연세대학교 정보대학원 | 비즈니스 AI 빅데이터 분석 전공

2025년 12월 석사 학위논문

양자화된 대규모 언어 모델에서의 간접 프롬프트 인젝션 탐지 강건성 분석

Analysis of Robustness for Indirect Prompt Injection Detection in Quantized LLMs

연구 배경: 문제와 기회

보안 위협: 프롬프트 인젝션

LLM 에이전트가 외부 문서를 처리할 때, 악의적인 명령어가 포함된 **간접 프롬프트 인젝션**에 노출되어 의도치 않은 동작을 수행할 위험이 있습니다.

⚠ 사용자 의도 왜곡 및 정보 유출 위험

현실적 제약: 자원 효율성

서비스 배포 환경에서는 VRAM 용량과 추론 지연 시간(Latency)의 한계로 인해 **모델 양자화(Quantization)**가 선택이 아닌 필수로 요구됩니다.

FP32 대비 메모리 절감 필수

핵심 과제: 탐지 강건성

양자화로 정밀도 손실이 발생하여 활성화 값(Activation) 분포가 변할 때, **기존 보안 탐지 기법이 유효한가**에 대한 검증이 필요합니다.

INT8 환경에서의 탐지 성능 분석

RESEARCH OBJECTIVES

연구 목적과 핵심 질문

PRIMARY OBJECTIVE

실제 서비스 배포 환경을 고려한 **INT8 양자화 환경**에서 활성화 기반 **간접 프롬프트 인젝션 탐지 기법**의 성능과 강건성을 실증적으로 분석하고 검증하는 것.

1

탐지 성능 유지

FP32 원본 모델 대비 INT8 양자화 모델에서도 **탐지 성능 (ROC-AUC)**이 유의미하게 유지되는가?

2

최적 레이어 이동

양자화로 인한 노이즈가 누적되면서 **최적의 탐지 레이어 위치**가 변화하거나 이동하는가?

3

실용적 비교

자원 제약 시, **소형 FP32 모델 vs 대형 INT8 모델** 중 보안과 성능 면에서 어느 쪽이 우월한가?

4

자원 효율성

탐지 성능을 희생하지 않으면서 달성할 수 있는 실질적인 **메모리(VRAM) 절감 효과**는 어느 정도인가?

METHODOLOGY

방법론: 활성화 Δ 기반 탐지 프로세스

INT8 양자화 환경 적용 (LLM.int8())

1 두 가지 입력 비교

Clean Input

Baseline

(기본 태스크만)

VS

Potential Attack

With External

(외부 컨텐츠 포함)

동일한 프롬프트에 대해 **외부 컨텐츠 유무**에 따른 두 가지 입력을 준비합니다.

2 활성화 델타(Δ) 계산

Δa

Difference Calculation

Extract Layers:

L15 L23 L31

특정 레이어에서 두 입력 간의 **활성화 값 차이 (Δa)**를 추출하여 의도 이탈을 수치화합니다.

3 이상 탐지 분류

Logistic Regression

Clean | **Attack**

추출된 델타 값을 **선형 분류기**에 입력하여 최종적으로 공격 여부를 판별합니다.

EXPERIMENT SETUP

실험 설계: 모델, 데이터, 지표

Target Models

FP32 vs INT8 비교

- **Phi-3** · Microsoft — 3.8B
- **Mistral** · Mistral AI — 7B
- **Llama-3** · Meta — 8B

Datasets

태스크 및 공격 시나리오

Clean (Normal)

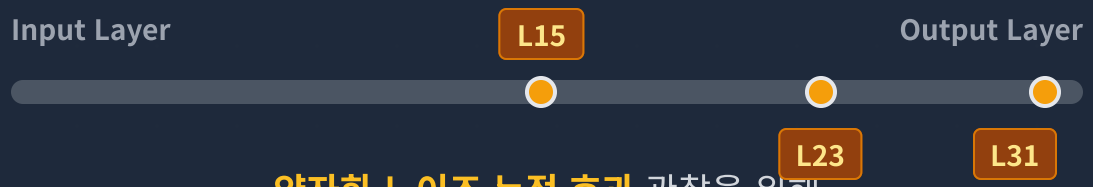
SQuAD · HotPotQA · Alpaca

Attack (Injection)

