Journal of KIIT. Vol. 22, No. 7, pp. 83-94, Jul. 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.7.83

의도 요소를 고려한 프롬프트 기반 고품질 AI 면접 데이터 선별 시스템

윤채원*, 정유철**

Selection System of High-Quality AI Interview Data based on Intent-Oriented Prompts

Chaewon Yoon*, Yuchul Jung**

본 연구는 2024년도 과학기술정보통신부와 한국전파진흥협회의 메타버스랩지원사업에 의한 연구임

요 약

LLM(Large Language Model)의 학습 성능은 사용되는 데이터셋의 품질에 크게 의존한다. 특정 도메인의 모델 학습 시 선별된 데이터의 품질에 따라 모델 성능에 직접적으로 영향을 미치기 때문에 고품질 데이터셋 구축은 모델 성능 향상을 위한 필수적 고려 요소이다. 따라서 본 논문은 AI 면접 도메인의 면접 분석에 사용 할 고품질 데이터 선별을 위한 프롬프트 기반 데이터 정제 시스템을 제안한다. 제안하는 시스템은 면접 답변 내 의도 요소 분석을 통해 정밀한 데이터 선별을 가능하게 하며, 각 의도 요소와 대응되는 개별 프롬프트를 추가하여 데이터 선별 근거의 타당성과 효율성의 개선 정도를 평가하는 것을 목표로 한다. 본 연구는 AI 면접 도메인에서 데이터셋의 신뢰성과 정확성을 향상시키는 새로운 데이터 선별 시스템을 제시함으로써 AI 면접 도메인에 특화된 분석 모델의 학습 효율성을 높이기 위한 토대를 형성한다.

Abstract

The training performance of a Large Language Models(LLM) significantly depend on the quality of the dataset used. As the quality of data within a specific domain directly impacts model efficacy, constructing a high-quality dataset is crucial for model improvement. Therefore, this paper proposes a prompt-based data refinement system for selecting high-quality data for analyzing AI interviews. The proposed system aims to enable precise data screening through the analysis of intent elements in interview answers, and to evaluate the validity of the data screening rationale and the degree of improvement in efficiency by adding individual prompts corresponding to each intent element. By presenting a novel data curation system that improves the reliability and accuracy of datasets in the AI interview domain, this research establishes the basis for improving the learning efficiency of analytical models specific to the AI interview domain.

Keywords

AI interview, generative AI, prompt engineering, intent instruction analysis

* 국립금오공과대학교 컴퓨터공학과 학사과정
- ORCID: http://orcid.org/0009-0007-5871-5366
** 국립금오공과대학교 인공지능공학과 부교수(교신저자)
- ORCID: http://orcid.org/0000-0002-8871-1979

· Received: Apr. 05, 2024, Revised: May 14, 2024, Accepted: May 17, 2024
· Corresponding Author: Yuchul Jung
Dept. of Computer Engineering, Kumoh National Institute of Technology, 61 Daehak-ro (yangho-dong), Gumi, Gyeongbuk, [39177] Korea
Tel.: +82-54-478-7536, Email: enthusia77@gmail.com

Ⅰ. 서 론

인공지능 기술이 급속도로 발전하면서 다양한 대규모 언어 모델(LLM, Large Language Model)이 등장하였으며, LLM의 가파른 성장은 다양한 산업과 학문 분야에 혁신적인 변화를 가져왔다. 여러 기업과 연구 기관들은 이러한 모델을 활용하여 복잡한 언어 기반 문제를 효과적으로 해결하거나 많은 인력과 시간을 요구하던 작업들을 자동화하여 시간과 비용을 절감하고 있다. 특히 여러 기업에서 LLM을 활용하여 챗봇과 같이 기존의 업무를 자동화하거나 새로운 유형의 응용 서비스를 제공하는데 크게 집중하고 있다.

코로나19 전염병 발병 이래로 비대면 서비스 기술의 급격한 발달로 채용시장에서도 비대면 면접 및 채용 프로세스가 크게 확산되었다. 기존 대면 면접 방식에서 벗어나 화상 통화 또는 AI 기반 면접 프로그램을 도입하여 채용 과정의 효율성과 접근성이 극대화되었다. 이러한 변화는 전염병 예방을 위한 사회적 거리 두기 규정을 준수하는 동시에 시공간의 제약을 두지 않기 때문에 채용 과정에서 소모되는 인력과 시간을 줄이며, 기간 내 원하는 시간에 면접을 볼 수 있어 채용 기업과 지원자 모두에게 편의를 제공하고 있다.

AI 면접 도메인에서 AI 기술은 면접관들이 파악하지 못했던 미세한 행동 패턴이나 반응을 포착하고, 축적된 데이터로 정보를 수치화하여 보다 객관적이고 정밀한 분석을 할 수 있도록 한다. 또한 취업 준비생들은 모의 면접을 통해 면접과 유사한 환경에서 실전과 같은 경험을 쌓고, 자신의 강점과 약점을 파악하여 효과적으로 면접을 대비할 수 있다.

따라서 AI 면접 도메인에서 면접 분석 보고서를 생성할 경우 객관적인 시선에서 면접자의 답변과 언어 등을 분석하여 면접자의 역량과 잠재 능력, 태도를 파악하는 것이 중요하다. 이러한 복잡한 분석을 수행하기 위해서는 다양한 상황을 반영하는 면접 질의응답 정보를 포함하며, 면접 분석 규칙을 도출하고 이를 수치화 가능한 기준을 설정할 수 있도록 돕는 특화 데이터셋을 구축하는 것이 필요하다.

따라서 본 논문에서는 AI 면접 분석 보고서 생성 모델 학습에 사용할 데이터셋을 구축하기 위해, 면접 답변 내 의도 정보를 활용하여 면접 질문에 대한 개별 의도 요소를 검출하고, 해당 의도 요소에 부합하는 개별 프롬프트 조합을 통해 답변의 적절성을 판단하는 데이터 정제 시스템을 제안한다.

