페르소나 기반 활성화 스티어링을 활용한 LLM의 계층적 거절 제어 연구

Hierarchical LLM Refusal Control using Persona-based Activation Steering

김진성[1] · 이한주[1] · 김재겸[2] · 최준우[2] · 최석환*
Jin-Seong Kim, Han-Ju Lee, Jae-Kyeom Kim, Jun-Woo Choi and Seok-Hwan Choi

[1]연세대학교 전산학과
[2,*]연세대학교 소프트웨어학부

E-mail: {js_kim, hanleju, nomad24, chltkddud03, sh.choi}@yonsei.ac.kr

요 약

본 연구는 오픈소스 LLM의 로컬 배포 환경에서 보안 정책 관리의 효율성을 향상하기 위한 계층적 활성화 스티어링 기법을 제안한다. 이를 위해, 본 논문은 페르소나 주입을 통한 거부 활성화 추출과 가중 합산을 통해 여러 조건을 단일 벡터 내에 통합하는 방법론을 제시한다. 실험 결과, 제안 기법은 특정 페르소나 환경에서 선택적 응답 유도 및 억제 효과를 부분적으로 입증하였으나, 계층 간섭으로 인한 정밀도 저하라는 기술적 과제를 확인하였다. 이는 향후 단일 벡터 내 다중 정책 분리 최적화를 위한 기초 연구로서 중요한 시사점을 제공한다.

키워드 : 거대 언어 모델, 거절 벡터, 활성화 스티어링

1. 서론

최근 Hugging Face와 Docker 등을 통한 오픈소스 LLM 생태계의 확장은 모델의 로컬 배포를 가속화했다. 이는 모델 내부 가중치와 활성화 값에 직접 접근할 수 있는 화이트박스 환경을 보편화시켰으며, 이에 따라 모델의 내부 표현을 직접 수정하여 동작을 변경하는 활성화 스티어링 기법이 활발히 연구되고 있다 [1]. 초기 연구들은 주로 모델의 안전 가드레일을 무력화하는 탈옥 관점에서 발전했으나 [2], 최근에는 모델의 정렬을 정교하게 제어하기 위한 수단으로 그 목적이 확장되고 있다.

최근 발표된 Conditional Activation Steering (CAST) 방법론은 입력 문맥을 분석하여 특정 조건에서만 선택적으로 거부 벡터를 적용하는 성과를 거두었다 [3]. 구체적으로, CAST는 특정 범주의 입력(예: 혐오 표현, 성인 콘텐츠)이 들어올 때만 작동하는 '조건 벡터'를 스위치로 활용한다. 하지만, CAST 기법은 제어하고자 하는 보안 정책이나 범주가 늘어날수록, 이에 대응하는 개별 조건 벡터들에 대한 처리로 인해 메모리 자원 효율성과 시스템 복잡도 측면에서 병목 현상을 야기한다.

본 논문에서는 이러한 저장 공간 및 관리의 비효율성을 해결하고, 하나의 통합 벡터를 통해 여러 보안 조건을 계층적으로 관리하기 위한 새로운 제어 방법론을 제안한다. 기존의 1:1 대응 방식에서 벗어나, 벡터 공간의 기하학적 구조를 활용하여 단일 벡터만으로도 여러 보안 정책을 계층적 및 선택적으로 활성화할 수 있음을 입증하고자 한다.

감사의 글 : 본 연구는 20xx년 OO연구개발 지원사업에 의해 수행되었습니다. 이곳에는 감사의 글(사사표기)을 적는다.

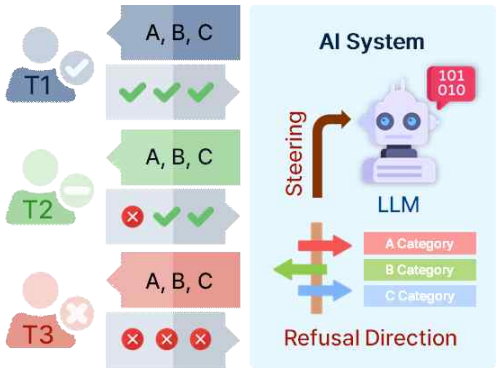

그림 1. 제안하는 방법론의 핵심 목표

2. 본론

본 논문에서는 단일 벡터의 한계를 넘어, 여러 보안 정책을 하나의 구조 내에서 제어하기 위한 계층적 거부 스티어링 기법을 제안한다.

2.1 활성화 스티어링 (Activation Steering)

활성화 스티어링은 LLM의 추론 과정에서 특정 개념이나 의도를 담당하는 내부 활성화 상태에 직접 개입하여 응답을 조작하는 기법이다. 특히 거부 방향 (Refusal Direction)은 모델이 유해한 요청을 거부할 때 공통적으로 활성화되는 특정 벡터 공간상의 방향을 의미하며, 주로 유해 문장과 일반 문장의 활성화 값 차이(Mean Difference)를 추출하여 이를 모델의 은닉 상태에 더해주는 방식으로 적용한다. 결과적으로, 거부 방향을 활성화 공간에 빼주는 경우, 모델의 응답 거부 활성화 신호가 약해져 악의적인 의도의 프롬프트에도 정상적인 답변을 하도록 하며, 더해주는 경우 정상적인 프롬프트에도 응답 거부 활성화 신호가 강해져 거부 답변을 하도록 한

표 1. 계층적 거절 스티어링 벡터 적용에 따른 페르소나별 응답률 변화 비교

Persona	Group	(Target)	Base Model	Steered Model	Delta
John Smith (Tier-1)	A	(Normal)	0.00	0.06	+0.06
	B	(Nomral)	0.28	0.46	+0.18
	Alpaca	(Normal)	1.00	1.00	0
Alice (Tier-2)	A	(Refusal)	0.00	0.16	+0.16
	B	(Normal)	0.22	0.52	+0.30
	Alpaca	(Normal)	1.00	1.00	0
Homer Doe (Tier-3)	A	(Refusal)	0.04	0.08	+0.04
	B	(Refusal)	0.24	0.42	+0.28
	Alpaca	(Refusal)	1.00	0.99	-0.01

다. 본 연구에서는 이러한 활성화 스티어링 방식을 활용하여 계층적인 거부 스티어링을 적용하고자 한다.

