이력서 연동 RAG 파이프라인과 LLM-as-a-Judge를 결합한 IT 직무 기술 면접 자동 평가 시스템 설계

유재민*, 신은선*, 허혜림*, 김무영*, 유한솔*, 이광형*
*서일대학교 Ai게임융합학과
dreamace@seoil.ac.kr

Design of an automatic IT job technical interview evaluation system combining resume-linked RAG pipeline and LLM-as-a-Judge

Yu Jae-Min*, Shin Eun-Sun*, Heo Hye-Rim*, Kim Moo-Young, Yu Han-Sol*,
Lee Kwang-Hyoung*
*Dept. of AI-Game Convergence, Seoil University

요 약

본 연구는 IT 직군 채용 시장에서 주니어 구직자가 기술 면접에서 겪는 어려움을 해결하기 위해, 검색 증강 생성(RAG) 파이프라인과 LLM-as-a-Judge 평가 방식을 결합한 IT 직무 특화 기술 면접 자동 평가 시스템 passit을 설계하였다. 제안 시스템은 지원자의 이력서 및 GitHub 데이터를 연동하여 개인화된 기술 질문을 생성한다. 또한 CRAG 기반으로 검증된 FAISS 벡터 데이터베이스를 활용해 IT 직무 지식을 검색하고, 이를 평가 컨텍스트로 사용하여 면접 답변을 4개 차원에서 자동 평가한다.1대3 압박 면접 및 다대다 면접 시나리오를 통한 실전 대비 환경을 제공하며, SSE 기반 스트리밍을 통해 실시간 응답성을 확보하고, 평가 후 취약점 분석 및 모범 답안을 포함한 개인화 리포트를 제공하도록 설계하였다. 본 연구는 실증 실험 이전의 시스템 설계 단계 연구로, RAG 기반 도메인 특화 지식 활용, 이력서 연동 개인화 질문 생성, 다면 평가 루브릭, 윤리적 설계 요소를 통합함으로써 기존 범용 LLM 면접 시스템의 한계를 보완할 수 있는 설계 방향을 제시한다는 점에서 의의를 가진다.

1. 서론

정보기술(IT) 산업의 급격한 성장과 함께 소프트웨어 개발자 수요가 지속적으로 증가하고 있으며, IT 직군 채용 과정에서 기술 면접의 중요성이 더욱 부각되고 있다[1]. CS 이론, 알고리즘, 데이터베이스, 네트워크 등 하드 스킬의 깊이 있는 검증을 요구하는 기술 면접은 포트폴리오나 코딩 테스트와 별개로 실무 역량을 종합적으로 평가하는 핵심 전형 단계로 자리 잡고 있다. 그러나 대학생 및 부트캠프 수료생과 같은 주니어 구직자들은 기술 면접에서 반복적으로 어려움을 겪고 있다. 이들은 포트폴리오와 코딩 역량을 갖추고 있음에도 불구하고 기술 면접에서 어려움을 겪는다. 특히 꼬리물기 압박 질문에 체계적으로 대응하는 기술 커뮤니케이션 능력이 부족하여 실전 면접에서 탈락하는 경우가 많다.[2]. 이는 기존 AI 면접 솔루션이 인성 및 기본 성향 파악 위주로 설계되어 있어, IT 직무의 하드 스킬 검증에 특화된 기능이 부재하기 때문이다. 이러한 한계를 해결하기 위해 최근 대규모 언어 모델(LLM)을 활용한 면접 지원 시스템 연구가 주목받고 있다. 천재성 외(2024)는 ChatGPT 기반 모의 면접 시스템을 구현하여 실시간 피드백의 유용성을 실증하였으나[3], 범용 LLM만을 사용하여 IT 직무 특화 지식이 부재하고 환각(Hallucination) 현상으로 인한 평가 오류 문제가 남아 있다. 표 1은 기존 시스템과 본 연구의 차별점을 나타낸다.

[표 1] 기존 면접 시스템과 본 연구 비교

구분	일반 AI면접	ChatGPT기반 (천재성 외,2024)	본 연구(passit)
도메인 특화 지식	없음	없음	IT 직무 RAG 지식베이스
꼬리물기 질문	없음	제한적	동적 생성
자동 채점	없음	없음	LLM-as-a-Judge 다차원
이력서 연동	없음	없음	이력서·깃허브 연동
사용 LLM	전용 모델	GPT-3.5/4	GPT-4o
환각 억제	없음	없음	RAG 컨텍스트 주입
윤리적 설계	없음	없음	개인정보 보호·공정성 반영

이를 해결하기 위해 본 연구는 검색 증강 생성(RAG) 파이프라인과 LLM-as-a-Judge 평가 방식을 결합한 IT 직무 특화 기술 면접 자동 평가 시스템 'passit'의 아키텍처를 제안한다. 본 설계

단계 연구의 학술적 기여는 다음과 같다.

