JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 4, pp. 1061-1068, Apr. 2026

Check for updates

LLM 기반 진로문장완성검사 자동 판정을 위한 프롬프트 엔지니어링 전략 비교 연구

최 예 슬[1,2] · 오 태 연[3*]
[1]전북대학교 교육학과 조교수
[2]서울과학종합대학원대학교 AI·빅데이터 석사
[3]서울과학종합대학원대학교 AI첨단학과 조교수

Comparative Study of Prompt Engineering Strategies for LLM-Based Automated Grading of Career Sentence Completion Test

Yeseul Choi[1,2] · Taeyeon Oh[3*]

[1]Assistant Professor, Department of Education, Jeonbuk National University, Jeonbuk 54896, Korea
[2]M.S., Seoul AI School, aSSIST University, Seoul 03767, Korea
[3]Assistant Professor, Seoul AI School, aSSIST University, Seoul 03767, Korea

[요 약]

본 연구는 파인튜닝 없이 프롬프트 엔지니어링만으로 대규모 언어모델(LLM)이 진로문장완성검사 응답에 대한 자동 등급 판정을 수행할 수 있는지를 검증하고, 프롬프트 구성요소가 판정 성능에 미치는 영향을 분석하였다. AI-Hub 진로문장완성검사 데이터셋에서 1,200건을 층화 표집하여, 역할 지시, 평가 루브릭, Few-shot 예시, Chain-of-Thought 추론을 누적적으로 추가한 네 가지 프롬프트 조건을 설계하고 GPT-4o-mini의 판정 성능을 비교하였다. 분석 결과, Few-shot 예시를 포함한 조건이 정확도 68.9%, QWK 0.725로 가장 높은 성능을 보이며 자동 채점의 실용적 수용 기준을 충족하였다. 특히 경계적 응답에 대한 판정 정확도가 크게 향상되었고, 과대·과소채점이 균형적으로 감소하였다. 본 연구는 진로상담 맥락에서 LLM 기반 자동 판정의 실용 가능성을 제시하며, 프롬프트 설계에서 구체적 예시 제공의 중요성을 시사한다.

[Abstract]

This study examines whether large language models (LLMs) can perform automated grade classification for Career Sentence Completion Test (SCT) responses using prompt engineering alone, without fine-tuning, and analyzes how prompt components affect classification performance. Using 1,200 stratified samples from the AI-Hub Career SCT dataset, four prompt conditions were designed by cumulatively adding role instructions, an evaluation rubric, few-shot examples, and chain-of-thought reasoning, and the performance of GPT-4o-mini was compared across these conditions. The results show that the condition including few-shot examples achieved the best performance, with an accuracy of 68.9% and a QWK of 0.725, meeting the practical acceptance criterion for automated scoring. In particular, the classification accuracy for borderline responses improved, and both over-grading and under-grading were reduced. These findings indicate the practical applicability of LLM-based automated grading in career counseling contexts and highlight the importance of providing concrete examples in prompt design.

색인어 : 대규모 언어모델, 프롬프트 엔지니어링, 진로문장완성검사, 자동 채점, Few-shot 프롬프팅

Keyword : Large Language Model, Prompt Engineering, Career Sentence Completion Test, Automated Grading, Few-Shot Prompting

http://dx.doi.org/10.9728/dcs.2026.27.4.1061

Received 28 January 2026; **Revised** 23 February 2026
Accepted 17 March 2026

***Corresponding Author; Taeyeon Oh**

Tel: +82-70-7012-2700
E-mail: tyoh@assist.ac.kr

Ⅰ. 서 론

2025년 고교학점제가 본격 시행되면서 학생 개개인의 흥미와 적성에 기반한 맞춤형 진로 지원의 중요성이 강조되고 있다. 그러나 진로 전담교사 1인이 다수의 학생을 담당하는 구조적 한계로 인해, 개별 학생의 진로 발달을 장기적으로 지원하는 데 현실적인 어려움이 지속되고 있다[1]. 이에 진로상담 교사의 전문적 판단을 보완하고 업무 부담을 경감할 수 있는 기술적 지원 방안에 대한 요구가 증가하고 있다.

