대규모 언어 모델의 숨겨진 약점: '중간 분실(Lost-in-the-Middle)' 현상이란?

1. 서론: LLM은 긴 글의 중간을 잊어버린다?

긴 책을 읽는 학생이 있다고 상상해 봅시다. 만약 이 학생이 책의 첫 부분과 마지막 부분은 생생하게 기억하지만, 정작 중요한 내용이 담긴 중간 부분은 흐릿하게 잊어버린다면 어떨까요? 시험 성적이 좋지 않을 것은 불 보듯 뻔합니다. 놀랍게도, 최첨단 대규모 언어 모델(LLM) 역시 비슷한 약점을 보입니다.이 현상을 바로 **'중간 분실(Lost-in-the-Middle)'**이라고 부릅니다. 이는 **"LLM이 긴 문맥(프롬프트)을 처리할 때, 중간에 위치한 정보를 제대로 활용하지 못하는 경향"**을 명확하게 정의하는 용어입니다.많은 사람들은 LLM의 '컨텍스트 창(Context Window, 한 번에 처리할 수 있는 글의 길이)'이 길어지면 당연히 성능도 좋아질 것이라고 생각합니다. 하지만 'Context Rot' 연구에 따르면, LLM은 입력 길이가 길어질수록 성능이 불안정해지는 더 넓은 의미의 문제를 겪습니다. '중간 분실' 현상은 바로 이 'Context Rot'의 가장 대표적인 증상 중 하나로, 모델이 입력된 문맥을 균일하게 처리하지 못한다는 사실을 명확히 보여줍니다. 즉, 컨텍스트 창이 아무리 길어도 그 안의 모든 정보를 동일한 중요도로 파악하는 것은 아니라는 의미입니다.그렇다면 LLM의 이러한 '건망증'은 실제로 어떻게 나타날까요? 연구 결과는 흥미로운 'U자형' 패턴을 보여줍니다.

2. 'U자형' 성능 곡선: 잊혀지는 중간 부분

'중간 분실' 현상은 LLM의 성능 그래프에서 뚜렷한 **'U자형 성능 곡선'**으로 나타납니다. Liu 등의 "Lost in the Middle" 연구에 따르면, LLM의 성능은 다음과 같은 특징을 보입니다.

- 성능이 가장 높은 구간**:** 중요한 정보가 입력 문맥의 시작**(beginning)** 또는 끝**(end)** 부분에 위치할 때.
- 성능이 가장 낮은 구간**:** 중요한 정보가 문맥의 중간**(middle)** 부분에 위치할 때.이는 마치 학생이 수업 시작과 끝에 들은 내용만 기억하는 것과 같습니다. 해당 연구에서 제시된 그래프(Figure 1)를 보면, 정답이 담긴 문서의 위치를 문맥의 시작에서 중간으로 옮길수록 정확도가 급격히 떨어지다가, 다시 끝으로 갈수록 정확도가 회복되는 뚜렷한 U자 형태를 확인할 수 있습니다.이러한 패턴은 심리학에서 오래전부터 알려진 두 가지 인지 편향과 매우 유사합니다.| 용어 | 정의 || ------ | ------ || 우선 효과 **(Primacy Bias)** | 문맥의 '처음'에 제시된 정보를 더 잘 기억하고 활용하는 경향. || 최신 효과 **(Recency Bias)** | 문맥의 '마지막'에 제시된 정보를 더 잘 기억하고 활용하는 경향. |

모델이 마치 사람처럼 처음과 끝만 중요하게 여기는 이유는 무엇일까요? 그 비밀은 LLM의 핵심 작동 방식인 '어텐션'에 숨어 있습니다.

3. 왜 이런 현상이 발생할까?: 'U자형 어텐션 편향'의 비밀

LLM의 핵심 기술 중 하나는 **'어텐션(Attention)'** 메커니즘 입니다. 이는 "LLM이 문맥 속에서 어떤 단어에 더 집중할지 결정하는 능력"으로 비유할 수 있습니다. 중요한 단어에 더 높은 '어텐션 점수'를 부여하여 의미를 파악하는 방식입니다.Hsieh 등의 "Found in the Middle" 연구는 '중간 분실' 현상의 근본 원인이 바로 이 어텐션 메커니즘의 내재적 편향, 즉 **'U자형 어텐션 편향(U-shaped attention bias)'** 때문임을 밝혀냈습니다. 이는 LLM이 정보의 실제 중요성이나 관련성과는 상관없이, 단지 위치 때문에 문맥의 시작과 끝 부분에 본능적으로 더 높은 어텐션 점수를 부여하는 경향을 의미합니다.해당 연구의 어텐션 가중치 시각화 그래프(Figure 4)를 보면, 실제 정답 문서가 어디에 있든 상관없이 모델의 어텐션은 항상

문맥의 양 끝에 높게 쏠려있는 U자 형태를 보입니다.이 편향의 핵심은 모델이 중간 부분을 아예 보지 못하는 것이 아니라는 점입니다. 오히려 문제는 위치 편향이 너무 강력해서, 중간에 있는 중요한 정보를 보고도 무시하게 만든다는 데 있습니다. 마치 양 끝에서 들려오는 정보의 '외침'이 중간에 있는 더 중요하지만 조용한 '속삭임'을 덮어버리는 것과 같습니다.**"LLM**은 중간에 있는 중요한 정보를 볼 수는 있지만**,** 시작과 끝에 있는 덜 중요한 정보에 의해 주의가 분산되어 결국 놓치게 됩니다**."**이러한 내재된 편향을 극복하기 위해 연구자들은 어떤 창의적인 해결책들을 제시하고 있을까요**?**

4. 문제 해결을 위한 노력들

'중간 분실' 문제를 해결하기 위해 연구자들은 각기 다른 철학을 가진 세 가지 주요 접근 방식을 시도하고 있습니다.