첫째, 본 연구는 CRAG 기반의 신뢰도 검증 메커니즘을 적용한 도메인 특화 지식베이스를 구축하였다. 또한 SSE 스트리밍 기술을 적용하여 실시간 응답성을 확보하였다. 이를 통해 환각을 억제하면서도 안정적인 평가 체계를 설계하였다.

둘째, 시맨틱 임베딩을 활용한 맥락 인식형 질문 생성, 다차원 평가 루브릭, 다대다·압박 면접 시나리오를 결합하여 개인화된 실전 환경을 구체화하였다.

셋째, 알고리즘 공정성 및 개인정보 보호 등의 윤리적 설계 요소를 프레임워크에 체계적으로 반영하였다.

2. 이론적 배경

2.1 검색 증강 생성(RAG)과 품질 검증 아키텍처

RAG는 외부 지식베이스를 활용해 LLM의 환각을 억제하는 기법이다[4]. 그러나 문서 무비판적 수용의 한계를 지닌 초기 Naive RAG를 극복하고자, 최근 CRAG[10], Self-RAG[11] 등 품질 검증이 강화된 Modular RAG[5]로 발전하고 있다. IT 기술 면접은 정답의 경계가 명확하기 때문에 높은 사실 정확성이 요구된다. 이에 본 시스템은 검색 문서의 신뢰도를 평가하고 교정하는 CRAG 메커니즘을 제어 레이어로 적용하였다. 이를 통해 평가의 타당성을 확보한다.

2.2 LLM 기반 자동 평가(LLM-as-a-Judge)

LLM-as-a-Judge는 특정 기준에 따라 LLM이 응답을 평가하는 방식으로, 다차원 루브릭 적용 시 GPT-4 계열은 인간 전문가와 80% 이상의 높은 일치율을 보여 객관적 평가 수단으로서의 학술적 근거가 입증되었다[6, 7]. 고유 한계인 위치 편향 및 자기 편향[6]을 완화하기 위해, 본 시스템은 세션마다 루브릭 제시 순서를 무작위화하고 단계별 추론(Chain-of-Thought)을 통해 평가 근거를 명시적으로 출력하도록 설계하여 공정성을 높인다.

2.3 AI 기반 기술 면접 시스템

최근 동적 상호작용 및 데이터 선별 등 LLM 기반 채용 지원 연구[8, 9]가 활발하나, 범용 LLM에 의존하거나 1:1 단일 면접 유형에 국한되는 한계가 있다. 본 연구는 IT 특화 지식베이스, CRAG 품질 검증, 압박 및 다대다 면접 시나리오를 포괄하는 다차원 자동 평가 시스템을 제안하여 기존 연구의 공백을 보완한다.

3. passit 시스템 설계

3.1 시스템 개요

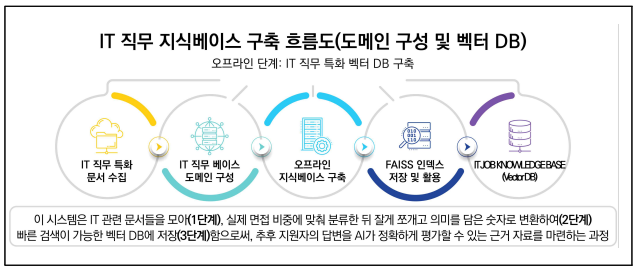
passit은 IT 직무 취업을 준비하는 주니어 구직자가 기술 면접에 체계적으로 대비할 수 있도록 지원하는 AI 기반 초개인화 기술 면접 플랫폼이다. 지원자의 이력서 및 깃허브 링크를 시맨틱 임베딩 방식으로 연동하여 개인의 기술 맥락에 맞는 면접 질문을 생성하고, CRAG 품질 검증 레이어가 적용된 RAG 기반 IT 직무 지식베이스를 활용하여 면접 답변을 다차원으로 자동 평가한다.

