

# 산업안전 대응을 위한 sLLM 질의응답 시스템의 응답 지연에 대한 상호작용 분석 연구

(Interaction Analysis Study on Response Delay of the sLLM Question-Answering System for  
Industrial Safety Response)

정도윤\*, 김남호\*\*

(Doyoon Jung, Namho Kim)

## 요약

본 연구는 산업안전 도메인에서의 AI 에이전트 기반 질의응답 시스템의 성능을 분석하는 것을 목적으로 한다. 산업안전 분야에서는 신속하고 정확한 정보 제공이 산업재해 예방과 대응에 중요한 역할을 한다. 이에 따라 본 연구에서는 대형 언어모델(LLM)을 활용한 질의응답 시스템의 응답 지연(latency) 성능을 비교·분석하였다. 성능 평가를 위해 두 가지 모델, LLaMA 3.1 Instruct와 Google Gemma 3 4B-IT를 동일한 조건에서 실험하였다. 실험 결과, Google Gemma 3 모델은 평균 응답 지연 시간이 88,583.71ms로 측정되었으며, LLaMA 3.1 모델은 47,534.75ms로 나타났다. 분석 결과, LLaMA 3.1 모델이 상대적으로 낮은 응답 지연을 보여 더 빠른 응답성을 보였다. 본 연구는 산업안전 관련 질의응답 환경에서의 대형 언어모델 성능 특성을 규명하고, 향후 산업안전 지식기반 시스템 개발에 유용한 기초자료를 제공한다는 점에서 의의가 있다.

■ 중심어 : 산업안전 ; AI 에이전트 ; 질의응답 시스템 ; 대형 언어모델 ; 성능 평가

## Abstract

This study aims to analyze the performance of an AI agent-based question-answering system in the industrial safety domain. In the industrial safety field, providing rapid and accurate information plays a critical role in preventing and responding to industrial accidents. Therefore, this study compared and analyzed the response latency performance of a question-answering system utilizing a large language model (LLM). To evaluate the performance, two models, LLaMA 3.1 Instruct and Google Gemma 3 4B-IT, were tested under identical conditions. The Google Gemma 3 model measured an average response latency of 88,583.71 ms, while the LLaMA 3.1 model measured 47,534.75 ms. The analysis revealed that the LLaMA 3.1 model exhibited a relatively low response latency, demonstrating faster response times. This study is significant in that it identifies the performance characteristics of large language models in an industrial safety-related question-answering environment and provides useful baseline data for the future development of industrial safety knowledge-based systems.

■ keywords : Industrial Safety ; AI Agent ; Question-Answering System ; Large Language Model ; Performance Evaluation ;

## I. 서론

산업안전 도메인에서의 AI 에이전트 기반 질의응

답 시스템은 산업 현장에서 요구되는 안전 규정, 법  
규, 사고 예방 지침 등을 신속하고 정확하게 제공하  
기 위한 중요한 수단으로 주목받고 있다[1]. 최근 산  
업현장은 복잡한 설비와 공정, 다양한 이해관계자가

\* 준회원, 호남대학교 컴퓨터공학과 박사과정, \*\* 정회원, 호남대학교 컴퓨터공학과

본 연구는 2025년도 교육부 및 광주광역시 지원으로 광주RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다.(2025-RISE-05-016)

