대규모 언어모델의 토큰 단위 연산 오류 분석 및 '패턴 간섭' 개념의 확립

윤종홍 송리안(멘토)

대전대신고등학교(Daejeon Daeshin High.) A.C.T.(KE)

ABSTRACT: 대규모 언어모델(LLM)은 뛰어난 자연어 처리 능력을 보임에도 불구하고, "9.11 - 9.9"와 같은 기본적인 산술 연산에서 "0.21"이라는 체계적 오류를 반복적으로 보인다. 본 연구는 이러한 현상이 단순한 계산 능력의 한계가 아니라, LLM 의 근본적인 언어 처리 메커니즘에서 비롯된다고 가정한다. LLM 은 수식을 수학적 구조가 아닌 언어적 패턴으로 처리하며, 이 과정에서 언어 규칙과 수학 규칙이 충돌하는 '패턴 간섭(Pattern Interference)' 현상이 발생하는 것을 확인했다. 본고는 '패턴 간섭'이라는 개념을 제안하여 LLM 의 연산 오류 메커니즘을 규명하고, LLM 이 '패턴 모방'을 넘어 진정한 의미 이해로 나아가기 위한 (1) 모듈화 설계, (2) 프롬프트 최적화, (3) 규칙 기반 파인튜닝 등의 개선 방향을 제시한다.

I. 서론

1.1. 연구 배경

대규모 언어모델(LLM)은 자연어 이해와 생성에서 놀라운 성능을 보이며, 번역, 요약, 코드 작성 등 다양한 영역에서 활용된다. 그러나 이러한 범용성에도 불구하고, LLM 은 기본적인 산술 연산에서 예상치 못한 오류를 보인다는 보고가 지속적으로 제기되어 왔다.

특히 흥미로운 점은, 같은 계산 문제라도 표현 방식에 따라 정확도가 달라진다는 현상이다. 예를 들어 "9.11 - 9.9"는 비교적 높은 정확도로 계산하지만, "9.11 빼기 9.9"로 표현하면 오답률이 증가한다. 또한 일부 모델은 "9.11 - 9.9 = 0.21"이라는 동일한 패턴의 오답을 반복적으로 출력한다. 이러한 현상은 단순히 "모델이 수학을 못한다"는 피상적 설명을 넘어, LLM 의 내부 처리 메커니즘에 대한 근본적 질문을 제기한다.

1.2. 연구의 필요성

본 연구는 다음 질문들에서 시작한다:

1. LLM 은 수식을 어떻게 인식하고 처리하는가?

2. 기호 표현과 자연어 표현에서 오류율이 달라지는 이유는 무엇인가?

3. "9.11 - 9.9 = 0.21"과 같은 체계적 오류는 어떤 메커니즘에서 발생하는가?

4. 이러한 오류를 설명할 수 있는 이론적 틀은 무엇인가?

본 연구는 LLM 의 수학적 오류를 언어 처리 메커니즘의 관점에서 분석함으로써, 모델의 한계를 구조적으로 이해하고 개선 방향을 제시한다. 특히 '패턴 간섭'이라는 새로운 개념을 제안하여, LLM 이 서로 다른 지식 체계를 어떻게 통합하고 충돌하는지를 설명하는 이론적 기초를 마련한다.

1.3. 기존 연구와의 차별성

기존 연구들은 주로 대규모 언어모델(LLM)의 산술 능력을 정량적으로 평가하는 데 집중해왔다. 예를 들어 Zhou et al.(2023), "Do Large Language Models Understand Arithmetic?" 은 여러 모델의 연산 정확도를 비교하며 오류 현상을 보고했지만, 표현 방식(예: "9.11 - 9.9" vs "9.11 빼기 9.9")에 따라 결과가 달라지는 이유를 설명하지 못했다.