1. 가장 간단한 해결책**:** 문서 재정렬 **(Re-ranking)** 사용자 입장에서 가장 직관적이고 즉각적인 **행동적 해결책(user-side fix)**입니다. 사용자가 직접 프롬프트를 작성할 때, 가장 중요한 정보나 지시사항을 문맥의 맨 앞이나 맨 뒤로 옮기는 방식입니다. 하지만 Hsieh 등의 연구가 지적하듯, 이 방법은 모델의 근본적인 약점을 해결하는 것이 아니라 그 약점에 맞춰주는 임시방편(workaround)에 가깝습니다. 모델의 정보 활용 능력을 근본적으로 개선하지는 못합니다.
2. 편향을 직접 보정하기**:** 어텐션 캘리브레이션 **(Attention Calibration)** 모델의 작동 방식에 직접 개입하는 **실시간 알고리즘 보정(post-hoc calibration)**입니다. Hsieh 등의 "Found-in-the-middle" 연구에서 제시한 이 접근법은 "LLM의 어텐션에서 위치 편향이라는 '노이즈'를 수학적으로 측정하고 제거하여, 정보의 '진짜 관련성' 점수만 남기는 방식"으로 비유할 수 있습니다. 마치 소음 제거 헤드폰처럼 위치로 인한 불필요한 집중을 걸러내고 내용의 중요도에만 집중하도록 보정하는 것입니다. 이 방법을 통해 모델은 시작이나 끝뿐만 아니라 중간에 숨겨진 정보도 효과적으로 찾아낼 수 있게 됩니다.
3. 집중하는 법 가르치기**:** 분해적 학습 **(Decompositional Training)** 모델을 처음부터 다르게 가르치는 **근본적인 훈련 기반 접근법(foundational training approach)**입니다. He 등의 "Never Lost in the Middle" 연구는 'PAM QA'라는 훈련 방식을 제안했습니다. 이는 복잡한 문제를 풀 때 단계별로 생각하도록 가르치는 것과 같습니다. 모델에게 복잡한 질문에 바로 답하게 하는 대신, 먼저 질문을 다시 반복하고(Question Repetition), 정답이 포함된 문서의 번호를 예측하게 한 뒤(Index Prediction), 마지막으로 예측한 문서를 바탕으로 최종 답변을 요약하도록(Answer Summarization) 훈련시킵니다. 이러한 단계별 훈련 과정은 모델이 문맥의 특정 위치에 의존하지 않고, 능동적으로 필요한 정보를 찾아 집중하는 법을 배우도록 만듭니다.이처럼 활발한 연구들은 '중간 분실' 문제가 점차 개선될 수 있음을 보여주며, 현재 우리가 LLM을 더 현명하게 사용하는 방법에 대한 중요한 힌트를 줍니다.

5. 결론: LLM을 더 잘 이해하고 활용하기

'중간 분실' 현상은 LLM을 처음 접하는 학습자들에게 두 가지 중요한 실용적 교훈을 줍니다.

- 핵심은 맨 앞이나 맨 뒤에**:** 현재 LLM에게 긴 글을 요약시키거나 복잡한 자료를 바탕으로 질문할 때, 가장 중요한 정보, 지시사항, 또는 핵심 문서는 프롬프트의 시작이나 끝에 배치하는 것 이 모델의 성능을 최대한으로 끌어올리는 가장 확실한 방법입니다.
- **LLM**은 완벽하지 않다**:** 이 현상은 LLM이 아무리 발전해도 여전히 완벽한 기술이 아니며, 고유한 기술적 한계와 특성을 가지고 있음을 보여주는 좋은 예시입니다. 이러한 특성을 이해하면 LLM의 답변을 맹신하지 않고, 그럴듯해 보이는 답변 속에서

중요한 정보가 누락되지는 않았는지 비판적으로 검토하는 데 큰 도움이 됩니다.물론 이 분야의 연구는 지금도 활발히 진행 중입니다. 어텐션 편향을 보정하고, 더 효과적인 훈련 방법을 개발하려는 노력이 계속되고 있는 만큼, 미래의 LLM은 문맥 전체를 지금보다 훨씬 더 균일하고 효과적으로 활용하게 될 것입니다. 그전까지는 이 똑똑한 학생의 '건망증'을 이해하고, 중요한 내용은 따로 메모하게 하거나(문서 재정렬), 집중하는 법을 훈련시키는(분해적 학습) 것이 우리의 역할일 것입니다.