접수일자 : 2025년 11월 28일

수정일자 : 2025년 12월 19일

게재확정일 : 2026년 01월 07일

교신저자 : 김남호 e-mail : nhkim@honam.ac.kr

공존하는 고위험 환경으로 변화하면서, 안전 관련 정보에 대한 실시간 접근성과 신뢰성 확보가 필수 요소로 부상하였다. 이러한 배경에서 산업안전 지식을 효율적으로 제공할 수 있는 AI 기반 질의응답 시스템의 필요성이 증대되고 있다. 특히 대형 언어모델의 발전과 함께 AI 에이전트는 자연어 이해 및 추론 능력을 바탕으로 사용자의 질의를 해석하고, 방대한 안전 규정 및 지침 데이터를 기반으로 맥락에 맞는 응답을 생성할 수 있는 수준에 도달하였다. 그러나 산업안전 도메인은 잘못된 응답이 실제 인명 피해와 직결될 수 있는 고위험 분야이므로, 모델의 응답 정확도뿐만 아니라 응답 지연 시간(latency), 일관성, 신뢰성 등 다양한 관점에서 성능을 체계적으로 평가하는 연구가 요구된다[2-3]. 이에 따라 AI 에이전트 기반 질의응답 시스템의 성능 특성을 정량적으로 분석하고, 산업안전 환경에 적합한 모델 선택 기준을 마련하는 것은 실무 적용을 위한 핵심 과제이다. 본 연구는 산업안전 도메인에서 활용할 수 있는 AI 에이전트 질의응답 시스템을 대상으로, 대표적인 대형 언어모델인 LLaMA 3.1 Instruct와 Google Gemma 3 4b-it의 응답 지연 성능을 비교·분석하는 데 목적을 둔다. 산업안전 분야에서 시간 민감도가 높은 질의응답 서비스를 설계할 때 모델 선택 또는 시스템 아키텍처 구성에 실질적인 기준을 제공하며, 향후 정확도·신뢰성·안전성 지표를 포함한 종합 성능 평가 연구로 확장될 수 있는 기초를 마련한다[4-5].

## II. 관련 연구

### 1. Llama 3.1 및 Gemma 2 기반 AI 에이전트 성능 분석 연구

최근 몇 년간 대규모 언어 모델(LLM)은 광범위한 주제와 고급 추론 능력을 요구하는 Massive Multitask Language Understanding (MMLU) 벤치마크를 비롯하여 여러 벤치마크에서 인간 수준을 뛰어넘는 언어 이해 능력을 입증하였다[6-7]. 그러나 이러한 LLM은 주로 영어와 같은 고자원 언어로 훈

련되었기 때문에, 우크라이나어와 같은 저자원 언어에서는 정확도 및 견고성(robustness) 저하를 초래하며 그 기능에 한계를 드러낸다. 더욱이, 수백억 개에 달하는 거대한 모델 크기는 광범위한 하드웨어 역량을 요구하여 자원이 제한적인 개인이나 조직에는 적합하지 않다는 문제가 있다.

본 연구는 이러한 문제점을 해결하기 위해 매개변수 효율적 미세 조정(PEFT) 및 사고의 사슬(Chain-of-Thought, CoT) 프롬프팅 방법을 결합하여, LLaMA 및 Gemma와 같은 소형 오픈 소스 언어 모델의 성능을 향상하는 데 그 목적을 둔다.

PEFT(Parameter-Efficient Fine-Tuning)와 CoT 추론을 결합한 제안된 방법이 LLaMA 및 Gemma와 같은 소형 모델의 성능을 크게 향상할 수 있음을 입증하며, 저자원 설정 및 저자원 언어 분야에 중요한 이바지를 한다. 특히, 단일 A100 GPU 및 4-bit 양자화 기술을 사용하여 고급 자연어 처리 시스템을 계산 제약 환경에서 훈련할 가능성을 보여주었다는 점에서 실질적인 의의를 가진다.

### 2. 지속 가능한 엣지 AI를 위한 LLM 추론 연구

대규모 언어 모델(LLM)을 엣지 디바이스에 배포하는 것은 계산 제약, 메모리 한계, 추론 속도 및 에너지 소비 측면에서 상당한 도전 과제를 제시한다. 이러한 제약을 완화하기 위해 모델 양자화(Model Quantization)는 모델 크기와 계산 오버헤드를 줄여 효율적인 LLM 추론을 가능하게 하는 핵심 기술로 부상하였다. 그러나 양자화는 종종 모델 정확도에 영향을 미쳐 개발자들은 계산 효율성과 예측 성능 간의 상충 관계(trade-off)를 신중하게 조정해야 한다[8-9]. 기존 연구는 양자화 방법론, 정확도 상충 관계 또는 다른 모델 압축 기법에 중점을 두었으나, 실제 엣지 환경 배포 시 양자화된 LLM의 에너지 소비에 대한 연구는 제한적이었다. 또한, 다양한 LLM 모델군과 작업 유형에 걸쳐 여러 양자화 기법이나 수준을 체계적으로 비교한 연구가 부족했다. 본 연구는 이러한 문제점을 해결하기 위해 엣지 디바이스에 배포된

양자화 LLM에 대한 체계적인 비교해 분석을 제시한다. 주된 목적은 다른 양자화 설정 하에서 에너지 효율성을 측정 및 평가하고, 이러한 구성이 다중 데이터 세트 및 작업에 걸쳐 추론 정확도와 지연 시간에 미치는 영향을 평가하는 것이다.