이에 따라 2절에서는 이전 연구들 중에서 심리나 의도를 분류한 사례들과 프롬프트 엔지니어링 기법을 사용한 데이터셋 구축 사례를 분석한다. 3절에서는 제안하는 프롬프트 기반 데이터 선별 시스템을 소개하며, 데이터셋 전처리 과정, 주요 의도 요소 선정, 데이터셋 내 의도 요소 검출 및 답변 적절성 판단 과정을 설명한다. 4절에서는 제안한 시스템의 판별 정확성과 타당성을 검증하기 위한 실험 과정과 한계를 설명하고, 5절에서는 결론을 서술한다.

Ⅱ. 관련 연구

2.1 심리 및 의도 분류 체계

심리 분석 모델은 대화나 몸짓 속에 내포되어 있는 화자의 감정을 탐지하는 데 사용된다. 이러한 모델들은 초기에 간단한 긍정/부정 분류에서 시작하여 현재는 단순한 이분법적 분류를 넘어 다양한 감정과 심리 상태를 정밀하게 분류하는 연구가 진행되고 있다. 이 과정에서 특정 감정 유형을 식별하기 위한 다양한 감정 분류 모델이 개발되고 활용된다[1][2]. 이와 같은 심리 분석 모델은 고객 서비스 모니터링, 건강 관리 등 경제학과 심리학을 비롯한 여러 분야에서 응용되고 있다[3][4].

예로 온라인 쇼핑몰의 구매자 리뷰 예측 연구[5]에서는 아마존 리뷰 데이터를 활용하여 긍정/부정 분류 모델을 제작했다. 리뷰 평점에서 1점과 2점을 부정적 리뷰로, 5점을 긍정적 리뷰로 분류하여 총 평과 리뷰 텍스트와 함께 학습 데이터셋을 구축한 후, 양방향 LSTM(Bidirectional LSTM) 알고리즘으로 모델을 개발하여 리뷰의 긍정/부정 분류 작업을 수행했다.

또한 영화 리뷰 내 감정을 분석하기 위해 감정 어휘 사전을 이용하여 9가지 감정을 분류하는 모델을 제작한 연구도 있다[6].

이 연구에서는 9가지 감정을 식별하기 위해, 감정 어휘들을 형태 분석한 감정 어휘 사전을 활용하여 사전과 영화평 말뭉치에서 나타난 어휘가 일치할 경우 감정을 주석하여 데이터셋을 구축했다. 이후 구축한 데이터셋으로 KcBert(Korean comments Bert) 모델[7]을 학습시켜 '기쁨, 슬픔, 공포, 분노, 혐오, 놀람, 흥미, 지루함, 통증'의 감정에 대해 분류했다.

심리 분석 모델은 AI 면접 도메인에서도 활용할 수 있다. 그 예로 KoBERT 모델[8]을 기반으로 면접 답변의 의도를 분석한 연구가 있다[9]. 해당 연구에서는 인사혁신처 국가공무원인재개발원과 국가직무능력표준(NCS, National Competency Standards)[10]등의 기준을 토대로 '지식/기술, 태도, 인성, 개인 배경, 기타'의 5가지 주요 의도 요소를 정의하여 면접 답변의 의도를 분류하는 작업을 수행했다.

본 연구에서는 [9]의 연구와 같이 AI 면접 도메인에서 지원자의 답변 의도를 분석하기 위해 특정 의도 요소를 정의한다. 이를 위해 AI-Hub의 채용면접 인터뷰 데이터[11] 가이드라인에 정의된 의도 요소를 바탕으로 5가지 주요 의도 요소로 재정의하고, 각 답변에 내재된 의도를 더 명확하게 분류할 수 있도록 돕는다. 이러한 의도 요소들은 면접 답변의 적절성을 효과적으로 평가하기 위한 자료로 사용된다.

2.2 프롬프트 엔지니어링

프롬프트 엔지니어링은 입력 프롬프트를 최적화하여 LLM에서 원하는 수준의 응답을 유도하는 모델 제어 기술을 말한다. 효과적인 프롬프트 엔지니어링을 통해 추가 학습 없이도 높은 성능과 효과를 얻을 수 있어 간편하면서도 시간과 비용을 절약할 수 있다. 이러한 이점으로 프롬프트 엔지니어링은 zero-shot 및 few-shot 프롬프트를 통해 새로운 작업에 대응하거나 모델의 응답 오류를 최소화하는 추론 과정을 개선하는 등 다양한 응용 분야에서 활용된다[12]-[14].

프롬프트 엔지니어링은 데이터셋 구축 과정에서도 자주 사용된다. 단순히 데이터의 양을 늘리는 것뿐만 아니라 기존 공개 데이터셋의 품질을 개선하거나 학습에 불필요한 데이터를 제외해 효과적인 데이터셋을 구축하는 전략으로 활용된다. 그 예로 제한된 데이터만으로 좋은 성능을 얻기 위해 자동으로 Instruction Tuning 데이터의 효율성을 강화하는 DEITA 모델을 구축한 연구가 있다[15]. DEITA는 복잡성과 품질, 다양성을 종합적으로 고려하여 기존 데이터셋를 정제하고 정렬함으로써 제한된 데이터로도 모델이 사용하는 데이터의 활용을 깊이 있도록 하여 모델의 성능을 향상시켰다.

또한 감정 분석 데이터셋 구축 과정에서 부족한 데이터를 보강하기 위해 데이터 증강 기법으로 프롬프트 엔지니어링을 활용할 수도 있다. 감정 분석 모델 성능 향상을 위한 데이터 증강 연구[16]에서는 'Generative 프롬프트'를 사용해 새로운 주제의 텍스트를 생성함으로써 데이터의 다양성을 확장하고 모델의 창의성을 증진시켰으며, 'Conditional 프롬프트'로 특정 감정 타깃 텍스트 생성을 목표로 하는 자연스러운 문장을 생성할 수 있도록 했다.

프롬프트 엔지니어링은 데이터를 분류하는 데에도 활용할 수 있다. 한 연구에서는 대학 졸업생을 대상으로 채용 공고의 직무적합성을 판단하기 위해 프롬프트 엔지니어링 기법을 사용했다[17]. 이때 채용 공고의 제목과 설명에서 요구하는 직무 경험 또는 기술을 반영하여 적합성 판단 조건을 설정하며, 모델의 이름과 역할, 응답에 대한 긍정적 호응을 포함하여 편향 없는 최적의 프롬프트를 구축함으로써 직무적합성을 효과적으로 판별했다.