AdvBench · TrustLLM · BeaverTails · Do-Not-Answer

Observation Layers

활성화 패턴 분석 지점



양자화 노이즈 누적 효과 관찰을 위해
중반, 후반, 마지막 레이어 선정

Key Metrics

성능 및 효율성 지표

ROC-AUC
탐지 정확도
& 강건성

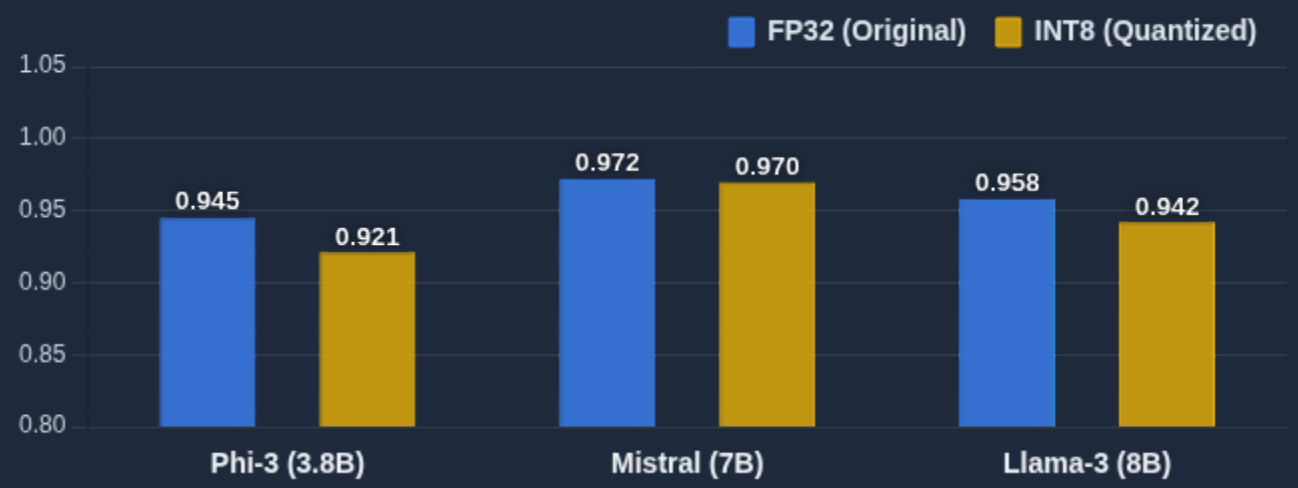
VRAM
메모리 사용량
(효율성)

KEY RESULTS

주요 결과: 성능 유지 + 자원 효율

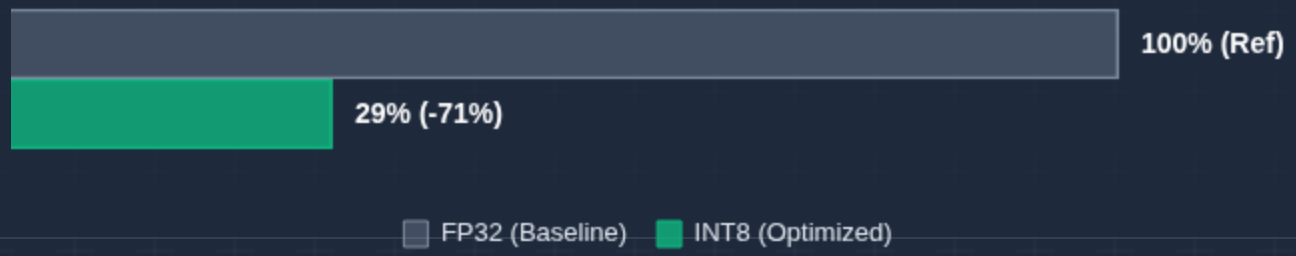
탐지 성능 비교 (ROC-AUC)

FP32 vs INT8



메모리 사용량 (VRAM)

약 70% 절감



높은 탐지 강건성 유지

INT8 양자화 시에도 모든 모델이 **ROC-AUC 0.92 이상**을 기록했습니다.

획기적인 자원 효율성

FP32 대비 VRAM 사용량을 **69.2% ~ 71.3%** 절감하여 효율성을 극대화했습니다.

최적 탐지 레이어 이동

양자화 노이즈로 인해 최적 탐지 지점이 앞당겨지는 경향을 확인했습니다.

결론 및 시사점

연구 요약 및 향후 전략 가이드

자원 효율성과 보안 강건성의 동시 확보

본 연구는 **INT8 양자화**가 LLM의 간접 프롬프트 인젝션 탐지 능력을 저해하지 않음을 입증했습니다. 오히려 메모리 자원을 70% 절감함으로써, 더 우수한 성능의 대형 모델을 보안 시스템에 도입할 수 있는 **실용적인 기회**를 제공합니다.

STRATEGY

실무 권고사항

- ✓ "Small FP32" 대신 "Large INT8" 모델 채택 권장
- ✓ 양자화 강건성이 입증된 아키텍처(Mistral) 우선 고려

OPERATION

운영 가이드라인

- ! **Mid-Layer:** 탐지 포인트를 중간(L15~20)으로 설정
- ! 이상치 처리를 위한 혼합 정밀도(FP16) 유지

NEXT STEP

한계 및 향후 연구

- → QAT 등 다양한 양자화 기법 비교 연구 필요
- → 실시간 환경에서의 탐지 지연(Latency) 최적화