정보가 중간 부분에 위치할 경우 모델의 성능은 현저히 저하되는 패턴을 보입니다.다수의 문서를 참고하여 질문에 답하는 다중 문서 질의응답(Multi-document QA) 실험은 이러한 성능 저하를 명확하게 보여줍니다. 예를 들어, GPT-3.5-Turbo 모델의 경우, 정답 문서가 20개의 문서 중 중간에 위치할 때 정확도가 **20%** 이상 하락 하는 것으로 나타났습니다. 심각한 경우, 여러 참고 문서를 제공했음에도 불구하고 아무런 문서 없이 모델 자체의 지식으로만 답변하는 '폐쇄형(closed-book)' 설정보다도 성능이 낮아지는 역설적인 결과가 발생함을 명확히 합니다. 이러한 성능 저하는 검색된 문서의 순서와 무관하게 핵심 정보를 찾아 합성해야 하는 RAG 시스템의 신뢰성에 직접적이고 부정적인 결과를 초래합니다.이러한 성능 저하는 모델이 정보를 단순히 잊어버리는 것이 아니라, 특정 위치의 정보를 본능적으로 더 중요하게 여기는 내재적 편향 때문에 발생합니다. 다음 장에서는 이 문제의 근본 원인인 모델의 주의 편향에 대해 깊이 있게 분석하겠습니다.

3. 근본 원인 분석: U자형 위치적 주의 편향

앞서 설명한 U자형 성능 저하는 임의적인 실패가 아니라, 모델 아키텍처에 내재된 보다 근본적인 메커니즘, 즉 **'U**자형 위치적 주의 편향**'** 의 직접적인 증상입니다. 모델의 내부

주의 가중치를 분석한 결과, 성능 곡선이 모델이 본질적으로 컨텍스트 내에서 '어디를 보는가'를 거의 완벽하게 반영하는 거울 이미지임이 드러났습니다. 이 편향을 이해하는 것은 문제 해결의 실마리를 찾는 핵심 열쇠입니다. 이 편향은 모델이 내용의 실제 관련성이나 중요도와는 상관없이, 입력된 프롬프트의 시작과 끝부분에 위치한 토큰에 본능적으로 더 높은 주의(attention) 가중치를 할당하는 경향을 의미합니다.이러한 주장은 여러 연구를 통해 시각적, 정량적으로 입증되었습니다.

- 시각적 증거**:** 모델의 평균 주의 가중치를 시각화하면 뚜렷한 **U**자형 패턴 이 관찰됩니다. 문서들이 어떤 순서로 배열되어 있든, 심지어 무작위로 섞어도 이 U자형 패턴은 일관되게 유지됩니다. 이는 모델이 내용이 아닌 '위치' 그 자체에 편향되어 있음을 보여주는 강력한 증거입니다.
- 정량적 증거**:** 모델이 생성한 답변과 입력 문서 간의 TF-IDF(단어 빈도-역 문서 빈도) 유사도를 측정한 결과, 모델의 답변이 첫 번째 위치의 문서 내용에 강력하게 의존하는 경향이 확인되었습니다. 이는 첫 번째 문서가 정답과 무관하더라도 모델의 답변 생성에 큰 영향을 미친다는 것을 의미합니다.본질적으로, 컨텍스트의 시작과 끝에 위치한 토큰에 대한 강한 위치적 편향이 중간에 위치한 정보의 실제 관련성에서 비롯된 약한 주의 신호를 **압도(overwhelm)**하여 '중간 분실' 현상을 야기하는 것입니다.이러한 편향을 직접적으로 보정하여 문제를 해결하려는 '중간 발견(Found-in-the-Middle)' 접근법을 다음 장에서 구체적으로 소개하겠습니다.

4. 해결 방안 1: '중간 발견(Found-in-the-Middle)'을 통한 주의 보정

'중간 발견(Found-in-the-Middle)' 메커니즘은 문제의 근본 원인인 위치적 주의 편향을 직접적으로 모델링하고 분리하여 제거하는 혁신적인 해결 방안입니다. 이는 기존 모델의 구조 변경 없이 추론 단계에서 적용할 수 있는 사후 보정(post-hoc calibration) 방식입니다.'중간 발견' 메커니즘은 다음과 같은 두 단계로 작동합니다.

- 편향 모델링 및 분리**:** 핵심 아이디어는 편향을 독립적으로 측정하는 것입니다. 내용적으로 중립적인 '더미 문서(dummy document)'를 다양한 위치에 삽입함으로써, 모델이 해당 문서에 할당하는 주의 가중치는 내용 관련성이 배제된 순수한 위치적 편향을 드러냅니다. 이렇게 측정된 기준 편향 값을 실제 문서에 할당된 주의 가중치에서 빼면, 진정한 관련성을 반영하는 '보정된 주의(calibrated attention)' 점수가 남게 됩니다.
- 주의 재분배**:** 계산된 '보정된 주의' 점수를 기반으로 전체 토큰에 대한 주의 가중치를 재분배합니다. 이 과정을 통해 모델은 더 이상 문서의 위치에 현혹되지 않고, 실제 내용의 관련성에 따라 정보에 집중하게 됩니다. 즉, 중간에 있더라도 중요한 문서에 더 높은 주의를 할당하도록 유도하는 것입니다.이러한 주의 보정 메커니지은 실험을 통해 괄목할 만한 성과를 입증했습니다.
- 다양한 RAG(검색 증강 생성) 작업에서 기존 모델 대비 성능을 최대 **15%p** 까지 향상시켰습니다.
- 가장 어려운 시나리오, 즉 정답 문서가 컨텍스트 중간에 위치하는 경우에도 성능이 6~15%p 향상되었습니다.
- 이 기술은 기존의 문서 재정렬(re-ranking) 방법론과 함께 사용될 수 있으며, 재정렬 기법에 추가로 적용하여 성능을 더욱 끌어올릴 수 있는 보완적인 해결책임이 확인되었습니다.'중간 발견'이 기존 모델에 적용할 수 있는 강력한 사후 보정(post-hoc) 방법을 제공하는 반면, 또 다른 대안 전략은 편향 문제를 보다 근본적인 수준에서 해결합니다. 즉, 모델 자체를 처음부터 위치에 구애받지 않도록 재훈련하는 것입니다.