진로상담에서 활용되는 문장완성검사(Sentence Completion Test, SCT)는 학생이 미완성 문장을 완성하는 방식으로 진로 관련 사고를 서술하도록 유도하는 텍스트 기반 검사이다[2]. 그러나 SCT 응답은 문장이 짧고 맥락이 불완전하며, 해석 과정에서 상담자의 전문적 판단에 대한 의존도가 높아 대규모 학생 집단을 대상으로 활용하는 데 제약이 있다.

최근 대규모 언어 모델(Large Language Model, LLM)이 교육 평가 영역에서 활용 가능성이 탐색되고 있다[3]. 기존 연구들은 주로 미세조정(fine-tuning)을 통해 서술형 응답의 자동 분류를 시도해 왔으나[4], 이는 도메인별 학습 데이터 확보와 추가 컴퓨팅 자원을 요구한다는 한계가 있다. 이에 따라 모델 파라미터를 갱신하지 않고 프롬프트 설계만으로 등급 판정을 수행하는 프롬프트 기반 접근이 주목받고 있다[5],[6].

그러나 기존 연구들은 대부분 단일 프롬프트 전략의 효과를 검증하는 데 그쳤으며, 역할 부여, 루브릭 제시, Few-shot 예시, Chain-of-Thought 추론 등 프롬프트 엔지니어링의 핵심 구성요소가 등급 판정 성능에 어떻게 기여하는지를 분해하여 분석한 연구는 제한적이다. 이러한 배경에서 본 연구는 파인튜닝 없이 프롬프트 엔지니어링을 통해 LLM이 진로문장완성검사 응답에 대한 등급 판정을 수행할 수 있는지를 탐색하고, 프롬프트 구성요소가 판정 성능에 미치는 영향을 체계적으로 분석하고자 한다.

Ⅱ. 관련 연구

2-1 LLM 기반 자동 채점 연구

자동 채점(Automated Essay Scoring, AES)은 1960년대부터 통계적 회귀 기반 접근에서 출발하여 이후 기계학습 및 딥러닝 기반 방법으로 발전해 왔다[7]. 최근 GPT, Claude 등 LLM의 등장으로 추가 학습 없이 프롬프트 기반 추론만으로 평가 과제를 수행할 수 있는 새로운 접근 방식이 제시되고 있다[3]. Nkoyo et al.[8]의 종합적 서베이에 따르면, GPT-4는 K-12 단답형 문항에서 인간 평가자와 유사한 수준의 일치도(Cohen's κ = 0.70)를 달성하였으며, 수동 채점 대비 약 81%의 시간 절감 효과를 보여 LLM이 교육 평가에서 실용적 수준에 도달할 수 있음을 실증하였다. 또한 Carpenter et al.[9]은 학생 서술형 응답 평가에서 Few-shot 프롬프팅 기반 GPT-4가 매크로 F1에서 가장 높은 성능을 달성하였으며, 프롬프트에 포함되는 구성 요소(rubric, exemplar, 학생 응답 예시)의 조합에 따라 성능이 크게 달라짐을 보고하였다. 특히 rubric 정보 추가가 일관되게 성능 향상을 가져온 반면, 예시의 특성에 따라 오히려 성능이 저하될 수 있음을 확인하여, Few-shot 프롬프팅에서 예시 설계의 중요성을 강조하였다. Yavuz et al.[10]은 영어 학습자 에세이 채점에서 프롬프트 구체화와 온도(temperature) 조정만으로 ChatGPT가 인간 평가자를 상회하는 매우 높은 신뢰도(ICC = 0.972)를 달성할 수 있음을 확인하였다. 이러한 결과들은 LLM 기반 자동 채점에서 프롬프트 설계의 정교화가 핵심 전략임을 시사한다. 이와 유사하게, 최근 연구들에서는 Bronlet[11]이 제시한 자아발달 단계나 Huang et al.[12]의 우울증 검사를 위한 SCT 기반의 문장 분류 과제에서도 LLM을 활용한 자동 채점이 전문가 채점과 높은 일치도를 보여, 프롬프트 기반 접근의 실용적 활용 가능성을 뒷받침하고 있다.