본 연구에서는 AI 면접 도메인 내에서 프롬프트 엔지니어링 기법을 사용하여 면접 질의응답 내 5가지 주요 의도 요소를 검출하고, 이를 기반으로 면접 답변의 적절성을 평가하는 데이터 정제 시스템을 제시한다. 면접 데이터 내에서 파악 가능한 의도 요소를 검출하기 위해 5가지 주요 의도 요소의 특성을 반영한 공통 프롬프트를 사용하며, 검출된 의도별 추가 프롬프트를 제공하여 면접 질의응답 데이터의 답변 적절성을 효과적으로 판단할 수 있도록 한다.

III. 제안 시스템

본 절에서는 입력 데이터셋을 정제하여 면접 분석 보고서 생성 모델 학습에 사용할 데이터를 선별하는 과정에서 다양한 질의응답 및 분석 결과를 포함시키기 위한 프롬프트 기반 데이터 정제 시스템에 대해 소개한다. 제안하는 정제 시스템은 크게

'의도 요소 검출'과 '답변의 적절성 판단' 과정으로 나뉜다. 공개 데이터셋에서 질문과 답변, 감정 및 의도 라벨링 데이터를 정제하여 의도 요소를 검출한 후 검출된 의도 요소의 개별 프롬프트를 추가하여 답변의 적절성을 판단하는 시스템이다.

3.1 문제 정의 및 목표

제안하는 프롬프트 기반 데이터 정제 시스템은 그림 1과 같다. 1차로 공개 데이터셋에서 면접 질의응답 정보 내에서 파악할 수 있는 5가지 주요 의도 요소를 검출한다. 이때 검출 정확도를 높이기 위해 공개 데이터셋에서 제공하는 감정 및 의도 라벨링 데이터를 함께 제공하며, 5가지 의도 요소 검출 조건을 담은 프롬프트를 활용하여 의도 요소를 검출한다. 이후 각 의도 요소와 대응되는 개별 프롬프트를 기본 프롬프트에 추가하여 면접 질의응답 정보에 대한 데이터의 적절성을 판단하는 구조이다. 제안하는 정제 시스템은 기존 정제 시스템 대비 면접 질의응답의 적절성 판단 정확도 향상 확인과 더불어 판단 분석 근거의 타당성 개선 수준을 확인하는 것을 목표로 한다.

3.2 데이터셋

3.2.1 활용 데이터셋 소개

본 연구에서는 정제 과정에 활용할 공개 데이터셋으로 한국지능정보사회진흥원(NIA)의 AI-Hub에서 제공하는 '채용면접 인터뷰 데이터' [9]를 채택했다. 한국어 면접 도메인에 특화된 공개 데이터셋은 매우 드물며, 채용면접 인터뷰 데이터는 경력직 및 신입 지원자의 면접 데이터 뿐만 아니라 실무와 인성, 직종별 질문과 답변 등 광범위한 면접 정보를 포함하고 있어 다양한 직종과 질의응답 정보에 효과적으로 대응할 수 있는 면접 답변 분석 보고서 모델을 학습시키기 위한 베이스 데이터로 적합했다.

채용면접 인터뷰 데이터의 직무 및 질문 유형별 데이터 분포는 표 1과 같다. 채용면접 인터뷰 데이터는 7가지 직무 카테고리에 대한 직무별 맞춤 질문을 포함하고 있으며, 실무 및 인성 질문도 골고루 포함하여 훈련 데이터로 활용할 경우, 모델이 다양한 유형의 질의응답 정보에 대해 더욱 효과적으로 대응할 수 있게 한다.

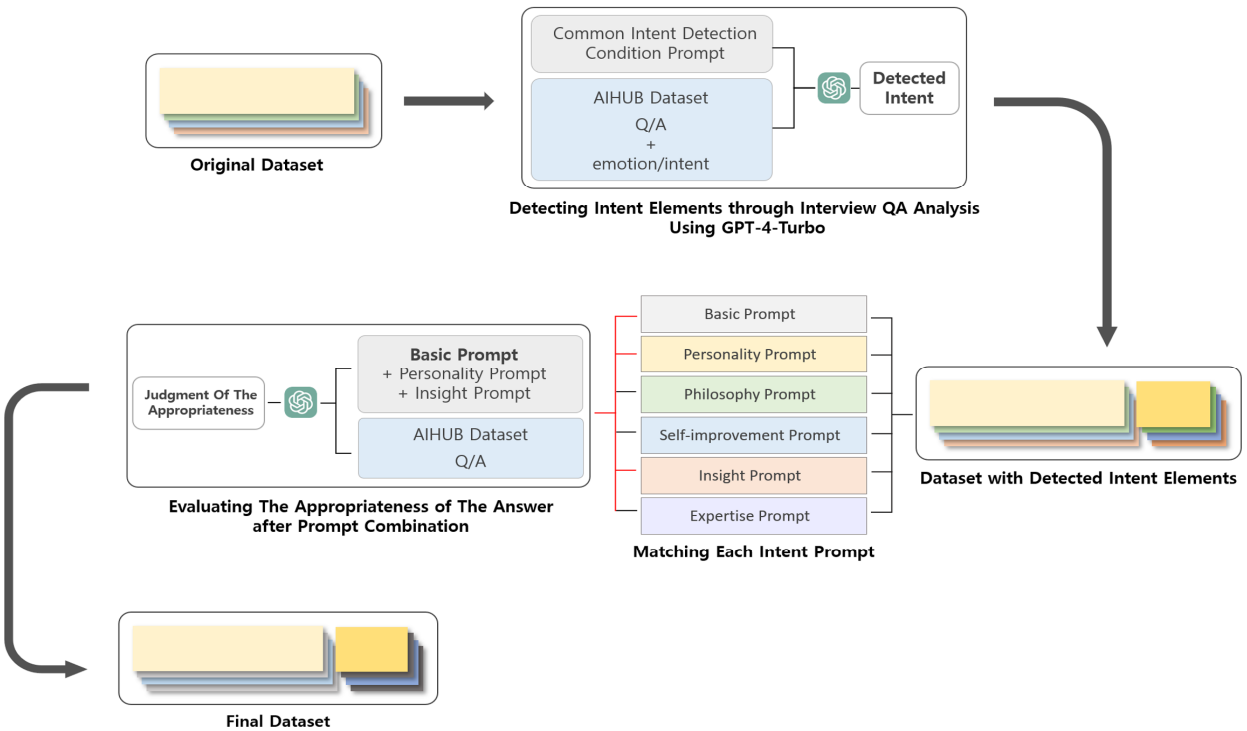

그림 1. 제안하는 프롬프트 기반 데이터 정제 시스템 구조
Fig. 1. Structure of our proposed prompt-based data curation system

표 1. 직무 및 질문 유형별 데이터 분포
Table 1. Data distribution by job and question type

By job type	Art & design	9.69%
	Business & administration	22.57%
	ICT	8.68%
	Production & maintenance	10.22%
	Public service	25.26%
	Construction & manufacturing	25.26%
	Sales & marketing	7.07%
	Total	100%
By question type	Practical skills	74.90%
	Personality	25.10%
	Total	100%

면접 답변 분석 모델은 직종별 맞춤 질문에 대한 분석 외에도 인성 또는 전문성과 같은 핵심 역량을 세밀하게 평가할 수 있는 능력을 갖추어야 한다.

이를 위해, AI-Hub의 채용면접 인터뷰 데이터를 바탕으로 지원자의 역량이 효과적으로 드러나는 답변에 대한 분석 데이터뿐만 아니라, 부적절한 답변에 대한 분석 데이터 모두를 포괄할 수 있는 프롬프트 기반 데이터 정제를 수행한다. 데이터 정제에 사용되는 채용면접 인터뷰 데이터의 요소는 면접 질문 및 답변 데이터이며, 면접 질의응답 데이터의 적합성 분석을 향상시키기 위해 의도 요소 검출 시 개별 감정과 의도 요소 라벨링 데이터를 사용하였다. 이를 통해 다양한 면접 정보를 종합적으로 분석하여 모델이 면접 답변의 적합성을 정밀하게 평가할 수 있는 기준을 마련하고, 답변에 대한 분석의 깊이를 풍부하게 하여 모델이 지원자의 면접 성공 가능성을 보다 정확하게 예측할 수 있도록 한다.

3.2.2 채용면접 인터뷰 데이터셋 전처리

AI-Hub의 채용면접 인터뷰 데이터의 면접 답변에 태깅 된 감정 및 의도 라벨링은 각 응답의 특성을 반영하는 고유한 값들로 구성되어 있다. 예를 들어, '사실'에 해당하는 답변은 'u-fact'로 표기하며, '회계감각'을 나타낼 경우 'm-acct'로 라벨링 되어 각 요소의 특징을 함축한다. 감정 요소의 경우, 긍정, 부정, 중립 감정에 따라 해당 감정이 드러날 경우 각각 p, n, u로 라벨링 값을 시작하도록 하여 검출된 감정의 성향을 명확하게 나타내었다. 또한 의도 요소의 경우, 직종별 의도에 따라 해당 직종을 상징하는 이니셜을 라벨 앞에 추가하여 해당 답변이 어떤 직무의 의도 또는 역량인지 확인할 수 있도록 되어있다.

따라서 본 연구에서는 LLM 모델이 면접 답변의 감정과 의도 요소를 더 잘 이해할 수 있도록, 각 요소에 한국어 설명을 추가하는 작업을 진행했다. 실질적 정보를 담고 있는 요소 명을 한국어로 번역하고, 기존 정보는 괄호 내에 표기하여 감정 및 의도 데이터를 정제했다. 즉, 표 2의 예와 같이 기존 'u-fact'는 '사실(Fact)'로, 'm-aact'은 '회계감각(Technology)'으로 변환했으며, 이후 각 면접 답변에 대한 감정 또는 의도 요소를 정확하게 대응시키기 위해, 모든 분류 결과를 "답변-감정(의도) 라벨링 결과"의 형식으로 정리했다. 또한 하나의 답변에 다수의 감정 및 의도 요소 라벨링 값이 존재하는 경우를 고려하여, 각 요소를 구분하기 위해 숫자를 할당했다.

3.3 주요 의도 요소 선정

다양한 역량에 대한 역량별 조건 프롬프트 설정과 정제의 한계로 인해 의도 요소의 범위를 좁혀 검출에 사용할 주요 의도 요소를 선정했다. 이때, 일부 데이터에 대한 의도 요소를 수동 검출하여 자주 발견되고 분류 오류가 잦은 의도 요소를 식별하여 가장 빈번히 검출되며 잦은 오류를 보인 의도 요소 상위 5가지를 주요 검출 요소로 선정했다. 검출에 사용한 의도 요소는 성격, 가치관, 자기계발, 통찰력, 전문성으로 이외의 요소들은 기타 요소로 처리했으며, 5가지 의도 요소는 국가직무능력표준(NCS)과 내부 보유 역량 사전 등을 활용해 정의한 AI-Hub의 채용면접 인터뷰 데이터 가이드라인 내 역량 요소를 참고했다.

선정한 5가지 주요 의도 요소에 대한 정의는 표 3과 같다. 성격과 가치관, 자기계발 요소는 채용면접 인터뷰 데이터 가이드라인의 의도 요소 정의와 일치하지만, 통찰력과 전문성 요소의 경우, 데이터의 의도 요소를 수동 분석하는 과정에서 의도 요소 검출 조건에 대한 확장의 필요가 있어 정의를 수정했다.

표 2. 정제 전 채용면접 인터뷰 데이터셋(a)과 정제 후 데이터셋(b)의 예
Table 2. Example of the (a) pre-refined and (b) post-refined dataset of interview

question_raw.text	기업이 자금을 조달하는 직간접적인 방법에 대해서 아는 만큼 최대한 자세한 설명 부탁드리겠습니다
answer_raw.text	기업이 자금을 조달하는 방식은 첫 번째는 채권을 발행하여 타인에게 자금을 빌려오는 차입이 있습니다. … (생략)
answer_raw_emotion.text	채권은 타인의 돈을 빌리는 것이기 때문에 … (생략) … 온전히 기업의 돈으로 남게 됩니다.
answer_raw_emotion.expression	u-fact
answer_raw_emotion.category	neutral
answer_raw_intent.text	기업이 자금을 조달하는 방식은 첫 번째는 채권을 발행하여 타인에게 자금을 빌려오는 차입이 있습니다.
answer_raw_intent.expression	m_acct
answer_raw_intent.category	technology
⋮	⋮
(a) 정제 전 채용 면접 인터뷰 데이터셋 (a) Recruitment interview dataset before refinements	
question_raw.text	기업이 자금을 조달하는 직간접적인 방법에 대해서 아는 만큼 최대한 자세한 설명 부탁드리겠습니다
answer_raw.text	기업이 자금을 조달하는 방식은 첫 번째는 채권을 발행하여 타인에게 자금을 빌려오는 차입이 있습니다. . . . (생략)
answer.emotion	1) 채권은 타인의 돈을 빌리는 것이기 때문에 … (생략) … 온전히 기업의 돈으로 남게 됩니다.-answer_emotion:사실(neutral)
answer.intent	2) 기업이 자금을 조달하는 방식은 첫 번째는 채권을 발행하여 타인에게 자금을 빌려오는 차입이 있습니다.-answer_intent:회계감각(technology) 3) … (생략)
(b) 정제 후 채용 면접 인터뷰 데이터셋 (b) Recruitment interview dataset after refinements	

표 3. 5가지 주요 의도 요소 정의
Table 3. Pefinition of 5 Key intent elements

Personality	지원자의 성격 또는 성향을 확인할 수 있는 질문 또는 답변
Philosophy	자신의 신념 또는 중요하게 여기는 관념, 생각 등이 드러나는 질문 및 답변
Self-improvement	자기계발에 대한 질문 또는 자신을 개선하고자 하거나 자기주도적으로 무엇인가를 해내려는 내용이 포함된 답변
Insight	어떤 상황이나 현상에 대한 지원자의 생각을 묻는 질문 또는 현 상황에 있어 최적의 선택을 할 수 있는 해결 방법을 포함한 질문 및 답변
Expertise	특정 직종 관련 전문용어나 개념에 대한 지식 확인과 관련된 질문 또는 지원자의 직종 전문성을 확인할 수 있는 질문 또는 답변

통찰력의 경우, 어떤 안건에 대해 여러 가지 변수들을 검토하는 것 외에도 어떤 상황이나 현상에 대한 지원자의 생각을 묻거나 다방면적 사고를 통해 지원자의 해결 방안을 확인할 수 있는 답변을 포함하여 종합적인 사고에 대해 포괄적 평가가 가능하도록 했다.

또한 전문성의 경우, 지원자가 보유한 기술 및 역량이 드러나는 질문뿐만 아니라 직종별 전문용어와 개념을 묻는 질문에 정확히 이해하고 대응한 답변을 포함하여 전문성의 개념을 확장시켰다.

3.4 의도 요소 검출

의도 요소 검출 과정은 면접 질의응답 데이터의 답변 적절성을 판단할 때 모델의 답변 데이터에 대한 이해를 돕기 위해 사용된다. 기존 채용면접 인터뷰 데이터에서 질의응답 정보와 검출된 감정 및 의도 라벨링 값은 그림 2와 같이 5가지 주요 의도 요소의 특성을 반영한 프롬프트를 사용하여 OpenAI의 GPT-4-Turbo 모델[18]을 통해 질문별 의도 요소가 검출된다.

여기서, [ ] 값은 각 의도 요소별 특징이 서술되며, { } 값은 각 질의응답 정보가 차례로 입력된다. 의도 요소 검출 프롬프트에는 성격, 가치관, 자기계발, 통찰력, 전문성 각각의 정의와 특성에 대한 설명, 그리고 5가지 의도 요소 외 기타 요소에 대한 간략한 설명을 포함한다. 5가지 주요 의도 요소에 대한 설명은 해당 역량이 잘 드러나는 질문과 답변의 정보와 특성을 상세히 서술하였으며, 5가지 의도 요소로 분류되지 않는 기타 요소들을 제대로 판별해 내기 위해 채용면접 인터뷰 데이터의 가이드라인 내 5가지 의도 요소 이외의 역량들을 함께 제공했다.

예를 들어, 자기계발의 경우 '자신의 장단점을 언급하며 단점을 개선하고자 하는 의지와 현실적 시행으로 발전한 경우' 또는 '취미 또는 개인 활동에서 유익한 활동을 통해 자기계발을 행하고 있는 경우'와 같이 질문 또는 답변 내에서 각 역량을 파악할 수 있는 답변 특징을 서술하여 모델이 이러한 특성을 참고하여 더 정확한 판정을 할 수 있도록 도왔다.

해당 의도 요소 검출 프롬프트는 모든 면접 질의응답 데이터에 대해 공통으로 적용되며, 이후 검출된 의도 요소와 연관된 개별 의도 조건 프롬프트를 연결하여 면접 질의응답 데이터 분석의 깊이를 더한다.

```
Please analyze the question, answer, and answer breakdown to detect the intent elements that can be identified in the question and answer by referring to the intent element characteristics. If an element is not included in the five intent elements, categorize it 'Other'.

###
<<5 Intent Element Traits>>
[Personality characteristic]
[Philosophy characteristic]
[Self-improvement characteristic]
[Insight characteristic]
[Expertise characteristic]

Other (Work Experience, Self-Description, Adaptability, ...(omit))
###

###
<<Interview question and answers with answer breakdown>>
Question: {question information}
Answer: {answer information}
Answer breakdown: {answer breakdown information}
```

그림 2. 의도 요소 검출 프롬프트 구조
Fig. 2. Structure of intent element detection prompt

3.5 답변 적절성 판단

면접 답변의 적절성을 판단하기 위해 사용되는 프롬프트는 2.4절에서 언급된 과정을 통해 각 면접 질의응답 데이터별로 식별된 의도 요소를 기반으로 구축되며 그림 3과 같다. 여기서 사용되는 [ ] 값은 적절한 답변과 적절하지 않은 답변의 특징 설명과 각 의도 요소별 고려 사항 프롬프트가 서술되며, { } 값은 각 질의응답 정보가 차례로 입력된다.

답변의 적절성 판단 시 모든 역량에 공통적으로 적용되는 기본 조건 프롬프트에는 적절한 답변과 적절하지 않은 답변의 특징의 일부가 담겨있다. 기본 조건 프롬프트는 의사소통, 언어 형식, 역량, 의지 등과 같이 면접 태도를 판단할 수 있는 특징과 더불어, 경험과 능력 등을 평가할 수 있는 답변의 품질과 관련된 특징을 포함한다.

5가지 주요 의도 요소에 대한 프롬프트는 각 요소별로 답변의 적절성을 평가할 때 기본 평가 조건 외에 고려해야 할 추가적인 조건이나 답변의 특징을 제공한다. 예를 들어 가치관과 관련된 프롬프트는 지원자가 자신의 견해를 얼마나 설득력 있고 명확하게 표현하는지, 그리고 구체적 예시 또는 사례를 들 때 지나치게 개인적인 내용을 포함하지 않는지 등을 확인하는 추가 조건을 포함한다. 전문성 관련 프롬프트의 경우, 전문용어나 개념에 대한 답변 분석 시, 구체적인 예시나 경험의 언급이 필요하지 않음에도 이러한 언급의 부재로 인해 답변이 잘못 분류되는 것을 방지하기 위해 조건을 추가했다. 전문용어나 개념에 대한 지식을 확인하는 질문에 대한 답변에서 개념의 이해와 정확성을 중점적으로 평가해야 한다는 조건을 명시했다.

이와 같이 의도 요소 검출 과정을 통해 검출된 의도 요소는 해당 의도 요소에 대한 추가 조건 프롬프트와 연결되어 기본 프롬프트 아래에 배치되며, OpenAI GPT-4-Turbo 모델을 통해 모델은 판단 근거와 함께 답변의 적절성을 평가한다. 위의 과정으로 구축된 프롬프트는 면접 평가 과정에서 모델에게 면접 답변의 적절성을 보다 정확하게 판단할 수 있는 근거를 제공하며, 모델이 면접 답변 데이터를 보다 정밀하게 분석하고 평가할 수 있도록 돕는다.

IV. 실험 및 분석

4.1 제안 시스템 적용

본 연구에서는 5가지 주요 의도 요소에 대한 답변 판단 근거의 타당성이 개선을 효과적으로 확인하기 위해, 채용면접 인터뷰 데이터 6.8만 건의 질의응답 데이터 중 5가지 주요 의도 요소가 잘 드러나는 1,591건의 데이터에 대해 제안하는 정제 시스템을 적용했다.

```
Using the characteristics of appropriate and inappropriate responses below, please evaluate the interviewee's interview demeanor and the quality of their responses as evidenced by the given question and answer information to determine the appropriateness of the answer. Please circle an O for appropriate and an X for inappropriate. And also include 3-4 lines of your analysis.

###
[Characteristics of an appropriate answer]

[Characteristics of an inappropriate answer]

* [Expertise essential considerations]
###

###
<<Interview question and answers with answer breakdown>>
Question: {question information}
Answer: {answer information}
Answer breakdown: {answer breakdown information}
```

그림 3. 답변 적절성 판단 시 사용되는 프롬프트의 예 (전문성 의도 요소 검출의 경우)

Fig. 3. Example of a prompt used for assessing answer appropriateness (for detecting the expertise intent element)

사용되는 1,591건의 데이터는 answer.intent 요소 중 5가지 주요 의도 요소가 2개 이상 검출된 질의응답 데이터를 포함하고 있으며, 2.4절과 2.5절의 과정을 거쳐 각 질의응답에 대한 의도 요소와 답변 적절성에 대한 판별 결괏값을 도출했다.

구축한 1,591건의 질의응답 데이터셋에 대한 데이터 통계는 다음과 같다.

표 4. 의도 요소 검출 개수별 데이터 분류 분포
Table 4. Data distribution by detected Intent elements

Intent detection count	Classified data count
2	1,040
3	353
4	141
5	34
6	14
7	3
8	4
9	2
Total	1,591

표 4는 데이터셋 내에서 검출된 의도 요소 개수에 따라 분류된 데이터의 분포를 보여준다. 검출된 의도 요소 개수가 2개일 때 1,040개로 가장 많았으며, 해당 데이터에는 서로 다른 5가지 주요 의도 요소를 가진 질의응답 데이터가 포함되어 있다.

표 5. 경력 유무에 따른 데이터 개수
Table 5. Data entries by career experience status

Experience status	Classified data count
Entry-level	1,331
Experienced	260
Total	1,591

표 6. 연령대별 분류된 데이터의 개수
Table 6. Data entries by age groups

Age group	Data count
Under 34	713
35-44	340
45-54	201
55 and over	337
Total	1,591

표 5-6은 구축한 데이터셋에서 경력 유무와 연령대별 데이터의 분포를 보여준다. 표 5를 통해 신입과 경력직 면접자 데이터 비율이 약 6:1로, 데이터셋이 경력직 면접자의 데이터도 일부 포함하고 있음을 알 수 있다. 또한 표 6을 통해 34세 이하부터 55세 이상에 이르는 다양한 연령대 지원자의 질의응답 데이터가 균등하게 포함되어 있음을 확인했다. 이를 통해 본 연구의 데이터셋이 다양한 유형의 면접자에 대한 질의응답 분석까지 포괄적으로 지원할 수 있는 구조를 갖추고 있음을 확인할 수 있다.

4.2 실험

본 실험에서는 제안하는 데이터 선별 시스템이 기존 선별 시스템 대비 데이터 판별 정확성과 선별 근거의 타당성 개선 정도를 확인한다. 이를 위해 제안하는 데이터 정제 시스템을 거쳐 얻은 판별 결과와 그 근거에 대해 인간 검증 및 GPT-4-Turbo 모델을 활용한 점수 측정 방식을 사용했다.

이를 위해 구축한 1,591건의 질의응답 데이터셋 중 약 10%의 데이터를 무작위로 선택하여 수동 평가 방식으로 답변 데이터의 적절성을 평가하여, 의도 분석 과정을 거치지 않은 기본 프롬프트로 얻은 답변 데이터의 적절성 판단 결과와 제안하는 정제 시스템을 거쳐 얻은 답변 판별 결과에 대한 정확도를 비교했다.

표 7. 수동 검증을 통한 프롬프트별 정확도
Table 7. Prompt accuracy through manual verification

Prompt type	Accuracy
Basic prompt	83.33%
Our prompt	87.33%

의도 분석 과정을 거치지 않고 공통으로 제공되는 기본 프롬프트로 답변 데이터를 판별했을 때와 제안한 데이터 정제 시스템을 거쳐 판별된 답변 데이터의 정확도는 표 7과 같다. 표 7에서 Basic 프롬프트는 기본 공통 프롬프트의 정확도를 나타낸다.

제안하는 데이터 정제 프롬프트 적용 결과, 기본 공통 프롬프트 사용 대비 답변의 적절성 판별 정확도가 4% 향상되었음을 확인할 수 있었다. 특히, '통찰력' 및 '전문성' 의도 요소에 대한 판별에 있어 뚜렷한 개선 결과를 보여주었다.

이는 제안하는 데이터 정제 프롬프트가 답변의 다양성과 복잡성을 포괄적으로 고려할 수 있는 평가 기준을 제공함으로써, 모델이 보다 정밀한 판단을 할 수 있도록 도울 수 있음을 시사한다.

제안하는 데이터 정제 시스템에서는 얻은 답변 판별 근거의 타당성의 향상 정도를 파악하기 위해, OpenAI의 GPT-4-Turbo 모델을 활용하였으며, 면접 질의응답 정보와 관련된 답변의 적절성 판단 결과와 근거 정보, 답변 평가 시 참고했던 의도 요소 고려 사항 정보를 제공한 후 답변 판별 근거의 타당성을 Likert 척도 [19]로 평가했다.

기본 프롬프트를 사용하여 답변 데이터를 판별한 근거와 제안하는 데이터 정제 시스템을 거친 데이터에 대한 GPT-4-Turbo를 활용한 답변 판별 근거의 타당성 평가 결과의 평균 점수는 표 8과 같다.

기존의 공통 프롬프트를 활용한 답변 판별 근거의 평균 타당성 점수는 3.27점으로 나타났다. 반면, 본 연구에서 제안하는 데이터 정제 프롬프트 적용 결과 평균 타당성 점수가 4.09점으로 기존 대비 0.82점 높은 결과를 보였다. 또한 표 8에서 확인할 수 있듯 제안하는 데이터 정제 프롬프트는 주요 5가지 의도 요소를 포함하는 답변의 타당성 평균 점수에서 대부분 4점 이상을 기록했다. 이는 제안하는 데이터 정제 프롬프트가 면접 답변의 적절성 판별 과정에서 5가지 주요 의도 요소에 대한 깊이 있는 이해와 분석을 증진시켜 기본 공통 프롬프트 대비 보다 정교하고 타당한 근거를 기반으로 답변을 평가하는데 중요한 기반을 제공함을 짐작할 수 있다.

표 8. GPT-4-Turbo 모델을 활용한 답변 근거 타당성 평가 Likert 척도 평균 점수
Table 8. Answer validity asessment with GPT-4-Turbo: average likert scale

		Score avg
Basic prompt		3.27
Our prompt		4.09
Our prompt	Personality	3.96
	Philosophy	4.11
	Insight	4.00
	Self-improvement	4.72
	Expertise	4.64

4.3 한계

제안하는 정제 시스템의 한계는 크게 두 가지로 요약할 수 있다.

첫 번째로, 제안하는 프롬프트 사용 시 질의응답 데이터 내 의도 요소를 검출하는 과정에서 기타 의도 요소의 분류가 완전히 이루어지지 않는 경우가 일부 존재한다. 기타 의도 요소에 대한 간략한 정보를 제공했음에도 불구하고, 의도 요소 검출 과정에서 질의응답 데이터가 주요 5가지 의도 요소 중 하나를 포함하여 분류되려는 경향이 있음을 발견했다.

두 번째로, 답변에 두 가지 이상의 의도 요소가 존재하는 경우, 일부 데이터의 답변 적절성 판별 과정에서 프롬프트 사이에 충돌이 발생했다. 구체적인 예시나 경험을 요구하는 의도 프롬프트와 그렇지 않은 의도 프롬프트가 함께 제공될 경우, 모델이 답변의 적절성을 판별하는 과정에서 충돌이 발생했다.

이러한 한계점들은 향후 연구에서 의도 요소별 프롬프트의 고려 사항을 더 명확히 구체화하고 추가 정제 과정을 통해 모델을 보다 면밀하고 정확한 판단을 할 수 있도록 개선할 계획이다.

V. 결 론

본 연구에서는 AI 면접 분석 보고서 생성 모델 학습을 위한 고품질 데이터셋 구축을 목표로, '성격, 가치관, 통찰력, 자기계발, 전문성'의 5가지 주요 의도 요소를 검출하고, 검출된 의도 요소와 대응되는 개별 프롬프트를 추가하여 보다 타당한 근거를 기반으로 답변의 적절성을 평가할 수 있는 데이터 정제 시스템을 제안했다. 2개 이상의 의도 태그를 가진 1,591개의 데이터로 제안하는 정제 시스템을 적용시킨 후 무작위 10%의 데이터를 통한 실험 결과, 기존 프롬프트 대비 4%의 판별 정확도 향상과 4.09점의 평균 답변 근거 타당성 점수로 대부분의 의도 요소가 4점 이상의 점수를 기록했음을 확인할 수 있었다.

제안하는 정제 시스템을 통해 AI 면접 분석 보고서 생성 모델이 다양한 유형의 질의응답 정보에 효과적으로 대응할 수 있도록 하며, 면접 평가 과정에

서 질의응답 정보를 보다 정밀히 분석하여 사용자 답변에 대한 판단 적절성과 타당성의 성능을 향상시키는 데 기여하는 주는 데이터셋 정제 방식임을 확인했다.

하지만 본 연구는 일부 데이터의 기타 요소의 분류 불완전성과 두 가지 이상 검출된 의도 요소의 개별 프롬프트 간 충돌 문제를 포함하는 한계가 존재한다. 향후 프롬프트 개선 및 추가 정제 시스템을 도입하고 의도 요소를 확장하여 다양한 면접 유형에 대한 모델의 면접 분석 성능 안정성을 향상시키는 연구를 진행하고자 한다. 제안된 정제 시스템으로 얻은 데이터는 다양한 면접 질의응답 정보와 분석 결과를 포함하여 모델 성능을 극대화할 수 있는 다양하고 풍부한 데이터셋을 구축하는데 기여할 것으로 기대한다.

References

[1] O. Bruna, H. Avetisyan, and J. Holub, "Emotion models for textual emotion classification", Journal of Physics: Conference Series, Vol. 772, No. 1, Nov. 2016. https://doi.org/10.1088/1742-6596/772/1/012063.

[2] Z. Wang, S. B. Ho, and E. Cambria, "A review of emotion sensing: categorization models and algorithms", Multimedia Tools and Applications, Vol. 79, No. 47-48, pp. 35553-35582, Jan. 2020. https://doi.org/10.1007/s11042-019-08328-z.

[3] S. An and O. Jeong, "A Study on the Psychological Counseling AI Chatbot System based on Sentiment Analysis", Journal of Information Technology Services, Vol. 20, No. 3, pp. 75-86, Jun. 2021. https://doi.org/10.9716/KITS.2021.20.3.075.

[4] J. Shin, H. Im, and B. Lee, "Topic modeling and sentiment analysis of service quality for integrated resorts in Korea", International Journal of Tourism and Hospitality Research, Vol. 35, No. 11, pp. 191-206, Nov. 2021. https://doi.org/10.21298/ijthr.2021.11.35.11.191.

[5] H. Kim, "Classification and Prediction of Fashion Product Online Review", Korean Journal of Human Ecology, Vol. 32, No. 6, pp. 767-782, Dec. 2023. https://doi.org/10.5934/kjhe.2023.32.6.767.

[6] Y. Jang, J. Choi, and H. Kim, "KcBert-based Movie Review Corpus Emotion Analysis Using Emotion Vocabulary Dictionary", Journal of KIISE, Vol. 49, No. 8, pp. 608-616, Aug. 2022. https://doi.org/10.5626/jok.2022.49.8.608.

[7] J. Lee, KcBERT: Korean comments BERT, https://github.com/Beomi/KcBERT-Finetune [accessed: Mar. 12, 2024].

[8] SKTBrain, KoBERT, https://github.com/SKTBrain/KoBERT [accessed: Mar. 13, 2024].

[9] Y. Heo, J. Lee, Y. Shin, and C. Kim, "Implementation of an intention classification model for interviewer answers for AI interviews", Journal of Computing Science and Engineering, pp. 1296-1298, Dec. 2022.

[10] National Competency Standards, https://www.ncs.go.kr/th01/TH-102-001-02.scdo [accessed: May. 10, 2024].

[11] Recruitment interview data, AI-Hub, https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71592 [accessed: Mar. 10, 2024].

[12] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications", arXiv:2402.07927, Feb. 2024. https://doi.org/10.48550/arXiv.2402.07927.

[13] Z. Wang, R. Panda, L. Karlinsky, R. Feris, H. Sun, and Y. Kim, "Multitask Prompt Tuning Enables Parameter-Efficient Transfer Learning", The Eleventh International Conference on Learning Representations, Kigali Rwanda, Mar. 2023. https://doi.org/10.48550/arXiv.2303.02861.

[14] Z. Zhang, S. Wang, W. Yu, Y. Xu, D. Iter, Q. Zeng, Y. Liu, C. Zhu, and M. Jiang, "Auto-Instruct: Automatic Instruction Generation and Ranking for Black-Box Language Models", Association for Computational Linguistics, pp.

9850-9867, Oct. 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.659.

[15] W. Liu, W. Zeng, K. He, Y. Jiang, and J. He, "What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning", The Twelfth International Conference on Learning Representations, Vienna Austria, Jan. 2024. https://doi:10.48550/arXiv.2312.15685.

[16] C. Kim and H. Jung, "GPT-based Data Augmentation Method for Improving Emotion Analysis Model Performance", Journal of Korean Institute of Information Technology, Vol. 22, No. 1, pp. 61-69, Jan. 2024. https://doi.org/10.14801/jkiit.2024.22.1.61.

[17] B. Clavié, A. Ciceu, F. Naylor, G. Soulié, and T. Brightwell, "Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification", International Conference on Applications of Natural Language to Information Systems, pp. 3-17, Jun. 2023. https://doi.org/10.1007/978-3-031-35320-8_1.

[18] OpenAI GPT-4-Turbo (gpt-4-0125-preview), https://openai.com/blog/new-embedding-models-and-api-updates [accessed: Mar. 24, 2024].

[19] R. Likert, "A technique for the measurement of attitudes", Archives of Psychology, 1932.

저자소개

윤 채 원 (Chaewon Yoon)

2021년 3월 ~ 현재 : 국립금오공과대학교 컴퓨터공학과 학사과정

관심분야 : 자연어처리, 합성데이터 자동 구축, LLM(fine-tuning)

정 유 철 (Yuchul Jung)

2011년 2월 : 한국과학기술원 (KAIST) 전산학과(공학박사)

2013년 7월 : 한국전자통신연구원 (ETRI) 선임연구원

2017년 8월 : 한국과학기술정보연구원(KISTI) 선임연구원

2022년 2월 : 국립금오공과대학교 컴퓨터공학과 조교수

2022년 2월 ~ 현재 : 국립금오공과대학교 인공지능공학과 부교수

관심분야 : 거대언어모델, 자연어처리, 지식그래프, 한국어 음성 인식/합성, AI응용