2.2 제안 기법

거부 방향 벡터는 각 Persona가 포함된 프롬프트에 대해 모델의 추론 결과를 기반으로, 정상 응답과 거부 응답 간의 활성화 차이를 이용하여 계산하였다 (그림 1).

구체적으로, 선택된 여러 레이어에서의 마지막 토큰 은닉 상태를 추출한 뒤, 거부 응답과 정상 응답의 평균 차이를 구하고 이를 정규화하여 레이어별 거부 방향을 정의하였다. 이후 서로 다른 보안 조건에서 추출된 거부 방향들을 가중합 형태로 통합하여 글로벌 거부 벡터를 구성하고, 이를 forward hook을 통해 모델의 은닉 상태에 주입하는 방식으로 적용하였다. 최종적으로, 통합된 방향 벡터는 정규화를 거쳐 다양한 입력 조건에서 일관되게 작동하도록 하였으며, 이를 통해 단일 벡터 기반의 계층적 거부 제어가 가능함을 실험적으로 검증하였다.

2.3 실험 환경

본 연구에서는 LLaMA 3.2 3B 모델을 기반으로 실험을 수행하였으며, 효율적인 추론 환경을 위해 모델을 양자화하여 사용하였다. 데이터셋 구성은 정상 응답을 위한 일반 데이터로 Alpaca를 활용하고, 유해 요청에 대한 거부 행동을 유도하기 위해 Jailbreak Bench를 사용하였다. 특히, Jailbreak Bench 내 일부 카테고리(A, B)를 각각 5개씩 샘플링하고, 나머지 데이터는 Alpaca로 구성하여 정상/유해 분포를 혼합한 형태로 실험 데이터를 구성하였다. 또한, 다양한 조건에서의 거부 행동을 유도하기 위해 임의의 Persona (John Smith, Alice, Homer Doe)를 프롬프트 prefix로 추가하여 입력을 구성하였다.

2.4 실험 결과

표 1은 제안한 계층적 거부 스티어링 벡터를 각 페르소나에 적용했을 때의 응답 변화 결과를 보여준다. 전반적으로 제안한 방법은 일부 조건에서 의도한 방향으로 동작하는 경향을 보였으나, 모든 계층에서 일관된 제어 성능을 달성하지는 못하였다.

구체적으로, Tier-1의 경우 A 그룹에 대해서는 정상 응답을 유도하는 방향으로 의도대로 동작하였으며, B 그룹에서도 응답률이 증가하는 등 전반적으로 긍정적인 성능 향상을 보였다. 그러나 Tier-2에서는 B 그룹에 대해서는 정상 응답 유도가 효과적으로 이루어진 반면, A 그룹에 대해서는 여전히 거부를 충분히 유도하지 못하는 한계를 보였다. 마지막으로 Tier-3의 경우, 상위 Tier (Tier-1, Tier-2)의 영향으로 A와 B 그룹 모두 억제되어야 함에도 불구하고, 오히려 해당 그룹들에서 응답이 발생하는 경향이 확인되었다. 또한, Benign 데이터에 대해서도 응답률이 소폭 감소하는 현상이 나타났으나, 이는 의도된 선택적 제어라기보다는 전반적인 억제 효과에 가까운 제한적인 결과로 해석된다.

이러한 결과는 제안한 단일 통합 벡터 기반 접근이 일부 조건에서는 유효하게 작동할 수 있으나, 계층 간 정책을 정밀하게 분리하고 안정적으로 제어하는 데에는 여전히 한계가 존재함을 시사한다.

3. 결론

본 논문에서는 기존 활성화 스티어링 기법의 한계를 극복하기 위해, 다수의 보안 정책을 단일 벡터 내에서 통합적으로 표현하고 제어할 수 있는 계층적 거부 스티어링 방법을 제안하였다.

실험 결과, 제안한 방법은 일부 Tier 설정에서 의도한 제어 동작을 성공적으로 수행하였으며, 특히 특정 조건에서는 응답 유도 및 억제 효과를 부분적으로 달성할 수 있음을 확인하였다. 이는 단일 벡터 기반 접근이 다중 보안 정책을 표현하는 데 있어 잠재적인 가능성을 지님을 시사한다. 그러나 계층 간 정책이 동시에 작용하는 환경에서는 조건 간 간섭이 발생하여, 특정 그룹에 대한 선택적 제어가 의도대로 이루어지지 않는 한계 또한 확인되었다.

따라서 향후 연구에서는 벡터 간 표현 충돌을 완화하고, 각 보안 조건을 보다 명확히 분리할 수 있는 구조적 설계가 필요하다.

참 고 문 헌

[1] A. Arditi, et al,, “Refusal in Language Models is Mediated by a Single Direction,” Proc. of the 38th International Conference on NeurIPS, no. 4322, pp. 1-47, 2024.

[2] A. Wei, N. Haghtalab, and J. Steinhardt, “Jailbroken: How Does LLM Safety Training Fail?,” Proc. of the 37th International Conference on NeurIPS, no. 3508, pp. 1-32, 2023.

[3] B. W. Lee, et al., “Programming Refusal with Conditional Activation Steering,” International Conference on Learning Representations (ICLR), 2025.