국내 진로상담 및 교육평가 영역에서도 최근 LLM을 활용한 자동평가 연구가 점차 확대되고 있다. 강혜림과 백재파[6]는 GPT-4o 기반 프롬프트 엔지니어링을 적용하여 한국어 학습자의 말하기 전사 자료를 자동 채점하는 모델을 설계하고, 인간 채점자와의 일치도를 분석한 결과 QWK 0.81 수준의 높은 일치도를 달성하였다. 김병학과 이수안[4]은 진로문장완성검사 데이터를 활용하여 오픈소스 LLM을 LoRA 방식으로 미세조정한 결과, 피드백 생성 품질이 크게 향상됨을 확인하였다. 이서진과 민경식[13]은 GPT-3.5를 활용하여 진로성숙도 응답을 자동 정량화하고 문항반응이론에 적용한 결과, 재태그된 데이터가 초기 태그 대비 더 높은 정보함수와 변별도를 보임을 보고하였다. 이는 LLM 기반 자동 정량화가 단순 분류를 넘어 측정 모형 수준에서도 활용 가능함을 시사한다. 이원태와 윤수연[14]은 진로문장완성검사 데이터를 기반으로 청소년 진로상담을 위한 LLM 기반 챗봇 설계 가능성을 탐색하였으며, 진로상담 맥락에서 LLM의 응용 범위가 응답 채점이나 피드백 생성을 넘어 대화형 상담 지원 서비스로 확장될 수 있음을 제시하였다.

그러나 이러한 진로문장완성검사 영역에서의 국내 선행연구들은 주로 미세조정 기반 피드백 생성에 초점을 맞추거나, 단일 프롬프트 전략의 효과를 검증하는데 그치고 있다. 프롬프트 엔지니어링의 핵심 구성요소가 등급 판정 성능에 어떻게 기여하는지를 체계적으로 분석한 연구는 제한적이다. 특히 한국어 진로문장완성검사와 같이 응답 길이가 짧고 맥락 의존성이 높은 과제의 경우, 별도의 미세조정 없이 프롬프트 설계만으로 실용적 수준의 자동 등급 판정이 가능한지에 대한 체계적 검증이 부족한 상황이다. 또한 학교급 및 등급 수준에 따른 성능 편차를 교차 분석하여 현장 적용 가능성을 구체화한 연구 역시 미흡하다.

2-2 프롬프트 엔지니어링 전략

프롬프트 엔지니어링은 LLM의 출력을 원하는 방향으로

유도하기 위해 입력 프롬프트의 구조와 내용을 체계적으로 설계하는 과정이다[3]. 대표적인 프롬프트 전략으로는 역할 부여(role prompting), 루브릭 조건화(rubric-conditioned prompting), Few-shot 프롬프팅, Chain-of-Thought(CoT) 프롬프팅 등이 있다.

Jiang & Bosch[5]는 GPT-4 기반 단답형 자동 채점 연구에서 루브릭이 포함된 프롬프트가 기본 채점 도구로 활용 가능함을 확인하였다. 그러나 루브릭이 지나치게 추상적이거나 일반화된 경우에는 오히려 모델의 판단 정확도가 저하될 수 있음을 보고하여, 루브릭의 구체성이 LLM 채점 성능의 핵심 변수임을 시사하였다.

Kojima et al.[15]과 Wei et al.[16]은 Chain-of-Thought (CoT) 프롬프팅이 LLM의 추론 능력을 유도하는 데 효과적임을 제시하였다. CoT 프롬프팅은 모델이 최종 답변에 도달하기 전에 중간 추론 과정을 명시적으로 생성하도록 유도하는 방법으로, 복잡한 추론 과제에서 성능 향상을 가져올 수 있다. 그러나 단순한 분류 과제에서는 그 효과가 제한적일 수 있다는 지적도 있다.

Brown et al.[17]은 few-shot 예시 제공이 zero-shot 대비 성능 향상을 가져올 수 있음을 보고하였다. Few-shot 프롬프팅은 과제의 입출력 예시를 제공함으로써 모델이 과제의 패턴을 학습하도록 유도하는 방법이다. 박선우 외[18]는 한국어 환경에서 지시적(instructive) 프롬프트가 zero-shot 조건에서도 안정적인 성능을 보임을 확인하였으며, 프롬프트의 언어적 특성이 성능에 영향을 미칠 수 있음을 시사하였다.

Ⅲ. 연구 방법

3-1 분석 자료

본 연구에서는 AI Hub에서 제공하는 '진로문장완성검사 텍스트 데이터셋'을 활용하였다. 이 데이터셋은 한국직업능력연구원과 AI Hub가 공동으로 구축한 것으로, 중·고등학생의 진로문장완성검사 응답과 전문가 평가 결과를 포함하고 있다. 진로문장완성검사는 진로성숙도의 여러 하위 영역(자기이해, 진로탐색, 진로계획 등)을 측정하는 문항으로 구성되어 있으며, 각 응답에 대해 전문 상담교사가 '상', '중', '하' 3단계 등급으로 평가한 결과가 포함되어 있다.

전체 데이터셋에서 '자기이해 및 긍정적 자아상' 영역의 중·고등학생 응답을 분석 대상으로 선정하였다. 이 영역은 학생이 자신의 강점, 흥미, 가치관 등에 대한 이해 수준을 평가하는 문항으로 구성되어 있다. 학교급(중학교, 고등학교)과 전문가 등급(상, 중, 하)에 따른 층화 무선 표집을 실시하여 총 1,200건(학교급 2×등급 3×200건)을 최종 표본으로 구성하였다. 표 1은 분석 자료의 구성을 나타낸다.

표 1. 분석 자료의 구성
Table 1. Composition of analysis data

School level	High	Medium	Low	Total
Middle school	200	200	200	600
High school	200	200	200	600
Total	400	400	400	1,200

표 2. 실험 조건별 프롬프트 구성요소
Table 2. Prompt components by experimental condition

Condition	Added component	Specific content
Baseline	Role instruction	Career counselor role assignment
+Rubric	+Evaluation rubric	Natural language criteria for High/Medium/ Low grades
+Few-shot	+Grade examples	2 examples per grade (6 total)
+CoT	+CoT reasoning instruction	4-step explicit reasoning process

3-2 프롬프트 설계

본 연구에서는 프롬프트 구성요소를 누적적으로 추가하는 4가지 실험 조건을 설계하였다(표 2, 실제 프롬프트 부록 참고). 이러한 누적적 설계를 통해 각 구성요소의 개별적 기여도를 분석할 수 있다.

첫째, Baseline 조건은 역할 지시(role instruction)만 부여하였다. 역할 지시는 LLM에게 '진로상담 전문가'로서 학생의 응답을 평가하도록 요청하는 것으로, 과제의 맥락을 설정하는 최소한의 프롬프트이다.

둘째, +Rubric 조건은 Baseline에 평가 루브릭을 추가하였다. 루브릭은 '상', '중', '하' 각 등급의 판정 기준을 자연어 형태로 명시한 것으로, 원본 데이터셋의 평가 기준을 참조하여 구성하였다. '상' 등급은 자기이해가 구체적이고 긍정적 자아상이 명확하게 표현된 경우, '중' 등급은 자기이해가 부분적이거나 표현이 모호한 경우, '하' 등급은 자기이해가 부족하거나 부정적 자아상이 표현된 경우로 정의하였다.

셋째, +Few-shot 조건은 +Rubric에 등급별 예시를 추가하였다. 각 등급별로 2개씩 총 6개의 예시를 제공하였으며, 예시는 해당 등급의 전형적인 응답 패턴을 대표하도록 선정하였다. 예시 선정 시 응답 길이, 내용의 구체성, 표현의 명확성 등을 고려하였다.

넷째, +CoT 조건은 +Few-shot에 Chain-of-Thought 추론 지시를 추가하였다. 모델이 등급을 판정하기 전에 (1) 응답 내용 요약, (2) 자기이해 수준 분석, (3) 긍정적 자아상 표현 여부 확인, (4) 최종 등급 판정의 4단계 추론 과정을 명시적으로 수행하도록 지시하였다.

3-3 분석 방법

실험은 OpenAI의 GPT-4o-mini 모델을 사용하여 수행하

였다. GPT-4o-mini는 GPT-4 시리즈의 경량화 버전으로, 비용 효율성이 높으면서도 우수한 성능을 보이는 것으로 알려져 있다. 실험의 재현성을 위해 temperature를 0으로 설정하여 출력의 일관성을 확보하였으며, 각 응답에 대해 단일 예측값을 생성하였다.

평가 지표로는 정확도(Accuracy), Quadratic Weighted Kappa(QWK), Macro F1-Score를 사용하였다. 정확도는 전체 예측 중 정답과 일치하는 비율을 나타내며, 가장 직관적인 성능 지표이다. QWK는 순서형 데이터의 평가자 간 일치도를 측정하는 지표로, 자동 에세이 채점 분야의 표준 지표로 사용되고 있다[19]. QWK는 단순 일치뿐 아니라 오분류의 심각도(등급 간 거리)를 반영하므로, 순서형 분류 과제에서 더 적합한 평가 지표이다. 일반적으로 QWK 0.70 이상이면 실용적으로 활용 가능한 수준으로 간주된다. Macro F1-Score는 각 등급별 F1 점수의 평균으로, 등급 간 불균형에 영향받지 않고 전체적인 분류 성능을 평가할 수 있다.

또한 오분류 유형을 분석하기 위해 과대채점(over-scoring)과 과소채점(under-scoring) 비율을 산출하였다. 과대채점은 실제 등급보다 높은 등급으로 예측한 경우, 과소채점은 실제 등급보다 낮은 등급으로 예측한 경우를 의미한다. 이를 통해 각 프롬프트 조건의 채점 경향성을 파악할 수 있다.

Ⅳ. 연구 결과

4-1 프롬프트 조건별 채점 성능 비교

4가지 프롬프트 조건에 따른 자동 등급 판정 성능은 표 3과 같다. 전체 1,200건을 분석한 결과, +Few-shot 조건이 정확도 68.9%, QWK 0.725, Macro F1 0.695로 모든 평가 지표에서 가장 높은 성능을 보였다. 이는 본 연구에서 설정한 실용적 수용 기준(QWK≥0.70)을 충족하는 결과로, 프롬프트 엔지니어링만으로도 진로문장완성검사 자동 판정이 실무적으로 활용 가능한 수준에 도달할 수 있음을 시사한다(조건별 QWK 비교 그림 1 참조).

그러나 +Rubric 조건에서 오히려 성능이 하락하여, Baseline 대비 QWK가 0.706에서 0.638로 감소하였으며, 정확도와 Macro F1도 모두 하락하였다. 이러한 결과는 추상적인 평가 기준의 텍스트 제시만으로는 모델이 등급 간 경계를 명확히 학습하지 못하며, 오히려 과도하게 엄격한 판정 기준을 적용하게 되어 과소채점 비율이 80.0%까지 급증한 것으로 해석된다.

+Few-shot 조건에서는 등급별 구체적인 응답 예시 6개(등급당 2개)를 제공함으로써 성능이 크게 회복되어 최고 성능(QWK 0.725)을 기록하였다. 이는 추상적 기준보다 구체적인 예시가 모델의 등급 판정에 더 효과적인 가이드라인을 제공함을 시사한다. 특히 과대채점(50.1%)과 과소채점(49.9%)이 거의 균등하게 나타나, 가장 균형 잡힌 판정 경향을 보였다.

표 3. 프롬프트 조건별 채점 성능 비교
Table 3. Comparison of grading performance by prompt condition

Condition	Accuracy (%)	QWK	Macro F1	Over (%)	Under (%)
Baseline	65.8	0.706	0.641	56.9	43.1
+Rubric	62.4	0.638	0.599	20.0	80.0
+Few-shot	68.9	0.725	0.695	50.1	49.9
+CoT	67.4	0.719	0.670	72.9	27.1

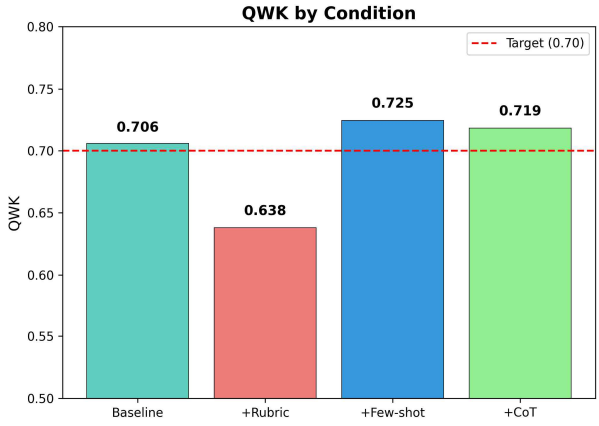

그림 1. 프롬프트 조건별 QWK 비교
Fig. 1. QWK comparison by prompt condition

+CoT 조건은 단계별 추론 과정을 명시적으로 요구하였으나, +Few-shot 대비 추가적인 성능 향상을 보이지 않았다(QWK 0.719). 오히려 정확도가 소폭 하락하고(68.9%→67.4%), 과대채점 비율이 72.9%로 증가하였다. 이는 진로문장완성검사와 같이 응답이 짧고 맥락 의존적인 과제에서는 명시적 추론 과정의 강제가 오히려 긍정적 방향으로의 과잉해석을 유발할 수 있음을 시사한다.

4-2 프롬프트 구성요소별 기여도 분석

프롬프트 구성요소의 누적적 추가에 따른 성능 변화를 분석한 결과, 각 구성요소의 기여도가 상이하게 나타났다. 표 4는 각 단계별 성능 변화량을 보여준다.

분석 결과, Few-shot 예시 추가가 가장 큰 개선 효과(ΔQWK +0.087)를 보였다. 이는 추상적인 평가 기준보다 구체

표 4. 프롬프트 구성요소별 성능 변화량
Table 4. Performance changes by prompt component

Component added	ΔQWK	ΔAccuracy (%p)	ΔMacro F1
+Rubric (from Baseline)	−0.068	−3.4	−0.042
+Few-shot (from +Rubric)	+0.087	+6.5	+0.096
+CoT (from +Few-shot)	−0.006	−1.5	−0.025

적인 예시가 모델의 판정 정확도 향상에 더 효과적임을 시사한다. 진로문장완성검사와 같이 응답이 짧고 맥락 의존적인 과제에서는 등급별 전형적인 응답 패턴을 예시로 제공하는 것이 모델의 판정 기준 학습에 도움이 되는 것으로 해석된다.

루브릭 추가는 역효과(ΔQWK −0.068)를 나타냈다. 이는 자연어로 기술된 루브릭이 모델의 해석 과정에서 불확실성을 증가시켰을 가능성이 있다. 특히 자기이해가 '구체적'이거나 '긍정적 자아상이 명확'하게 표현된 등의 기준은 경계가 모호하여, 모델이 과도하게 보수적인 판정을 하게 만든 것으로 보인다.

CoT 추론 추가는 유의미한 효과가 없었다(ΔQWK −0.006). 이는 본 연구의 과제가 복잡한 추론보다는 패턴 인식에 가까운 분류 과제이기 때문으로 해석된다. CoT 프롬프팅은 수학 문제 풀이나 논리적 추론이 필요한 과제에서 효과적이지만, 단순 분류 과제에서는 오히려 불필요한 복잡성을 추가할 수 있다.

4-3 등급별 성능 차이

등급별 F1-Score 분석 결과, '하' 등급의 F1이 가장 높고(0.824), '중' 등급의 F1이 가장 낮은(0.592) 것으로 나타났다. 표 5는 +Few-shot 조건에서의 등급별 상세 성능을 보여준다.

'중' 등급의 F1이 낮은 것은 경계적 응답의 판정이 어렵기 때문으로 해석된다. '상'과 '하' 등급은 비교적 명확한 특성을 가지는 반면, '중' 등급은 양쪽의 특성을 부분적으로 포함하거나 모호한 경우가 많아 판정이 어려운 것으로 보인다. 그러나 +Few-shot 조건에서 '중' 등급의 F1이 Baseline 0.431에서 0.592로 개선되어(+0.161), 세 등급 중 가장 큰 향상 폭을 보였다는 점은 주목할 만하다(등급별 개선 폭: High −0.026, Medium +0.161, Low +0.026). 이는 Few-shot 예시가 경계적 응답의 판정 기준 학습에 특히 효과적임을 보여준다.

표 5. +Few-shot 조건의 등급별 성능
Table 5. Performance by grade (+Few-shot condition)

Grade	Precision	Recall	F1-Score
High	0.698	0.640	0.668
Medium	0.544	0.650	0.592
Low	0.876	0.778	0.824

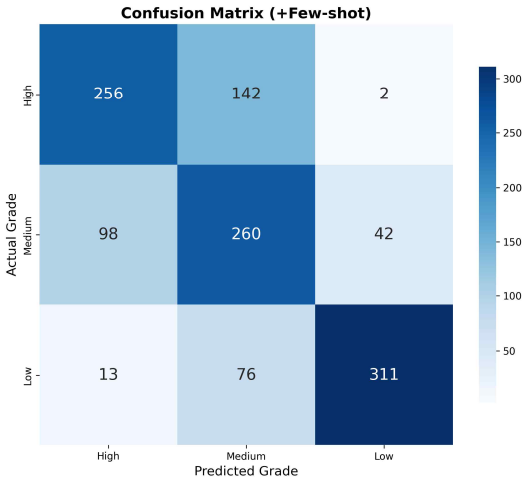

그림 2. +Few-shot 조건의 혼동행렬
Fig. 2. Confusion matrix (+Few-shot condition)

그림 2에서 확인할 수 있듯이, 정분류율은 68.9%(827/1,200)이며, 대각선에 위치한 정분류 건수가 각 등급에서 가장 높게 나타났다. 특히 '하' 등급(311건, 77.8%)의 정분류율이 가장 높았고, '상' 등급(256건, 64.0%)이 가장 낮았다.

오분류 패턴을 분석한 결과, 인접 등급 간 오분류가 전체 오분류의 96%(358/373건)를 차지하였다. 구체적으로 살펴보면, '상'→'중' 오분류가 142건(38.1%), '중'→'상' 오분류가 98건(26.3%), '중'→'하' 오분류가 42건(11.3%), '하'→'중' 오분류가 76건(20.4%)으로 나타났다. 반면 원거리 오분류('상'↔'하')는 극히 드물어(15건, 4.0%), '상'→'하'가 2건, '하'→'상'이 13건에 불과하였다.

이러한 결과는 모델이 등급의 순서적 관계(상>중>하)를 적절히 학습하고 있음을 시사한다. 비록 인접 등급 간의 미세한 차이를 완벽하게 구분하지는 못하더라도, 오분류가 발생하더라도 1단계 차이의 인접 등급으로 분류되는 경향이 강하며, 심각한 오류(2단계 이상 차이)는 거의 발생하지 않음을 보여준다. 이는 실무적 관점에서 중요한 의미를 가지는데, 자동 판정 결과가 전문가 판정과 다소 차이가 있더라도 그 차이가 인접 등급 수준에 머무른다면, 1차 스크리닝 도구로서의 활용 가치는 충분히 확보될 수 있기 때문이다.

4-4 학교급별 성능 차이

표 6과 그림 3에서처럼 고등학교(QWK 0.735)가 중학교(QWK 0.716)보다 다소 높은 성능을 보였다. 두 학교급 간 QWK 차이는 0.019로 크지 않으나, 모든 프롬프트 조건에서 일관되게 고등학교의 성능이 높게 나타났다. 두 학교급 모두에서 +Few-shot 조건이 최고 성능을 기록하였으며, +Rubric 조건의 역효과와 +CoT 조건의 추가 효과 부재가 동일하게 관찰되었다. 이러한 일관된 패턴은 본 연구에서 발견된 프롬프트 전략별 효과가 중등학교 학교급에서 일반화될 수 있음을 시사한다.

고등학교의 성능이 상대적으로 높은 것은 고등학생의 응답이 더 명확하고 구체적인 경향이 있어 판정이 용이했기 때문으로 해석된다. 고등학생은 인지적·언어적 발달 수준이 높아 자신의 생각과 감정을 더욱 정교하게 표현할 수 있으며, 진로에 대한 탐색 경험도 상대적으로 풍부하다. 반면 중학생의 경우 언어 표현력이 상대적으로 제한적이고, 자기이해의 수준도 발달 단계상 차이가 있을 수 있다. 또한 중학생은 진로에 대한 고민이 본격화되기 시작하는 시기로, 응답이 다소 추상적

이거나 불명확한 경우가 많아 자동 판정의 난이도가 높아진 것으로 보인다.