본 연구는 산업안전 도메인에서 시간 민감도가 높은 질의응답 서비스를 제공하는 것을 목적으로 하며, 경량 대형 언어모델(sLLM)이 Edge 컴퓨팅 환경에서 효율적으로 구동되어 서비스 속도를 향상할 수 있다는 가정을 참고하였다.

관련 연구에서는 Edge AI 환경의 대표적인 플랫폼인 Raspberry Pi 4 (4GB RAM)를 사용하여 LLM 추론 연구를 수행하였다. 이 연구는 Ollama 라이브러리에서 제공되는 Gemma 2 (2B), LLaMA 3.2 (1B) 및 Qwen 2.5 (0.5B, 1.5B) 기반 모델들을 포함한 28가지 양자화된 LLM 변형을 활용하였다.

해당 연구에서 Ollama 프레임워크는 PTQ (Post-Training Quantization)와 가중치 전용 양자화 (weight-only quantization) 기법을 기본적으로 적용하여 모델의 경량화를 실현하였다. 실험 결과, 긴 응답 길이를 요구하는 HumanEval (코드 생성) 작업은 가장 높은 지연 시간(평균 94.84)을 보였으며, 이는 응답 길이가 길수록 에너지 소비 및 지연 시간과 높은 상관관계가 있음을 입증한다. 이는 본 연구에서 응답 지연 시간 분석을 위해 sLLM 모델을 비교하고 응답 길이의 영향을 분석하는 근거가 된다.

### III. 제안

#### 1. 산업안전 도메인에서의 AI 에이전트 질의응답 시스템 성능 분석

본 연구에서는 산업안전 도메인에서의 AI 에이전트 질의에 대한 멀티 에이전트 시스템을 제안하고자 한다.

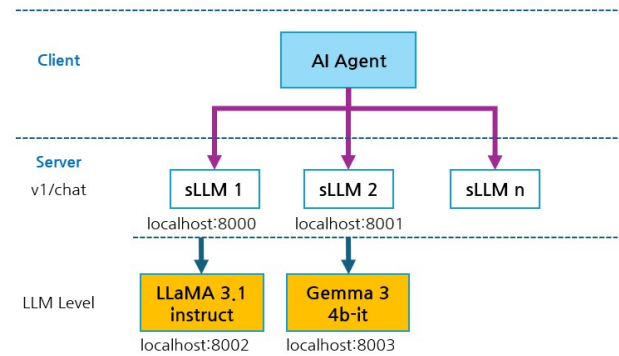


그림 1. Edge AI Service System Architecture Based on a Lightweight, Large-Scale Language Model

그림 1은 본 연구에서 제안하는 AI 에이전트 시스템이다. 본 시스템 아키텍처는 클라이언트의 AI 에이전트가 다중 sLLM 서버(v1/chat)로 요청을 분산 처리하며, 각 sLLM 서버는 독립적으로 로컬에서 특정 경량 LLM(LLaMA 3.1 instruct, Gemma 3 4b-it 등)을 구축하여 서비스한다. 각 서버는 서로 상이한 LLM 모델과 포트를 통해 제공되며, 계층적으로 서버와 LLM 인스턴스가 연결되어 확장할 수 있는 경량 대형 언어 모델 환경을 구성한다.

그림 2는 본 시스템에서 사용한 멀티 에이전트 LLM의 처리 방법에 대한 흐름도이다[13].