또한 Lin et al.(2025, ICML), "Critical Tokens Matter" 는 숫자 토큰화 방식이 계산 정확도에 미치는 영향을 분석했으나, 이를 언어적 간섭 구조로 해석하지는 않았다. Patel et al.(2024, ICLR), "Language Models Are Not Abstract Reasoners" 역시 LLM 이 수학적 규칙을 일반화하지 못한다고 지적했으나, 내부 처리 과정의 상호작용은 다루지 않았다.

이에 반해 본 연구는 산술 오류를 단순한 성능 한계가 아닌 언어 패턴과 수학 규칙의 경쟁적 활성화로 해석했다. LLM 이 수식을 문자적 패턴으로 처리하면서 언어적 규칙이 수학적 추론을 방해하는 현상을 발견하고, 이를 '패턴 간섭(Pattern Interference)'이라는 새로운 개념으로 정의했다. 이 개념은 기존 연구들이 기술적으로만 접근했던 산술 오류 문제를 구조적·인지적 관점에서 재해석한다는 점에서 차별성을 지닌다.

II. 본론: LLM 의 연산 오류 메커니즘 분석

2.1. LLM 의 계산 처리: 토큰화와 패턴 예측

LLM 이 수행하는 계산 과정은 인간의 수학적 사고와 본질적으로 다르다. 인간은 "9.11 - 9.9"를 하나의 수식으로 인식하고, 소수점 정렬과 자릿수 계산이라는 명확한 알고리즘을 적용한다. 반면 LLM 은 입력된 식을 수학적 구조로 인식하지 않고, 문자열의 연속된 패턴으로 처리한다.

구체적으로, "9.11 빼기 9.9"는 모델 내부에서 다음과 같이 분해된다:

[9] [.] [11] [ ] [빼기] [ ] [9] [.] [9]

이 토큰 시퀀스는 모델에게 수식이 아닌 문장의 일부처럼 인식된다. 모델은 '빼기'라는 단어가 '빼기 연산'을 의미한다는 수학적 개념을 이해하지 못하며, 단순히 해당 문자가 학습 데이터에서 어떤 맥락에 등장했는지를 참조해 다음 토큰을 확률적으로 예측한다. 즉, 계산이 아닌 패턴 완성을 수행한다. 이러한 구조적 한계로 인해, LLM 은 수학적 정확성보다 언어적 개연성을 우선시하게 되며, 이는 체계적인 오류 패턴으로 나타난다.

2.2. 표현 방식에 따른 오류율 차이 (기호 vs 자연어)

LLM 테스트 중 흥미로운 결과도 있었다. 같은 의미의 수식이라도 표현 형식에 따라 모델의 처리 방식이 달라진다. 특히 수학 기호("-")와 자연어 표현("빼기") 사이에는 오류율에서 유의미한 차이가 관찰된다.

"9.11 - 9.9"처럼 기호를 사용한 경우, 모델은 학습 과정에서 빈번히 노출된 수식 형태의 패턴을 활성화한다. 이 경우 숫자와 기호의 배열이 비교적 명확하게 유지되어, 모델이 수학적 맥락을 부분적으로 인식할 가능성이 높다.

반면 "9.11 빼기 9.9"처럼 자연어로 표현된 경우, 모델은 이를 문장 구조로 해석한다. 이때 활성화되는 것은 수학 패턴이 아닌 언어적 비교·서술 패턴이다. 예를 들어:

1. "A 는 B 보다 크다/작다“

2. "A 와 B 의 차이는...“

3. "A 에서 B 를 제거하면...“

이러한 언어 패턴이 우세하게 작용하면서, 모델은 정확한 산술 연산 대신 의미적 관계 판단을 수행하게 된다. 결과적으로 "빼기"라는 자연어 표현은 수학적 연산보다 언어적 해석을 촉발하여, 오류를 증폭시키는 요인으로 작용한다.