본 시스템의 평가 엔진으로는 GPT-4o를 사용하며, 임베딩 모델로는 OpenAI text-embedding-3-small을 채택하였다. GPT-4o를 선정한 이유는 Zheng 외(2023)의 연구에서 GPT-4 계열이 LLM-as-a-Judge 평가에서 인간 평가자와 가장 높은 일치율을 보였기 때문이며, 비용 대비 성능 측면에서도 우수하다는 점을 고려하였다.

본 시스템은 오프라인 지식베이스 구축 단계와 온라인 면접 평가 단계의 두 단계로 구성된다. 오프라인 단계에서는 IT 직무 특화 문서를 수집 및 품질 검증하여 벡터 데이터베이스를 구축하고, 온라인 단계에서는 지원자와 실시간으로 상호작용하며 질문 생성 및 답변 평가를 수행한다.

3.2 IT 직무 지식베이스 구축

지식베이스는 실제 IT 기업 기술 면접 후기 340건을 분석하여 자료구조·알고리즘(30%), 데이터베이스(20%) 등 6개 핵심 도메인 비중을 실증적으로 설정하였다. 데이터는 공개 CS 교재 및 기술 블로그 등에서 수집하며, 향후 현업 면접관 설문을 통해 비중의 대표성을 추가 검증할 계획이다. 특히, 무분별한 외부 데이터 삽입으로 인한 환각(Hallucination) 현상을 방지하기 위해 오프라인 구축 단계에 CRAG 메커니즘을 차용한 문서 신뢰도 자체 평가 레이어를 새롭게 도입하였다. 수집된 문서는 실제 면접의 평균 답변 길이를 고려하여 LangChain을 통해 청크 크기 512토큰, 오버랩 64토큰으로 우선 분할된다. 이후 각 문서 청크는 인덱싱 전 GPT-4o를 통해 사실적 정확성을 평가받으며, 신뢰 점수가 임계값(0.7) 미만인 데이터는 선제적으로 배제되어 시스템의 기술 정확도(40%) 평가에 대한 신뢰 기반을 확고히 한다. 최종적으로 신뢰도 검증을 통과한 청크만 OpenAI 임베딩 모델(text-embedding-3-small)을 거쳐 FAISS 인덱스에 저장되며, 실시간 평가 시 코사인 유사도 기반 Top-K(K=5) 검색을 통해 고품질의 평가 컨텍스트로 제공된다.



[그림 1] IT 직무 지식베이스 구축 흐름도

3.3 이력서 연동 맞춤형 질문 생성 파이프라인

기존 이력서 파서의 단순 키워드 매칭(Keyword Matching)이 지닌 문맥 손실 한계를 극복하기 위해, 본 시스템은 시맨틱 임베딩 기반의 맥락 인식형(Context- aware) 파이프라인을 설계하였다. 이력서 및 깃허브 연동 시 단순 기술명 추출을 넘어, '프로젝트 문제 해결 과정이 담긴 문단 전체'를 text-embedding-3-small 모델로 벡터화한다. 이를 지식베이스와 대조해 의미론적으로 가장 적합한 개념을 식별하여 개인화된 초기 질문을 생성함으로써, 표면적으로 명시되지 않은 지원자의 숨겨진 역량까지 발굴한다. 더불어 면접 진행 중에는 Chain-of-Thought(CoT) 프롬프트를 활용하여 지원자 답변의 논리적 허점과 기술적 정확성을 파고드는 심층 꼬리물기 질문을 동적으로 생성하고 그 난이도를 정교하게 제어한다.

[표 2] 꼬리물기 질문 생성 프롬프트 구성 요소

프롬포트 구성 요소	내용
시스템 역할 정의	“당신은 IT 기술 면접관입니다.”
참조 지식 컨텍스트	FAISS Top-5 검색 결과 주입({retrieved_context})
이력서 맥락	지원자 이력서 임베딩 기반 주요 경험 요약({resume_context})
이전 답변	지원자 직전 답변({candidate_answer})
생성 지시	논리적 허점·예외 케이스·실무 적용 문제를 짚는 꼬리물기 질문 1개 생성
출력 형식	질문 텍스트 단독 출력 (JSON 불필요)

3.4 면접 유형별 시나리오 구성

본 시스템의 1대3 압박 면접 시나리오는 3명의 AI 면접관이 각각 독립적인 RAG 검색과 CoT 추론을 병렬로 수행하므로, 답변 생성까지의 대기 시간이 길어질 위험이 있다. 대화형 AI 시스템에서 가장 중요한 사용자 경험(UX) 지표 중 하나는 TTFT(Time to First Token)로, 사용자가 질문을 보낸 후 첫 번째 토큰이 화면에 출력되기까지 걸리는 시간을 의미한다. 프롬프트 길이가 증가하면 Prefill 단계의 처리 시간이 증가한다. 이로 인해 TTFT(Time to First Token)가 급격히 늘어난다. 이러한 지연은 실시간 면접 환경에서 대화 흐름을 단절시키고, 사용자 몰입도를 저하시킨다.

이를 해결하기 위해 본 시스템은 SSE(Server-Sent Events) 기반 스트리밍 방식을 채택하도록 설계하였다. 스트리밍 방식은 LLM이 첫 번째 토큰을 생성하는 즉시 클라이언트 화면에 순차적으로 출력함으로써, 전체 응답이 완성되기까지의 대기 시간을 시각적으로 상쇄한다. 특히 1대3 면접의 경우 3명의 AI 면접관 인스턴스를 비동기(Async) 방식으로 병렬 실행하되, 각각의 스트림을 독립된 UI 컴포넌트에 바인딩함으로써 면접관별 응답이 순차적으로 노출되는 효과를 구현한다. 표 3은 면접 유형별 AI 운용 방식과 스트리밍 처리 전략을 나타낸다.

[표 3] 면접 유형별 구성 및 AI 운용 방식

면접 유형	AI 구성	주요 특징	학습 목적
1대1 꼬리물기	AI 면접관(1인)	기술 개념 심층 압박 질문 동적 생성	핵심 기술 설명 능력 강화
1대3 압박 면접	AI 면접관(3인) 기술/심층/압박 역할 부여	다각도 기술 압박 및 허점 집중 질문	복합 질문 대응력 향상
다대다 면접	AI 면접관 + 가상 지원자	타 지원자 답변에 대한 돌발 질문 포함	비교 평가 상황 적응력 강화

4. LLM 기반 자동 평가 모듈 설계

4.1 평가 루브릭 설계

본 시스템의 자동 평가는 기술 면접에서 요구되는 역량을 반영한 4개 차원의 루브릭으로 구성된다. 기술 정확도(40%)는 CRAG로 품질이 검증된 RAG 검색 지식과 지원자 답변의 일치도 및 사실 오류 여부를 평가하며, 가장 높은 비중을 차지한다. 논리 구조(30%)는 개념 설명의 체계성과 인과관계의 명확성을, 심층도(20%)는 개념의 확장 및 실무 연결 능력을 평가한다. 커뮤니케이션(10%)은 설명의 명확성과 전달력을 평가한다. 가중치 배분은 IT 기술 면접에서 기술적 사실 검증이 가장 우선시된다는 점을 반영하였으며, 향후 현업 면접관 설문을 통해 가중치 타당성을 실증적으로 검증할 계획이다. 본 시스템의 루브릭 가중치는 실무 역량 평가에서 '하드 스킬(Hard Skill)'의 비중이 가장 높아야 한다는 기존 기술 면접 분석 결과에 기반한다. 특히 '기술 정확도'에 40%의 최고 가중치를 부여한 것은, IT 직무의 특성상 오개념 전달이 실무 협업에 미치는 부정적 영향을 최소화하기 위함이다.

[표 4] 자동 평가 루브릭

평가 차원	가중치	평가 기준
기술 정확도	40%	RAG 검색 정답 지식과의 일치도, 사실 오류 및 환각 여부
논리 구조	30%	개념 설명의 체계성, 인과관계 명확성
심층도	20%	개념의 확장 및 응용 가능성, 실무 연결 능력
커뮤니케이션	10%	설명의 명확성과 전달력

4.2 평가 프롬프트 및 출력 형식

평가 프롬프트는 [시스템 역할 정의] → [정답 참조 지식베이스 컨텍스트] → [평가 루브릭] → [면접 질문] → [지원자 답변] → [JSON 출력 형식]의 순서로 구성된다. Chain-of-Thought 방식으로 평가 근거를 단계적으로 생성하도록 유도하여 평가 투명성을 확보한다. LLM-as-a-Judge의 위치 편향을 완화하기 위해 평가 루브릭 항목의 제시 순서를 세션마다 무작위화하도록 설계하였다. 출력 형식은 {"scores": {"technical_accuracy": 0~100, "logic": 0~100, "depth": 0~100, "communication": 0~100},

"reasoning": "평가 근거", "weak_points": ["취약 개념 목록"], "model_answer": "모범 답안"}의 JSON 구조를 사용하며, 이를 기반으로 취약점 분석 및 모범 답안을 포함한 개인화 리포트를 자동 생성한다.

[그림 2] 평가 흐름 프롬포트

5. 기대효과

본 연구에서 제안한 passit 시스템은 IT 직무 주니어 구직자의 능동적 면접 대비를 지원하며, 다음 세 가지 측면에서 기여가 기대된다.

첫째, 기술적 신뢰성 및 UX 향상이다. CRAG 기반 문서 검증과 RAG 지식베이스를 결합하여 범용 LLM의 환각 현상을 이중으로 억제하며[4], SSE 스트리밍을 도입해 복잡한 추론 과정(CoT)에서 발생하는 응답 지연(TTFT)을 상쇄함으로써 실전과 유사한 대화 몰입도를 보장한다.

둘째, 초개인화된 실전 면접 학습이다. 시맨틱 임베딩 기반의 문맥 중심 이력서 분석으로 지원자의 숨겨진 역량을 파악하고, 동적인 꼬리물기 질문과 1대3 압박 및 다대다 면접 시나리오를 통해 실전 적응력을 극대화한다. 또한, 세밀한 다차원 평가 리포트와 모범 답안을 제공하여 자기주도적 역량 개선을 유도한다.

셋째, 산업적 확장성이다. 본 시스템은 구직자의 학습 도구를 넘어 B2B 채용 매칭 솔루션으로의 연계가 가능하여, 기업의 면접 자동화 및 IT 채용 시장 전반의 비용 효율성 증대에 기여할 수 있다.

6. 결론

본 연구에서는 IT 주니어 구직자의 기술 면접 대비를 위한 자동 평가 시스템 'passit'의 통합 아키텍처를 설계하였다. 시맨틱 임베딩 기반의 맥락 인식형 질문 생성, CRAG 품질 검증과 RAG를 결합한 환각 억제, LLM-as-a-Judge 기반 다차원 평가, SSE 스트리밍을 통한 대화 지연 상쇄, 실전 압박 시나리오 및 윤리적 가이드라인을 유기적으로 결합하였다. 본 연구는 프로토타입 구현 이전의 설계 단계로 실증 실험은 수행되지 않았으나, 최신 AI 평가 파이프라인의 구체적 타당성과 재현 가능성을 선제적으로 제시했다는 점에서 학술적 의의를 지닌다. 향후 실제 면접 데이터셋을 구축하여 CRAG 및 RAG 적용 여부에 따른 환각 발생률 비교 실험을 수행하고, 현업 개발자 평가와의 일치율(Cohen's Kappa)을 측정하여 제안 시스템의 신뢰성을 실증적으로 검증할 계획이다.

참고문헌

[1] 고용노동부, "2023 직종별 사업체 노동력 조사", 고용노동부, 2023.

[2] 윤채원, 정유철, "의도 요소를 고려한 프롬프트 기반 고품질 AI 면접 데이터 선별 시스템", 한국정보기술학회논문지, 2024.

[3] 천재성, 이재원, 황재원, "ChatGPT를 이용한 컴퓨터 분야 AI 모의 면접 시스템 설계 및 구현", 실천공학교육논문지, 16권 2호, pp. 223-230, 2024.

[4] Lewis, P., et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", Advances in Neural Information Processing Systems (NeurIPS), 2020.

[5] Gao, Y., et al., "Retrieval-Augmented Generation for Large Language Models: A Survey", arXiv:2312.10997, 2024.

[6] Zheng, L., et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena", NeurIPS, 2023.

[7] 정민수, 이정훈, "GPTScore를 이용한 구조화된 표 데이터 기반의 RAG 질의응답 시스템 응답 평가", 한국정보통신학회 종합학술대회 논문집, 28권 2호, 2024.

[8] Ko, Y., et al., "LLM-as-an-Interviewer: Beyond Static Testing Through Dynamic LLM Evaluation", arXiv:2412.10424, 2024.

[9] 정효정, 송주현, 서상훈 외, "RAG 기반 LLM 성능 평가 및 검증을 위한 LangChain 활용 RAGA 방법론 연구", 대한전자공학회 학술대회, 2024.

[10] Shi, W., et al., "CRAG: Corrective Retrieval Augmented Generation", arXiv:2401.15884, 2024.

[11] Asai, A., et al., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection", arXiv:2310.11511, 2023.