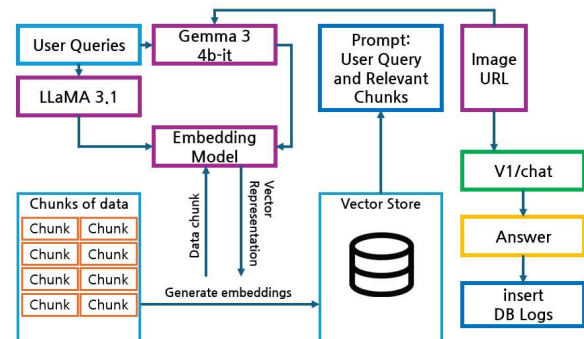


그림 2. Proposed multi-agent LLM processing model

본 시스템 구조는 사용자의 질의가 LLaMA 3.1 또는 Gemma 3 4b-it 모델로 전달되고, 데이터는 적절히 분할되어 임베딩 모델을 통해 벡터 표현으로 변환된다. 이 벡터 표현은 벡터 저장소에 저장되어, 질의 및 관련 데이터 청크와 결합한 프롬프트가 생성된다. 프롬프트와 이미지 URL 정보는 v1/chat API를 통해 대화형 예측을 수행하며, 결과로 생성된 응답은 데이터베이스 로그에 기록된다. 이러한 구성은 에지 환경에서 문서 연동 및 멀티모달 질의응답이 가능한 경량

대형 언어 모델 기반 시스템의 운영 흐름을 나타낸다 [14-15].

## 2. 사용자 인터페이스

본 연구에서는 모델과 기준 서버, API Key를 설정할 수 있는 형태로 멀티모달 클라이언트를 그림 3처럼 설계하였다. AI Agents - Multimodal Client는 사용자 질의에 따라 멀티모달 응답 기록 및 시스템 상태, 모델·서버 설정, 백업·리셋 기능 및 토큰 사용 현황 등을 GUI로 통합 관리한다. 사용자는 질의 입력, 프롬프트 및 이미지 URL 추가가 가능하며, 선택된 언어 모델과 API 키, 서버 주소에 따라 대화형 서비스를 실시간으로 제공받는다.

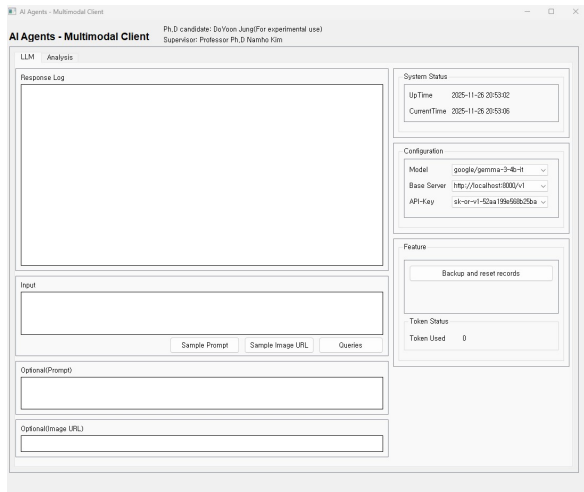


그림 3. UI design for multi-agent user clients  
시스템은 응답 로그와 데이터 입력 창을 분리하고, 운영 시간 및 현재 시각을 모니터링함으로써 안정적인 실험 환경을 지원한다.

## IV. 실험 결과 및 분석

본 연구에서는 Hugging Face의 오픈 소스 sLLM을 활용하여 응답 지연 시간과 답변 품질에 대한 성능 분석 실험을 수행하였다.

표 1은 본 실험에 사용된 모델들을 정리한 것이다.

표 1. 실험에 사용된 모델 정리

모델명	구분	파인튜닝된 데이터 세트 갯수
LLaMA 3.1 Instruct	Fine-tuned	2,052
Google Gemma 3 4b-it	Base Model	n/a

파인튜닝된 LLaMA 3 Instruct 모델과 Google Gemma 3 4b-it 모델을 대상으로 다양한 측면에서 성능을 비교·분석하였다.

### (1) 실험 환경

실험 환경은 표 1과 같으며, PyTorch 2.12를 사용하였으며, Cuda는 12.4버전을 사용하였다.

표 2. 실험 환경

구분	내용
OS	Windows 11 Pro
CPU	Intel Core i5 13세대 13400F(2.5Ghz)
RAM	80GB(DDR5-5600)
GPU 1	NVIDIA Geforce RTX 3090 (GDDR6 24GB)
GPU 2	NVIDIA T1000 (GDDR6 8GB)
Disk	SSD 6TB + HDD 4TB
Python	3.11
Framework	PyTorch 2.12 + Cuda 12.6

### (2) 응답 지연시간 측정 결과

산업안전 관련 표준 질의 20개를 각 모델에 입력하여 응답 시간을 측정하고, 그 평균값을 산출하였다. 질의 내용은 "중대재해처벌법 대응 방안", "화학물질 누출 시 조치 요령", "산재보험 처리 기준" 등 다양하게 구성되었다. 표 3은 모델별 평균 응답 지연시간 비교에 대한 것이다. 표 3에서 알 수 있듯이, LLaMA 3.1 모델은 Gemma 3 대비 절대적인 응답 시간뿐만 아니라 초당 토큰 생성 수(TPS) 측면에서도 우위를 보였다. 이는 LLaMA 3.1이 단순히 짧은 답변을 내놓기 때문에 빠른 것이 아니라, GQA(Grouped-Query Attention) 등 아키텍처 최적화를 통해 동일 자원 대비 추론 효율성이 더 높음을 시사한다.

표 3. 모델별 평균 응답 지연시간 비교

모델명	평균 지연시간(ms)	표준 편차 (ms)	환산 시간 (초)	TPS (Tokens /sec)
LLaMA 3.1 Instruct	47,534.75	15,200	약 47.53	3.78
Google Gemma 3 4b-it	88,583.71	24,150	약 88.58	2.82

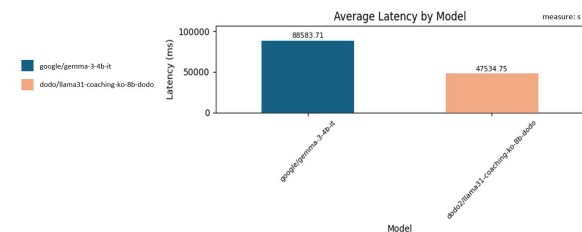


그림 4. Graph for average latency

그림 4는 Google의 Gemma 3 모델과 파인튜닝된 LLaMA 3.1 Instruct 모델의 질의 대기시간을 측정된 결과를 나타낸다. 그림 5에서는 두 모델에 대하여 각각 20회씩 LLM 질의 작업을 수행했을 때 기록된 DB 로그를 기반으로 대기시간을 비교하였다.

txt_response_msg	user_ip	regidate	token_used	latency_ms
어려움을 겪는 사람들을 돕고 싶다는 당신의 마음은 정...	119	2025-11-26 17:50:02	20	88,302
어려움을 겪는 사람들을 돕고 싶다는 당신의 마음은 정...	119	2025-11-26 17:56:57	20	102,952
사회적 기억을 통해 사람들은 어려움을 겪는 사람들...	119	2025-11-26 18:08:40	20	31,150
컴퓨터공학과 공부하는 약 4년이 소요됩니다. 학사 과정...	119	2025-11-26 18:10:03	20	34,358
어려움을 겪는 사람들을 돕고 싶다는 당신의 마음은 정...	119	2025-11-26 18:13:32	20	81,833
컴퓨터공학과는 일반적으로 4년제 학사 과정을 거쳐 중...	119	2025-11-26 18:15:19	20	79,871
중대재해처벌법 발령 시 대응 방안은 다음과 같습니다. *	119	2025-11-26 18:31:19	20	107,239
직장 내에서 발생한 산업재해를 산재보험으로 처리받으...	119	2025-11-26 18:34:23	20	79,640
중대재해처벌법은 안전 관리를 잘하면 중대재해가 발생...	119	2025-11-26 18:37:14	20	75,423
산업재해는 고의로 일으킨 경우가 존재합니다. 이러한 ...	119	2025-11-26 18:38:22	20	49,208

그림 5. Some views of latency times recorded in DB logs  
실험 결과, LLaMA 3.1 모델이 Gemma 3 모델 대비 약 1.86배 빠른 속도를 기록하였다. 두 모델 간의 지연시간 차이는 약 41초에 달하며, 이는 긴급 상황에서 매우 유의미한 차이이다. 본 실험에서 Google Gemma 3 모델의 88초와 LLaMA 3.1 모델의 47.53초의 응답 지연시간은 산업재해 대응과 같은 민감도가 매우 높은 긴급 상황에서 신속한 정보 제공을 위한 시스템 요구사항에 미치지 못하는 실정이다.

그러나 본 실험의 실험 환경이 Edge AI 환경(GPU 1: RTX3090, GPU 2: T1000)이며, 모델 양자화(4b-it) 및 RAG 연산이 포함된 End-to-End 지연시간임을 감안해야 한다.

LLaMA 3.1의 경우 미세 조정을 통해 답변의 길이

를 최적화하고 불필요한 서술을 줄이는 방향으로 학습되었을 가능성이 있으며, 아키텍처 자체의 연산 효율성(예: Grouped-Query Attention 등)이 영향을 미친 것으로 분석된다. 반면 Gemma 3 모델은 상대적으로 더 깊은 연산 과정을 거치거나, 생성하는 토큰의 수가 더 많아 지연 시간이 길어진 것으로 추정된다.

### (3) 답변 품질 및 내용 분석

본 연구에서는 산업안전 재해 도메인을 대상으로 한 gemma 3 모델과 Fine-Tuning된 LLaMA 3.1 Instruct 모델을 바탕으로 답변의 질적인 측면에 대해서도 실험하였다. 그림 6은 “중대재해처벌법이 발생하면 어떻게 대응해야 하나요??”라는 프롬프트를 입력했을 때 반응이다.

Response Log

질문 시간: 2025-11-26 18:31:19  
질문: 중대재해처벌법이 발생하면 어떻게 대응해야 하나요??  
답변: 중대재해처벌법 발령 시 대응 방안은 다음과 같습니다.

1. 상황 인지 및 초기 대응

- 1. 사고 발생 즉시: 사고 발생 사실을 즉시 인지하고, 관계 당국(고용노동부, 경찰 등)에 신고합니다. 사고 경위, 피해자 현황, 사고 원인 등을 정확하게 파악하고 기록합니다.
- 2. 대응 절차 수립: 사고 발생 시 대응 절차를 사전에 수립하고, 관계 당국과 조율하여 적극적으로 협조합니다. 필요한 자료를 제공하고, 조사에 대한 공문을 정확하게 작성하여 답변합니다.
- 3. 피해자 지원: 사고 피해자 및 그 가족에게 필요한 의료 지원, 경제적 지원, 심리 상담 등을 제공합니다.
- 4. 사후 보고 및 검토: 사고 발생 사실을 중정신에게 보고하고, 사고 원인 분석 및 재발 방지 대책 수립을 위한 내부 검토를 진행합니다.

2. 법적 책임 및 대응

- 1. 법적 책임 전담: 중대재해처벌법은 사업주 또는 경영책임자가 중대재해를 예방하기 위한

System Status

UpTime: 2025-11-26 18:28:58  
CurrentTime: 2025-11-26 18:31:34

Configuration

Model: google/gemma-3-4b-it  
Base Server: http://localhost:8000/v1  
API-Key: sk-or-v1-52ba193e560825ba

Feature

Backup and reset records

Token Status

Token Used: 20

Input

중대재해처벌법이 발생하면 어떻게 대응해야 하나요??

Sample Prompt Sample Image URL Queries

그림 6. Results of the GEMMA 3-4b-it inquiry regarding the Serious Disaster Punishment Act

Gemma 3 모델은 사고 발생 시 필요한 조치 방법에 대해 상세하고 체계적으로 설명할 수 있음을 확인하였다. 그림 7은 ‘직장 내 산업재해 발생’이라는 복잡한 질의에 대해 Gemma 3 모델을 활용한 사례를 보여준다. 예시 질의로 ‘직장 내에서 산업재해가 발생하였을 때 산재보험 처리에 있어 고의적 사고에 대한 판정 기준’이 제시되었으며, Gemma 3 모델은 이와 관련해 엄격한 법적 절차와 구체적 증거 확보의 필요성 등 다방면의 판단 요소를 질문의 해석으로 제시하였다.

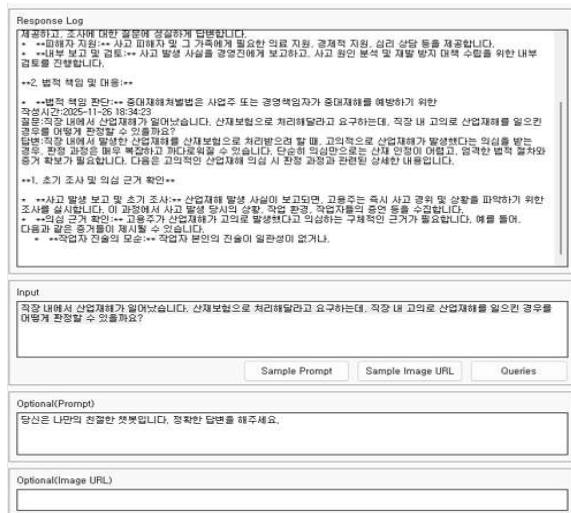


그림 7. Gemma 3-4b-it Inquiry Results on Industrial Accidents

그림 8은 중대재해처벌법에 대한 질의와 산업재해에 관한 질의를 LLaMA 3에 수행해 본 모습이다.

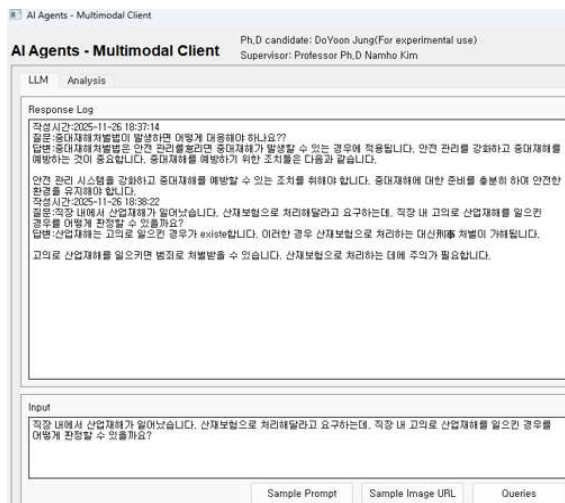


그림 8. Results of a fine-tuned LLaMA 3.1 Instruct inquiry on the Serious Accident Punishment Act and industrial accidents.

LLaMA 3.1 Instruct 모델은 명령형 구조의 응답 생성을 지원하는 언어 모델이며, Gemma 3b 4b-it 모델과 비교 시 더욱 명확하고 지시 기반의 언어 생성 결과를 산출할 수 있음을 확인하였다. 다만, 실험 과정에서 일부 할루시네이션 현상이 관찰되었으며, 이러한 오류를 개선하기 위해서는 추가적인 파인튜닝 작업이 필요하다.

## V. 결 론

본 연구는 산업안전 도메인에서의 AI 에이전트 도

입을 위한 기초 연구로서, 엣지 환경에서 구동되는 LLaMA 3.1과 Gemma 3 모델의 성능을 비교하고 분석하였다. 실험 결과, LLaMA 3.1 Instruct 모델은 약 47초의 지연 시간을 기록하여 신속성 측면에서 우수성을 보였고, Gemma 3 4b-it 모델은 약 88초로 속도는 느리지만 답변의 상세함과 논리적 완결성에서 강점을 보였다.

이러한 결과는 산업안전 지식기반 시스템을 구축할 때, 사용 목적(긴급 대응 vs 심층 분석)에 따라 적합한 모델을 선택해야 함을 시사한다. 특히 중대재해 처벌법과 같이 법적 민감도가 높은 영역에서는 환각 현상을 억제하고 답변의 근거를 명확히 제시할 수 있는 RAG 기술의 고도화가 필수적이다.

본 연구는 실험실 환경에서의 단일 턴 질의응답에 국한되었다는 한계가 있다. 향후 연구에서는 실제 시끄럽고 네트워크가 불안정한 산업 현장에서의 필드 테스트를 수행하고, CCTV 영상 등을 분석하여 위험을 감지하는 멀티모달(Multi-modal) 모델과의 연동성을 검증하는 방향으로 확장되어야 할 것이다. 또한, 지속적인 데이터 업데이트와 사용자 피드백을 통한 강화학습(RLHF) 파이프라인을 구축하여, 살아있는 안전 지능 시스템으로 발전시켜 나갈 계획이다.