2.3. "9.11 - 9.9 = 0.21" 오류의 발생 메커니즘

실제로 다수의 LLM 에서 "9.11 - 9.9"를 계산할 때 "0.21"이라는 오답이 반복적으로 관찰된다. 이 오류는 단순한 실수가 아니라, 모델의 내부 처리 로직에서 비롯된 체계적 편향이다. 모델은 다음과 같은 휴리스틱 패턴을 따른다:

1. 정수부 인식: "9.11"에서 정수부 '9' 감지, "9.9"에서 정수부 '9' 감지

2. 소수부 분리 처리: "9.11"의 소수부 '11'을 독립적 단위로 인식, "9.9"의 소수부 '9'를 독립적 단위로 인식

3. 정수부 상쇄: 9 - 9 = 0 (정확)

4. 소수부 차이 계산: '11'과 '9'의 차이 → '2' 도출 (자릿수 관계 무시)

5. 결과 조합: 정수부 '0' + 소수점 + '2' + (남은 자릿수) '1' → "0.21”

이 과정에서 모델은 소수점 자릿수 정렬이라는 수학적 규칙을 적용하지 않는다. 대신 '11'과 '9'를 문자 그대로 처리하여, 마치 "11 - 9 = 2"를 수행한 후 결과를 소수점 뒤에 배치하는 방식으로 패턴을 완성한다.

수학적으로는 명백한 오류지만, 언어 패턴의 관점에서는 다음과 같은 이유로 개연성이 있다:

• 학습 데이터에서 "A.BC - A.D = 0.E" 형태의 패턴이 빈번히 등장

• '11'과 '9'의 차이를 구하는 것은 언어적으로 자연스러운 처리

• 결과 형식("0.XX")이 소수 뺄셈의 일반적 출력 형태와 일치

따라서 모델은 논리적으로 잘못된 답을 내면서도, 확률적으로는 가장 그럴듯한 패턴을 선택한 것이다.

III. '패턴 간섭' 개념의 제안

3.1. '패턴 간섭(Pattern Interference)'의 정의

이러한 현상을 설명하기 위해, 본 연구에서는 '패턴 간섭(Pattern Interference)'이라는 개념을 제안한다. 패턴 간섭이란, 서로 다른 규칙 체계(언어적 패턴 vs 수학적 규칙)가 동시에 활성화되어, 한쪽의 정확한 판단을 방해하는 현상이다.

LLM 은 본질적으로 언어 패턴 예측 모델이다. 그러나 학습 데이터에는 수학 문제와 그 풀이도 포함되어 있어, 모델은 일정 수준의 수학적 추론 능력을 습득한다. 문제는 이 두 체계가 동일한 어텐션 메커니즘과 가중치를 공유한다는 점이다.

입력이 수학 문제로 주어졌을 때, 모델 내부에서는 다음 두 신호가 경쟁적으로 활성화된다:

• 어 패턴: "이런 형태의 문장 뒤에는 이런 단어가 온다"

• 학 규칙: "이런 연산에는 이런 절차를 따라야 한다“

만약 언어 패턴의 신호가 더 강하면, 수학적 정확성은 무시되고 언어적으로 그럴듯한 답이 출력된다.

3.2. 패턴 간섭의 특징

• 표면적 자연스러움: 오답임에도 불구하고, 언어적 구조는 자연스럽다.

- 일관된 오류 패턴: 같은 유형의 문제에서 같은 방식의 오류가 반복된다.
- 맥락 의존성: 표현 방식(기호/자연어)에 따라 간섭의 강도가 달라진다.

3.3. 패턴 간섭의 확장적 함의

패턴 간섭은 단순한 계산 오류를 넘어, LLM 의 논리적 추론 한계를 설명하는 핵심 개념이다.

IV. 연구의 의의

본 연구는 다음과 같은 학술적·실용적 의의를 갖는다.

4.1. 이론적 기여

LLM 의 수학적 오류를 단순한 모델의 기술적 한계가 아닌, 언어 패턴 처리 메커니즘의 구조적 부작용으로 재해석한 최초의 시도이다. 기존 연구들이 "모델이 계산을 못한다"는 현상 기술에 그쳤다면, 본 연구는 왜 못하는가에 대한 메커니즘을 '패턴 간섭'이라는 개념적 틀로 설명했다.