5. 해결 방안 2: 'PAM QA'를 통한 위치 불변 훈련

주의 보정과는 다른 접근법으로, 모델이 정보의 위치에 구애받지 않도록 근본적인 체질을 개선하는 특수 훈련 전략이 있습니다. **'PAM QA(Position-Agnostic Multi-step QA)'** 는 이러한 목적을 위해 설계된 훈련 방식으로, 복잡한 과제를 여러 단계로 분해하여 모델이 위치와 무관하게 핵심 정보를 식별하고 활용하는 능력을 학습시킵니다.'PAM QA' 훈련 방식은 다음과 같은 세 단계의 프로세스로 구성됩니다.

1. 질문 반복 **(Question Repetition):** 모델은 먼저 주어진 질문을 그대로 반복하여 생성합니다. 이는 모델이 후속 처리 과정을 핵심 과제에 고정시켜 불필요한 정보에 의해 주의가 분산되는 것을 방지합니다.
2. 인덱스 예측 **(Index Prediction):** 다음으로, 모델은 제공된 여러 문서 중에서 정답의 근거가 되는 문서의 **색인(index)**을 예측합니다. 이것이 위치적 편향을 깨는 결정적인 단계입니다. 모델이 관련 문서의 '색인'을 명시적으로 식별하도록 강제함으로써, 편향된 끝점을 가진 연속적인 시퀀스가 아닌, 검색 가능하고 주소 지정이 가능한 공간으로 컨텍스트를 처리하도록 학습합니다.
3. 답변 요약 **(Answer Summarization):** 마지막으로, 앞선 단계에서 파악한 질문(1단계)과 예측한 근거 문서(2단계)를 종합하여 최종 답변을 생성합니다. 이 마지막 단계는 사전 식별된 관련 정보에만 근거하여 답변을 종합하도록 훈련시켜 생성된 답변의 충실도(faithfulness)를 높입니다.이 훈련 방식을 거친 모델은 기존 모델과 뚜렷한 차이를 보입니다. PAM QA 훈련을 받은 모델의 주의 가중치는 컨텍스트 전반에 걸쳐 훨씬 더 균등하게 분포하며, U자형 편향이 거의 사라집니다. 그 결과, 관련 문서가 컨텍스트의 어느 위치에 있든 합성 과제**(Synthetic Task)**에서 거의 완벽에 가까운(99%) 안정적인 성능을 유지합니다. 이는 다른 모델들이 정답의 위치에 따라 성능이 급격히 저하되는 것과 매우 대조적인 결과입니다.

6. 결론 및 시사점

본 보고서는 장문 컨텍스트를 처리하는 LLM의 신뢰성을 저해하는 '중간 분실' 문제가 피상적인 현상이 아님을 밝혔습니다. 이 문제는 모델 아키텍처에 내재된 **'U**자형 위치적 주의 편향**'** 이라는 근본적인 원인에서 비롯되며, 이로 인해 모델은 입력 정보의 시작과 끝에 과도하게 의존하고 중간 내용은 소홀히 다루게 됩니다.이러한 문제를 해결하기 위해 본 보고서에서는 두 가지 주요 해결 방안을 제시했습니다.

- 주의 보정 **(Found-in-the-Middle):** 추론 시점에 모델의 주의 메커니즘에 직접 개입하여 위치적 편향을 제거하는 **'**사후 보정**(post-hoc calibration)'** 방식입니다. 기존 모델을 수정 없이 즉시 적용할 수 있는 유연한 접근법입니다.
- 특수 훈련 **(PAM QA):** 모델이 정보의 위치에 구애받지 않도록 분해적 과제를 통해 학습시키는 **'**사전 훈련**(specialized training)'** 방식입니다. 문제의 근본 원인을 훈련 단계에서부터 해결하여 모델의 내재적 능력을 강화합니다.이 연구 결과는 LLM 개발 및 평가에 중요한 시사점을 제공합니다. 앞으로 LLM의 성능을 평가할 때는 단순히 정답을 맞혔는지에 대한 정확도를 넘어, 정보의 위치와 같은 변수에 대해서도 일관된 성능을 유지하는지, 즉 **견고성(robustness)**을 측정하는 새로운 평가 프로토콜이 필요합니다.결론적으로, '중간 분실'과 같은 내재적 편향을 이해하고 해결하는 것은 단순히 성능 수치를 높이는 것을 넘어, LLM이 방대한 정보를 신뢰성 있게 처리하고 그 잠재력을 온전히 실현하기 위한 필수적인 과제입니다.