## REFERENCES

- [1] M. Jeon, J. Kim, H. Park, S. Lee, J. Noh and J. Park, "A Large Language Model-based Trajectory Prediction Algorithm for Industrial Accident Prevention," *Smart Media Journal*, vol. 14, no. 2, pp. 53-63, 2025.
- [2] C. Kim, M. Kim, H. Shin, S. Lee, and J. Choi, "An Efficient RAG Invocation Method Based on Query Type and Confidence," *The Journal of KINGComputing*, vol. 21, no. 5, pp. 45-56, 2025.
- [3] Y. Jeong, "Performance Evaluation of AI-IoT Platform Based on Blockchain in Cloud Environment," *Industry and Science*, vol. 4, no. 3, pp. 67-73, 2025.
- [4] J. Park and Y. Baek, "Fuzzing-based Security Vulnerability Detection Method in Model Context Protocol (MCP) Environments," *JKIOTS (Journal of Korean Internet of Things Society)*, vol. 11, no. 5, pp. 3, 2025.
- [5] J. Kwak and Y. Jeong, "Enhancing the



- Performance of Low-Resource Text Classification through KoBERT-based Category Mapping and LLM Integration,” *Journal of Korean Institute of Information Technology*, vol. 23, no. 1, pp. 1-11, 2025.
- [6] M. Syromiatnikov, V. Ruvinskaya, and N. Komleva, “Empowering Smaller Models: Tuning LLaMA and Gemma with Chain-of-Thought for Ukrainian Exam Tasks,” arXiv preprint arXiv:2503.13988, 2025.
- [7] S. H. Kim, Y. W. Yu, J. E. Park, S. H. Noh, and N. R. Son, “A Case Study on AI Comment Detection Based on Large Language Models: The 21st Korean Presidential Election,” *Smart Media Journal*, vol. 14, no. 10, pp. 149-160, 2025.
- [8] E. J. Husom, A. Goknil, M. Astekin, L. K. Shar, A. Kåsen, S. Sen, B. A. Mithassel, and A. Soyly, “Sustainable LLM Inference for Edge AI: Evaluating Quantized LLMs for Energy Efficiency, Output Accuracy, and Inference Latency,” *ACM Transactions on Internet of Things*, vol. 6, no. 4, pp. 1-35, 2025.
- [9] J. Lee, “Research on Generative AI Chatbot Service based on Academic Data,” *Smart Media Journal*, vol. 13, no. 12, pp. 122-129, 2024.
- [10] S. Park, W. Choi, J. Ahn, C. Yoo, and G. Yang, “Fuzzing-based Security Vulnerability Detection Method in Model Context Protocol (MCP) Environments,” *Proceedings of the Korean Institute of Information Scientists and Engineers Conference*, vol. 2025, no. 7, 2025.
- [11] S. Son, J. Kim, C. Cho, and S. Park, “Trends in Network Optimization Using Large Language Models,” *The Journal of Korean Institute of Communications and Information Sciences*, vol. 50, no. 7, pp. 1073-1084, 2025.
- [12] Y. Kim and M. Oh, “Implementation of an LLM-based voice chatbot for edge computing to monitor private sensing data,” *Proceedings of the 2025 Summer Conference of the Korean Institute of Communications and Information Sciences*, pp. 1953-1954, 2025.
- [13] D. Jung and N. Kim, “A Study on the Development of a Career Aptitude Test System in the Form of a Customized Multi-Layer Questionnaire Using sLLM,” *Smart Media Journal*, Vol. 14, No. 8, pp. 42-49, 2025.
- [14] H. Ahn, S. Son, W. Jung, and J. Kim, “Multi Modal LLM Technology Trends,” *Journal of Korean Institute of Information Scientists and Engineers*, Vol. 43, No. 5, pp. 29-39, 2025.
- [15] S. Im and Y. Ko, “Multi-Modal Deep Learning

Model for Drug-Drug Interaction Prediction Using Structural and LLM-Based Embeddings,” *Proceedings of the Korean Institute of Information Processing Society Conference*, Vol. 32, No. 1, pp. 711-712, 2025.

#### 저 자 소 개



정도윤(준회원)

2025년 호남대학교 대학원  
컴퓨터공학과 석사 졸업  
2025년 호남대학교 대학원  
컴퓨터공학과 박사과정

<주관심분야 : 컴퓨터비전, 자연어 처리, 인공지능>



김남호(정회원)

1997년 포항공과대학교 정보통신  
학과 석사 졸업  
2013년 전남대학교 전산통계학과  
박사 졸업  
1991년~1997 포스코DX(주) 선임

연구원

1998년~현재 호남대학교 컴퓨터공학과 부교수

<주관심분야 : AI에이전트, 사물인터넷, 인공지능, 응용SW, 스마트제조>