특히 '패턴 간섭'이라는 용어는 단순히 오류를 명명하는 것을 넘어, 두 개의 이질적 처리 체계가 단일 신경망 구조 내에서 충돌하는 현상을 포착한다는 점에서 개념적 독창성을 갖는다. 이는 LLM 이 '범용 지능'처럼 보이지만, 실제로는 여러 도메인의 패턴을 무차별적으로 혼합한 결과물임을 시사한다.

4.2. 학술적 활용 가능성

본 연구에서 제안한 패턴 간섭 개념은 다음 영역에서 추가 연구의 기초 자료로 활용될 수 있다:

1. 인지과학적 접근: 수학적 사고와 언어적 사고가 인간의 뇌에서도 분리된 처리 경로를 갖는지, LLM 의 패턴 간섭이 인간의 인지적 오류와 유사한지 비교 연구 가능

2. 계산언어학적 접근: 자연어 처리와 형식 언어 처리의 경계에 대한 이론적 논의 확장

3. AI 안전성 연구: 패턴 간섭이 발생하는 조건을 특정함으로써, 고신뢰성이 요구되는 영역(의료, 금융, 법률)에서 LLM 의 한계를 예측하고 보완책 마련

V. 결론

본 연구는 대규모 언어모델(LLM)이 단순한 산술 연산조차 안정적으로 수행하지 못하는 원인을 '패턴 간섭(Pattern Interference)'이라는 개념으로 설명했다.

분석 결과, LLM 은 수식을 수학적 구조로 이해하지 않고 언어적 패턴의 연속으로 처리하며, 이 과정에서 언어 규칙과 수학 규칙이 서로 간섭을 일으켜 체계적인 오류를 유발한다는 사실을 확인했다.

특히 "9.11 - 9.9 = 0.21"과 같은 반복적 오답은 계산 능력 부족이 아닌 언어 예측 구조의 확률적 편향에서 비롯된 것으로, 이는 모델 내부의 패턴 학습 메커니즘이 갖는 근본적 한계를 드러낸다.

따라서 본 연구는 LLM 의 오류를 단순한 기술적 결함이 아닌 언어 처리 구조의 부산물로 규정하며, 이러한 구조적 간섭을 줄이기 위한 새로운 연구 방향을 제시했다.

구체적으로는 (1) 수학 연산과 언어 처리를 분리한 모듈화 설계, (2) 표현 방식에 따른 프롬프트 최적화, (3) 수학 데이터의 규칙 기반 파인튜닝 전략 등이 LLM 의 연산 신뢰도를 높이는 핵심적 대안으로 제안되었다.

결국, 본 탐구는 LLM 의 "오류"를 비판적으로 해부함으로써 AI 가 '패턴 모방'을 넘어 진정한 의미 이해로 나아가기 위한 전제 조건을 제시한다. '패턴 간섭'은 향후 LLM 연구뿐 아니라 AI 인지구조, 계산언어학, 인공지능 안전성 연구 전반에서 중요한 분석 틀로 확장될 잠재력을 지닌다.

참고문헌

• Lin, Z., et al. (2025). "Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM's Reasoning Capability." Proceedings of the International Conference on Machine Learning (ICML).

• Nogueira, R., et al. (2024). "Large Language Models Struggle to Add Two Numbers." arXiv preprint arXiv:2401.07920.

• Pan, Y., et al. (2023). "Do Large Language Models Understand Arithmetic? An Analysis of the T-Equal-K Problem." arXiv preprint arXiv:2305.11116.

• Patel, A., et al. (2024). "Language Models Are Not Abstract Reasoners." Proceedings of the International Conference on Learning Representations (ICLR).

• Wei, J., et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." Advances in Neural Information Processing Systems (NeurIPS).

• Yuan, Z., et al. (2024). "Analyzing and Mitigating Arithmetic Errors in Large Language Models." arXiv preprint arXiv:2401.10186.

• Zhang, Z., et al. (2023). "Evaluating Arithmetic Capabilities of Large Language Models." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL).