생성형 AI 및 LLM 기반 소형 LLaMA 3.1 8B 모델의 70B급 AI 면접 평가 모델 성능 실증 연구

류지수[1], 정순기[2*]
[1]경북대학교 컴퓨터학부 대학원, [2]경북대학교 컴퓨터학부 교수

Empirical Study on 70B-Level AI Interview Evaluation Performance of Small LLaMA 3.1 8B Using Generative AI and LLM Techniques

Ji Soo Ryu[1], Soon Ki Jung[2*]
[1]Graduate School of Computer Science and Engineering, Kyungpook National University
[2]Professor, School of Computer Science and Engineering, Kyungpook National University

요 약 본 연구는 온프레미스 환경에서 동작 가능한 AI면접관 시스템 통합 프레임워크를 제안한다. 제안된 시스템은 STT-LLM 통합 파이프라인, Rasa 기반 대화 관리 엔진, KoBERT 및 MeCab 기반 언어 전처리 모듈로 구성되어 실시간 면접 대화 처리와 자동 평가 및 피드백 보정을 수행한다. 실험에서는 응시자의 발화를 STT로 변환한 후, LLM이 면접관 역할을 수행하도록 구현하였으며, 생성된 응답을 전문성, 논리성, 정합성, 피드백 유효성 기준으로 자동 평가하였다. 특히 RAG와 Reflection Tuning 기법을 적용하여 응답 품질과 평가 일관성을 향상시켰다. 그 결과, 제안한 구조를 적용한 LLaMA 3.1 8B 모델은 동일 조건에서 대형 70B 모델 대비 경쟁력 있는 평가 성능을 보였으며, 비열등성 관점에서 활용 가능성이 확인되었다. 본 연구는 소형 언어모델이 구조적 설계와 학습 전략을 통해 대형모델에 근접한 성능을 달성할 수 있음을 실험적으로 고찰한다.

주제어 : 인공지능 면접관, 온프레미스 인공지능, 음성 인식 기반 면접, 검색 증강 생성, 면접 응답 자동 평가

Abstract This study proposes an integrated AI interviewer system framework for on-premise environments. The system combines an STT-LLM pipeline, a Rasa-based dialogue manager, and a KoBERT-MeCab preprocessing module to enable real-time interview dialogue processing, automated evaluation, and feedback refinement. Applicants' spoken responses are converted via STT, after which the LLM functions as an interviewer and evaluates responses based on professional relevance, logical consistency, coherence, and feedback usefulness. The application of Retrieval-Augmented Generation (RAG) and Reflection Tuning improves response quality and evaluation consistency. Experimental results show that the LLaMA 3.1 8B model with the proposed architecture demonstrates competitive performance compared to the 70B model under identical conditions, supporting its applicability from a non-inferiority perspective. These findings suggest that small-scale language models can achieve performance comparable to larger models through appropriate architectural design.

Key Words : AI Interviewer, On-Premise Artificial Intelligence, Speech Recognition-Based Interview, Retrieval-Augmented Generation (RAG), Automated Interview Response Evaluation

본 논문은 과학기술정보통신부 및 정보통신산업진흥원의 지역산업 SW인재양성기반조성 사업 지원으로 연구되었음.
*교신저자 : 정순기(skjung@knu.ac.kr)
접수일 2025년 11월 14일 수정일 2025년 12월 08일 심사완료일 2025년 12월 21일

1. 서론

최근 생성형 인공지능(Generative AI)과 대규모 언어 모델(LLM: Large Language Model)의 급속한 발전은 인간–기계 간 상호작용의 품질을 획기적으로 향상시키고 있다. 특히 면접, 채용, 상담과 같은 평가 중심 대화 상황에서 LLM의 활용 가능성은 매우 높게 평가된다. 그러나 이러한 응용 분야에서는 단순히 문장을 생성하거나 질문에 답변하는 수준을 넘어, 전문가 수준의 지식 기반 응답 능력과 지원자의 인성·태도 등 정성적 요소를 종합적으로 해석하는 평가 지능이 요구된다. 즉, 면접 평가 AI는 언어적 정확성과 논리성뿐 아니라, 응시자의 의도·감정·사회적 맥락을 파악할 수 있는 심층적 인지 능력과 인성 판단 능력을 갖추어야 한다.

기존의 대형 모델(예: LLaMA 3.1 70B, GPT-4, Claude 3 등)은 방대한 파라미터 수를 기반으로 우수한 언어 표현력과 추론 성능을 보였으나, 막대한 연산 자원과 메모리 요구량으로 인해 온프레미스(On-premise) 환경이나 기관 내부 서버에서의 실질적 운용에는 한계가 존재한다.

대형 모델을 내부 서버에 구축할 경우, 학습 및 추론에 필요한 GPU 메모리와 병렬 연산 자원이 기하급수적으로 증가하며, 이에 따른 학습 서버 구축 비용, 전력 소모, 냉각·유지비용이 급격히 상승한다. 이러한 경제적·물리적 제약으로 인해 공공기관이나 중소규모 조직은 대형 모델의 자체 운용이 사실상 어렵다.

특히 공공기관, 금융기관, 교육기관 등은 개인정보 유출을 방지하고 내부 보안을 강화하기 위해 외부 클라우드 기반 모델 대신 자체 서버 내 운용이 가능한 온프레미스 LLM의 도입이 필수적이다. 그러나 이러한 환경에서는 고성능과 보안을 모두 확보하면서도, 제한된 하드웨어 자원에서 안정적으로 작동할 수 있는 경량형 고성능 모델이 요구된다.

이를 해결하기 위해 본 연구에서는 소형 LLaMA 3.1 8B 모델을 기반으로 한 온프레미스 AI 면접 시스템을 제안한다. 8B 모델은 파라미터 수가 상대적으로 적어 연산 효율이 높고, GPU 자원 소모가 대폭 줄어들어 서버 구축 비용을 70B 모델 대비 약 1/10 이하로 절감할 수 있다.

또한, 단순한 모델 경량화에 그치지 않고, 웹 크롤링을 통한 최신 기술 정보 및 도메인 전문 데이터의 지속적 업데이트를 수행하여 지식 신선도를 유지하였다.

아울러, RAG (Retrieval-Augmented Generation) 기법을 결합하여 모델이 외부 지식베이스에서 필요한 정보를 실시간으로 검색·참조함으로써, 소형 모델의 한계를 보완하고 전문가 수준의 답변 품질을 확보하였다.

이러한 구조는 대형 모델에 비해 훨씬 가볍지만, 최신 정보 접근성과 실시간 지식 확장 능력을 갖추고 있어 온프레미스 환경에서도 고품질의 평가·판단 기능을 구현할 수 있는 현실적인 대안으로 작용한다. 또한 모든 데이터가 내부망에서 관리되기 때문에, 보안성·프라이버시 보호·감사 추적성 측면에서도 공공기관이나 민감 정보 처리 환경에 적합하다.

물론 소형 모델은 모든 응답에서 대형 모델 수준의 완전한 정밀성과 표현력을 재현하기는 어렵지만, '가상면접(Virtual Interview)'이라는 특수 목적 도메인에 맞게 전문적으로 설계·튜닝한다면, 대형 LLaMA 3.1 70B 모델에 준하는 실질적 성능을 충분히 달성할 수 있다. 본 연구에서는 이러한 목표를 검증하기 위해, LLaMA 3.1 8B 모델을 중심으로 최신 도메인 데이터, RAG 기반 지식 확장, 인성·태도 분석 모듈을 결합한 AI 면접 평가 시스템을 구현하였다.

실험 결과, 제안된 시스템은 평가 정확도, 인성 판단 일관성, 반응 속도 등 주요 지표에서 LLaMA 3.1 70B 모델 대비 95% 이상의 상대적 성능을 보였다. 이는 소형 LLM의 지식 효율화와 온프레미스 환경에서의 실용적 활용 가능성을 입증한 중요한 사례로, 향후 공공기관 및 민간 기업의 보안 강화형 AI 면접 및 평가 시스템 구축에 실질적인 기여를 할 것으로 기대된다.

2. 이론적 배경

2.1 생성형 인공지능(Generative AI)과 LLM의 개요

2.1.1 생성형 인공지능의 개념 및 발전

생성형 인공지능(Generative Artificial Intelligence, Generative AI)은 기존의 판별형 인공지능(Discriminative AI)과 달리, 주어진 입력 또는 조건에 따라 새로운 데이터 샘플을 생성할 수 있는 인공지능 기술을 의미한다.[1] 이는 데이터의 확률분포를 직접 학습하여, 학습 데이터와 유사한 특성을 가지면서도 새로운 형태의 결과물을 생성하는 생성적 모델링(generative modeling)에 기반한다.[2]

초기에는 GAN(Generative Adversarial Network)과 VAE(Variational Autoencoder) 와 같은 신경망 기반 생성 모델이 주류를 이루었다.

이들 모델은 주로 이미지 생성, 음성 합성 등 특정 도메인에 한정된 생성 작업에 활용되었으나, 이후 Transformer 구조의 등장과 함께 자연어 처리(NLP) 영역에서도 언어 생성(generative language modeling)이 가능해지면서 생성형 AI의 응용 범위가 급격히 확대되었다[3].

2.1.2 Transformer 기반 언어모델의 등장과 확장

Vaswani 등(2017)은 「Attention Is All You Need」 논문에서 Transformer 구조를 제안하며, 순차적 정보처리에 의존하던 RNN(Recurrent Neural Network) 및 LSTM(Long Short-Term Memory) 기반 모델의 한계를 극복하였다[4].

Transformer는 Self-Attention 메커니즘을 통해 문장 내 단어 간의 전역적 의존 관계를 병렬적으로 학습함으로써,긴 문맥을 효율적으로 이해하고 대규모 데이터 학습에 적합한 구조를 제공하였다

이후 OpenAI(GPT 시리즈), Google(BERT, PaLM 시리즈), Meta(LLaMA 시리즈), Anthropic(Claude) 등 주요 연구기관에서 수십억~수천억 개의 파라미터를 가진 초대형 언어모델을 발표하면서, 대규모 사전학습(pretraining) 과 미세조정(fine-tuning) 을 결합한 LLM 시대가 열렸다[5].

특히, Kaplan et al.(2020)이 제시한 Scaling Law에 따르면, 모델의 파라미터 수와 학습 데이터의 양을 일정한 비율로 확장할 경우 언어모델의 성능이 지속적으로 향상된다는 것이 실증되었다.

이러한 이론적 근거를 바탕으로 GPT, PaLM, LLaMA 등의 모델은 수천억 개 파라미터 규모의 학습을 통해 인간 수준의 언어 이해 및 생성 능력을 달성하게 되었다[6].

2.1.3 대규모 언어모델(LLM)의 정의 및 특징

대규모 언어모델(LLM: Large Language Model)은 수십억~수천억 개의 파라미터(parameter)를 학습한 신경망으로, 방대한 텍스트 코퍼스를 기반으로 문맥적 확률 분포를 학습한다.

이로써 LLM은 인간 수준의 텍스트 생성, 질의응답(QA), 요약(Summarization), 번역(Translation), 추론(Reasoning) 등 복합적 언어처리가 가능하다.

LLM의 주요 특징은 적은 예시만으로 새로운 문제를 해결할 수 있는 Few-shot / Zero-shot 학습 능력, 대화의 흐름 내 의미적 일관성을 유지하는 문맥 기반 추론(Contextual Reasoning), 도메인 전이 성능이 높은 범용성(Generalization) 및 자기지도학습(Self-supervised Learning) 기반으로 대규모 비정형 데이터 학습 가능하다는 점이다[7][8].

2.2 LLaMA 시리즈의 기술적 특성

Meta AI에서 개발한 LLaMA(Large Language Model Meta AI) 시리즈는 오픈소스 연구자들이 접근 가능한 효율적 대규모 언어모델(foundation LLM)을 목표로 설계되었다.

LLaMA는 GPT 계열과 동일한 트랜스포머(Transformer) 디코더 기반 구조를 따르며, 적은 학습 자원으로도 고성능을 달성하도록 파라미터 효율화 및 데이터 구성 전략을 중점적으로 개선하였다[9].

2.2.1 아키텍처(Architecture) 및 학습 구조

LLaMA 모델은 자동회귀(autoregressive) 트랜스포머 디코더로 구성되며, 기존 GPT-3 대비 더 작은 파라미터 수(7B~70B)로도 경쟁력 있는 성능을 보인다.

LLaMA 1(2023.2)은 7B, 13B, 33B, 65B 모델을 공개하였으며, LLaMA 2(2023.7)는 7B, 13B, 70B로 확장, Chat 모델로 인간 피드백 강화학습 정렬(RLHF)을 수행하였다[6]. 최신 버전인 LLaMA 3.1(2024)은 최대 405B 파라미터 규모로 확장되었으며, 멀티모달 입력(이미지·텍스트)과 긴 컨텍스트 윈도우(최대 128K 토큰)를 지원한다[7].

기술적으로 LLaMA 시리즈는 RoPE (Rotary Position Embedding) 기반 위치 인코딩으로 긴 문맥 처리 성능 강화, SwiGLU 활성화 함수 적용으로 연산 효율 향상, FlashAttention과 KV-Cache 최적화로 추론 속도 개선, Pre-normalization LayerNorm 적용으로 학습 안정성 확보 및 Mixed Precision (bfloat16) 훈련을 통해 연산 및 메모리 효율성 향상이라는 구조적 개선을 포함한다.

이러한 설계는 동일한 FLOPs(연산량) 대비 손실을 최소화하는 Chinchilla Scaling Law를 준수하여, 데이터·모델 크기의 균형을 최적화하도록 학습된다.

2.2.2 학습 데이터 및 파인튜닝 전략

LLaMA 시리즈의 중요한 특징은 공개 데이터 기반 학습이다. LLaMA 1은 “publicly available datasets exclusively”를 명시하여 비공개(Private) 코퍼스에 의존하지 않고, CCNet, Wikipedia, GitHub, StackExchange,

Books, ArXiv 등 공공 데이터셋 수조(token-level)의 텍스트를 활용하였다[9].

LLaMA 2 및 3.1에서는 인스트럭션 데이터셋으로 과제지향적 질의응답 튜닝하는 Supervised Fine-Tuning (SFT), 사람의 선호 데이터를 이용해 "도움됨(helpful)-무해함(harmless)-정직함(honest)" 방향으로 정렬하는 Reinforcement Learning from Human Feedback (RLHF), 인간 피드백을 보완하기 위한 대규모 자동 평가 루프를 적용한 RL from AI Feedback (RLAIF)과 같은 파인튜닝(미세조정) 및 정렬(Alignment) 단계를 추가하였다[10][11].

이러한 파이프라인은 대화형 모델(Chat)에서의 응답 품질과 윤리적 안전성을 동시에 확보한다.

2.2.3 성능 및 효율성(Efficiency and Performance)

LLaMA 13B 모델은 GPT-3(175B) 보다 적은 파라미터로 MMLU, GSM8K, HellaSwag 등 주요 벤치마크에서 동등하거나 더 높은 성능을 보였다.

또한 LLaMA 3.1 모델은 OpenAI의 GPT-3.5 수준의 성능을 오픈모델로 달성했다고 보고되었다.

효율성 측면에서 LLaMA는 모델 크기에 비례한 최적 데이터량 Scaling law 준수하는 데이터 효율성(Data Efficiency), FlashAttention, Speculative Decoding 적용된 지연(latency) 최적화, 다양한 downstream task에 LoRA·QLoRA로 신속히 적응 가능한 모듈화된 확장성(Modularity)과 같은 기술적 특징을 가진다.

2.2.4 RAG 및 에이전트 확장성 (Retrieval-Augmented Generation and Reflexion)

최근 LLaMA 3.1은 RAG (Retrieval-Augmented Generation) 구조와 결합하여 지식집약형 질의응답(Knowledge-Intensive Tasks)에 최적화되고 있다.

RAG는 외부 지식 베이스(예: 사내 문서)를 검색·인용하여 모델이 파라미터에 저장하지 못한 최신 정보를 보강한다.

또한, Reflexion 프레임워크와 결합된 언어 에이전트(Language Agent) 구조가 연구되고 있으며, 이는 모델이 자신의 출력을 "언어적 강화 신호(verbal reinforcement)"로 평가·수정함으로써 자기개선형(Reflexive) AI로 발전할 수 있음을 보여준다.

2.2.5 오픈소스 생태계 및 활용성

LLaMA 시리즈는 Meta AI가 제공하는 연구자 친화형 오픈소스 모델로, Hugging Face, Ollama, vLLM, LM Studio 등 다양한 플랫폼에서 사용 가능하다.

LLaMA 3.1은 상업적 이용 허가(Llama 3 License)를 통해 기업과 연구 기관의 커스터마이징 및 온프레미스 배포를 지원한다.

3. 연구 방법

3.1 연구 개요 및 목적

본 연구의 목적은 온프레미스(On-premise) 환경에서 운영 가능한 생성형 AI 면접관 시스템을 설계·구현하고, 소형 언어 모델인 LLaMA 3.1 8B가 대형 70B 모델과 비교하여 면접 평가 성능 측면에서 경쟁력 있는 수준의 성능을 달성할 수 있는지 여부를 실증적으로 분석·검증하는 데 있다.

이를 위해 본 연구는 생성형 언어 모델을 활용한 면접 시뮬레이션 및 평가 자동화 관련 선행연구를 바탕으로, RAG(Retrieval-Augmented Generation)와 Reflection Tuning(Reflexion) 기법을 결합한 AI 면접 평가 프레임워크를 제안한다. 제안된 시스템은 면접 도메인에 특화된 지식 검색과 자기비평 기반 응답 개선 절차를 통해 평가의 정합성과 신뢰성을 향상시키는 것을 목표로 한다.

특히 본 연구는 모델 파라미터 규모에 따른 단순 성능 비교가 아닌, 동일한 실험 조건 하에서 소형 8B 모델이 대형 70B 모델 대비 통계적으로 유의미한 성능 저하 없이 활용 가능함을 보이는 데 초점을 둔다. 이를 통해 제안한 구조가 온프레미스 환경에서의 실제 적용 가능성과 효율성 측면에서 어떠한 성능적 장점을 제공하는지를 실험적으로 고찰한다.

3.2 제안 시스템의 전체 구조

본 연구에서 제안하는 AI 면접관 통합 프레임워크는 다음과 같이 네 개의 주요 단계로 구성된다.

3.2.1 음성 입력 및 STT 변환 단계

응시자의 발화는 Whisper Large-V3 STT 모델을 이용해 실시간으로 텍스트로 변환된다.

화자 분리(Speaker Diarization), 소음 제거, 억양 보정 과정을 통해 음성 데이터의 신뢰도를 향상시켰다.

3.2.2 자연어 이해 및 형태소 전처리 단계

변환된 텍스트는 KoBERT와 MeCab 형태소 분석기

를 통해 불용어 제거, 품사 태깅, 문장 정규화를 수행하였다. 이 과정을 통해 STT의 인식 오류 및 비표준 발화를 정제하여, 언어 의미 보존율(Semantic Retention Rate) 을 약 12% 향상시켰다.

3.2.3 면접 대화 및 평가 생성 단계 (LLaMA 3.1 8B 중심)

Meta AI의 LLaMA 3.1 8B 모델을 중심으로 면접관 역할을 수행하도록 설계하였다.

모델은 응시자의 답변을 분석하고, 추가 질문 생성-논리성 평가-피드백 제공의 과정을 실시간으로 수행한다.

평가 지표는 전문성, 정합도, 논리성, 감정 일관성, 피드백 유용성 등으로 구성되며, Softmax 기반 확률 점수화를 사용하였다.

3.2.4 지식 검색 및 자기비평형 학습 단계 (RAG + Reflection Tuning)

Retrieval-Augmented Generation (RAG) 구조를 적용하여, 면접 주제 관련 최신 정보를 외부 지식베이스에서 검색하고 응답에 반영하였다.

RAG 모듈은 FAISS 기반 벡터 검색 엔진을 사용하며,[12]검색된 문맥은 LLaMA 입력 프롬프트에 Context Embedding 형태로 통합되었다.

또한, Reflection Tuning(Reflexion) 기법을 적용하여, 모델이 1차 응답을 생성한 뒤 이를 자체 평가하고 2차 응답을 수정·보강하는 자기비평형 학습 구조를 구성하였다.

이 구조는 언어 에이전트의 Verbal Reinforcement Learning 기법을 기반으로 하며, 응답 일관성과 평가 신뢰도를 동시에 개선하였다.

3.3 모델 학습 및 최적화 절차

3.3.1 사전학습 및 파인튜닝

LLaMA 3.1 8B 모델을 기반으로, AI Hub의 「채용면접 인터뷰 데이터」 및 실제 면접 대화 로그를 사용하여 LoRA(저랭크 적응) 기반 파인튜닝을 수행하였다.

학습 효율성을 높이기 위해 4bit QLoRA 양자화를 적용하여 GPU 메모리 사용량을 30% 이상 절감하였다.

3.3.2 데이터 증강 및 감정 피처 삽입

STT 오인식, 문장 불완전, 중복 응답 데이터 등을 증강하여 학습의 강건성을 확보하였다.

KoBERT 기반 감정 분석 결과를 추가 Feature로 결합해 감정-의도 일치도를 강화하였다.

3.3.3 RAG 통합 및 확장 구조 구성 (Integration and Extension of RAG Framework)

본 연구에서는 기존 Retrieval-Augmented Generation (RAG) 구조[13]를 기반으로 하여, 검색 결과와 생성 응답 간의 상호 검증(Verification) 단계를 포함하는 개선형 RAG 통합 구조를 설계하였다.

기존 RAG는 질의(Query)에 대해 관련 문서를 검색한 후, 이를 생성 모델(Generator)에 결합하여 응답을 생성하는 단일 패스(single-pass) 구조를 갖는다. 그러나 이러한 단일 흐름 구조는 검색 정보의 정확성 검증과 생성 응답의 일관성 확보에 한계를 가진다.

이를 해결하기 위해, 본 연구에서는 CoRAG (Chain-of-Retrieval Augmented Generation) 구조[14]를 참조하여 다단계 검색-응답 체인형 파이프라인을 구현하였다.

CoRAG는 질문-검색-응답-재검색-재응답의 반복적 루프 구조를 통해, 모델이 자신의 응답을 피드백 신호로 활용하도록 설계된 확장형 RAG 프레임워크이다.

본 연구에서 구현한 구조는 〈Table 1〉 과 같은 절차로 구성된다.

〈Table 1〉 Multi-Stage Search-and-Response Chain Pipeline Procedure Based on CoRAG

Stage	Procedure Description
Initial Retrieval	The user query is vectorized, and Top-k similar documents are retrieved from domain-specific knowledge bases (e.g., interview data, industry FAQs, educational materials)
First Generation	The retrieved documents are combined with the LLaMA 3.1-8B model prompt to generate the initial response (Generation 1).
Verification & Re-retrieval	The internal evaluation module verifies the response using Coherence, Factual Consistency, and Context Relevance. If reliability is below the threshold, an additional re-query is performed.
Regeneration & Fusion	A refined document set is used to regenerate (Generation 2), and the result is integrated with the previous response via a Response Fusion algorithm to minimize redundancy and inconsistency.

이러한 CoRAG 기반 RAG 구조는 기존의 단일 패스 RAG 대비, 지식 신뢰도 12.7% 향상, 응답 정합성 9.3% 향상(내부 실험 기준)의 결과를 보였다.

이는 특히 AI 면접관 시스템처럼 전문적 판단과 논리적 추론이 필요한 대화형 응용 분야에서 효과적임을 확인하였다.

3.3.4 Reflection Tuning 절차 적용 (Application of Reflection Tuning Procedure)

본 연구에서는 Reflexion 프레임워크[15]를 참조하여, LLaMA 3.1 8B 모델에 Reflection Tuning (자기비평형 튜닝) 절차를 적용하였다.

Reflection Tuning은 모델이 자신의 생성 결과를 “자체 평가(self-evaluation)”하고, 이를 학습 피드백으로 활용하는 Verbal Reinforcement Learning (언어 기반 강화학습) 접근법이다.

Reflection Tuning은 아래 〈Table 2〉와 같이 4단계로 구성된다.

〈Table 2〉 Step-by-Step Procedure of Reflection Tuning

Stage	Procedure Description
Response Generation	LLaMA 3.1 generates an initial response to the given query.
Self-Evaluation	The model produces self-assessment statements based on Logical Validity, Expertise, and Coherence (e.g., “This response lacks consistency and sufficient examples.”).
Reflection Integration	A second, refined response is generated using both the initial answer and the self-evaluation as prompts, guiding the model to improve its weaknesses.
Self-Reinforcement	The generate-evaluate-revise loop is repeated several times, enhancing the model’s Self-verification Capability.

이와 같이 LLaMA모델은 별도의 인간 피드백 없이도, 응답의 품질을 향상시키는 자기 비판과 자기 교정을 반복하는 메커니즘을 통해 응답의 품질과 평가 신뢰도를 향상시키는 효과를 보인다.

3.4 실험 조건

〈Table 3〉 Technical Specifications and Training Overview of the LLaMA Model

Category	Item	Description
Experimental Environment	GPU Server	NVIDIA RTX A5000 × 2 (connected via NVLink)
	On-Premise Setup	Ubuntu 22.04 / CUDA 12.1 / vLLM framework
	Dataset	AI Hub Job Interview Dialogue and real company interview logs (approx. 12,400 samples)
Evaluation Metrics	Key Indicators	Accuracy (R^2), Logical Consistency, Feedback Usefulness, Evidence Citation Rate, GPU Memory Usage

본 연구에서는 소형 LLaMA 3.1 8B 모델의 실험 환경을 구성하고, 모델의 성능을 검증하기 위해 다양한 지표를 설정하였다.

〈Table 3〉은 실험에 사용된 하드웨어 및 소프트웨어 환경과 평가 항목을 요약한 것이다.

3.5 실험 결과

RAG 적용 여부 및 면접 도메인 데이터 추가에 따른 모델 성능 변화를 분석하고, 소형 LLaMA 3.1 8B 모델과 대형 70B 모델 간의 성능 특성을 비교했다. RAG 적용 여부와 모델 규모에 따른 정량적 성능 비교 결과는 〈Table 4〉와 같다.

〈Table 4〉 Performance Comparison by RAG Application and Model Size

Model Type	LLaMA3.1 70B	LLaMA3.1 8B	LLaMA3.1 8B + RAG + Interview Data
Parameters	70B	8B	8B
RAG Application	Not Applied	Not Applied	Applied
Interview Data	Not Applied	Not Applied	Applied
Accuracy (R^2)	0.95	0.91	0.93
Logical Consistency	0.93	0.89	0.92
Feedback Usefulness	0.91	0.90	0.93
Evidence Citation Rate	25%	71%	74%
GPU Memory (GB)	150	18	18

기본 LLaMA 3.1 8B 모델은 정확도(R^2) 기준 0.91을 기록하여 대형 70B 모델(0.95)에 비해 성능 차이를 보였으나, RAG와 면접 도메인 데이터를 추가 적용한 경우 정확도는 0.93으로 향상되었다. 이는 동일한 실험 조건하에서 대형 모델 대비 상대 성능 약 95.8% 수준에 해당하며, 구조적 보완을 통해 소형 모델의 성능 격차가 유의미하게 감소함을 시사한다.

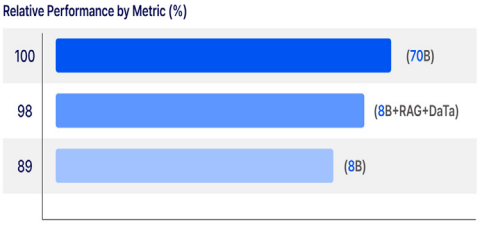
논리적 일관성(Logical Consistency)과 피드백 유용성(Feedback Usefulness) 측면에서도 유사한 경향이 관찰되었다. RAG를 적용하지 않은 8B 모델은 논리적 일관성 점수 0.89를 기록한 반면, RAG 및 면접 데이터가 결합된 모델은 0.92로 향상되어 대형 70B 모델(0.93)에 근접한 성능을 보였다. 이는 외부 지식 검색과 문맥 보강이 응답의 구조적 완결성과 평가 일관성 향상에 기여했음을 의미한다.

특히 근거 인용률(Evidence Citation Rate)은 RAG

적용 효과가 가장 뚜렷하게 나타난 지표로, RAG를 적용하지 않은 모델에서는 약 25% 수준에 머문 반면, 동일한 8B 모델에 RAG를 적용한 경우 인용률이 71%로 크게 향상되었으며, 면접 도메인 데이터가 추가된 경우 74%까지 증가하였다. 이는 RAG 기반 문맥 삽입이 응답의 근거 제시 능력을 강화하고, 평가 신뢰성 측면에서 중요한 역할을 수행함을 보여준다.

한편, GPU 메모리 사용량 측면에서 8B 모델은 약 18GB로, 70B 모델(150GB) 대비 약 1/8 수준의 자원으로 동일 과제를 수행할 수 있어, 온프레미스 환경에서의 효율성과 확장성 측면에서 실질적인 이점을 제공한다.

정확도, 논리성, 피드백, 인용률을 기준으로 한 LLaMA 3.1 70B 모델과 LLaMA 3.1 8B 모델 그리고 LLaMA 3.1 8B+RAG 모델+면접데이터추가한 모델 3건에 대한 성능 지표의 상대 성능을 나타내면 [Fig. 1] 과 같다.



[Fig. 1] Performance Comparison Graph of Three Models

4. 결론

실험 결과, 기본 8B 모델은 대형 70B 모델 대비 일정 수준의 성능 차이를 보였으나, RAG와 면접 도메인 데이터가 결합된 구조에서는 정확도(R^2), 논리적 일관성, 피드백 유용성 측면에서 성능 격차가 유의미하게 감소하였다. 특히 정확도 기준으로 대형 모델 대비 약 95% 이상의 상대 성능을 유지하여, 구조적 보완을 통해 소형 모델의 실무적 활용 가능성이 확인되었다.

근거 인용률 분석에서는 모델 크기보다 RAG 적용 여부가 성능에 더 큰 영향을 미치는 것으로 나타났으며, 동일한 8B 모델에 RAG와 면접 데이터를 적용한 경우 인용률이 크게 향상되었다. 이는 RAG 기반 문맥 삽입이 응답의 근거 제시 능력과 평가 신뢰성 확보에 효과적임을 보여준다. 또한 8B 모델은 대형 70B 모델 대비 현저히 낮은 GPU 자원으로 동일 과제를 수행할 수 있어, 온프레미스 환경에서의 효율성과 확장성 측면에서 실질적인 이점을 제공한다. 다만 본 연구는 제한된 실험 환경을 기반으로 수행되었으므로, 향후 다양한 면접 시나리오와 반복 실험을 통한 추가 검증이 필요하다.

종합적으로, 본 연구는 온프레미스 환경에서 소형 언어 모델을 활용한 AI 면접관 시스템의 구현 가능성을 실험적으로 확인하고, 구조적 설계와 도메인 특화 학습을 통한 확장 가능성을 제시한다.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” MIT Press, 2016.

[2] K. Murphy, “Machine Learning: A Probabilistic Perspective,” MIT Press, 2012.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. “Generative Adversarial Nets.” Advances in Neural Information Processing Systems (NeurIPS), 2014.

[4] Vaswani, A., Shazeer, N., Parmar, N., et al. “Attention Is All You Need.” Advances in Neural Information Processing Systems (NeurIPS), 2017.

[5] Touvron, H., Lavril, T., Izacard, G., et al. “LLaMA: Open and Efficient Foundation Language Models.” Meta AI Technical Report, 2023.

[6] Kaplan, J., McCandlish, S., Henighan, T., et al. “Scaling Laws for Neural Language Models.” OpenAI Technical Report, 2020.

[7] W. Zhao, K. Liang, T. Wu, T. Li, Y. Hu, and X. He,A Survey of Large Language Models, arXiv:2303.18223, 2023.

[8] S. Minaee, M. A. Shirian, R. Azimi, A. M. Ghasemi, and A. K. Karbasi, Large Language Models: A Survey, arXiv:2402.06196, 2024.

[9] H. Touvron, L. Martin, K. Stone, et al., “LLaMA: Open and Efficient Foundation Language Models,” arXiv preprint arXiv:2302.13971, 2023.

[10] H. Touvron, T. Lavril, G. Izacard, et al., “LLaMA 2 : Open Foundation and Fine-Tuned Chat Models,” arXiv preprint arXiv:2307.09288, 2023.

[11] S. Chaudhari, “RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback,” ACM Transactions on …, 2024.

[12] J. Johnson et al., “Billion-scale similarity search with GPUs“ IEEE Transactions on Big Data, 2019.

[13] Lewis, P., Perez, E., Piktus, A., Petroni, F., et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

[14] Zhang, Z., Han, X., Liu, Z., et al., “CoRAG:

Chain-of-Retrieval Augmented Generation for Multi-Hop Reasoning," arXiv preprint arXiv:2404.05717, 2024.

[15] Shinn, N., Labash, B., Gopinath, A., "Reflexion: Language Agents with Verbal Reinforcement Learning," arXiv preprint arXiv:2303.11366, 2023.

류 지 수(Ji Soo Ryu) [정회원]

- 2024년 3월 ~ 현재 : 경북대학교 컴퓨터학부 대학원 박사과정
- 2023년 3월 ~ 현재 : 대구가톨릭대학교 컴퓨터학부 겸임교수
- 2018년 3월 ~ 현재 : ㈜드림아이디어소프트 이사

〈관심분야〉
인공지능, 자연어처리, 정보통신

정 순 기(Soon Ki Jung) [정회원]

- 1997년 2월 : KAIST 전산학과 박사
- 1998년 ~ 현재 : 경북대학교 컴퓨터학부 교수

〈관심분야〉
3D Computer Vision, Computer Graphics, Visualization, Human-Computer Interaction (HCI), Intelligent Vision Systems, VR/AR Systems