표 6. 학교급별 채점 성능 비교(+Few-shot 조건)
Table 6. Grading performance by school level (+Few-shot)

School level	Accuracy (%)	QWK	Macro F1
Middle school	68.2	0.716	0.689
High school	69.7	0.735	0.700

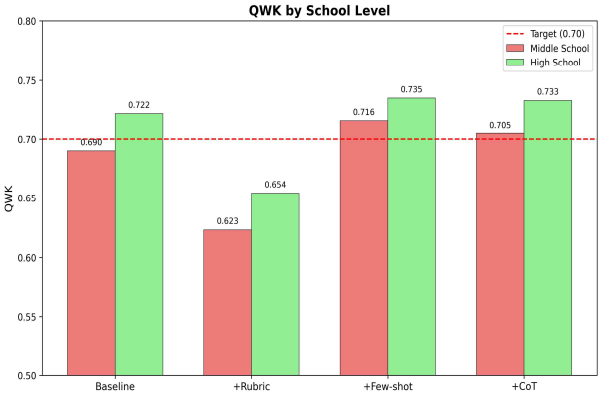

그림 3. +Few-shot 조건의 학교급 QWK 비교
Fig. 3. QWK comparison by school level (+Few-shot condition)

V. 결 론

본 연구는 파인튜닝 없이 프롬프트 엔지니어링만으로 LLM이 진로문장완성검사 응답에 대한 자동 등급 판정을 수행할 수 있음을 실증하였다. 연구 결과, +Few-shot 조건이 정확도 68.9%, QWK 0.725로 가장 높은 성능을 보였으며, 이는 자동 채점 분야의 실용적 수용 기준(QWK≥0.70)을 충족하는 수준이다.

본 연구의 주요 기여는 다음과 같다. 첫째, 프롬프트 구성 요소의 누적적 효과를 체계적으로 분석하여, 루브릭의 명시적 제시가 오히려 역효과를 초래할 수 있다는 점을 규명하였다. 이는 진로문장완성검사와 같이 응답이 짧고 맥락 의존적인 과제에서는 추상적 기준보다 구체적 예시가 더 효과적임을 시사한다. 둘째, Few-shot 예시가 경계적 응답('중' 등급)의 판정에 특히 효과적이라는 점을 확인하였다(F1 개선 +0.161). 이는 모호한 사례의 판정 기준을 학습하는 데 예시 기반 접근이 유용함을 보여준다. 셋째, 프롬프트 전략에 따라 과대/과소채점 패턴이 달라지며, Few-shot 조건이 가장 균형 잡힌 판정을 보인다는 점을 발견하였다.

본 연구의 결과는 진로상담 현장에 다음과 같은 실무적 시사점을 제공한다. LLM 기반 자동 판정 시스템은 진로상담 교사의 업무 부담을 경감하고, 대규모 학생 집단에 대한 진로 평가의 효율성을 높이는데 기여할 수 있다. 특히 프롬프트 설계 시 명시적 예시를 포함하는 것이 판정 정확도 향상에 중요하며, 추상적인 루브릭보다는 구체적인 응답 예시를 제공하는 것이 효과적이다.

다만 본 연구 결과의 현장 적용 시에는 자동 판정이 가지는 윤리적 함의에 대한 신중한 고려가 필요하다. 청소년기 학생에게 평가 결과는 진로 발달과 자아개념 형성에 민감한 영향을 미칠 수 있으며, 부정확한 등급 판정은 왜곡된 자기이해나 불필요한 불안을 유발할 수 있다. 특히 학생에게 자동 판정 결과가 직접 제공될 경우, 잘못된 진로 이해 형성이나 낙인 효과 등의 윤리적 위험이 발생할 수 있다. 따라서 자동 판정 시스템은 전문가의 판단을 대체하는 독립적 진단 도구가 아닌, 진로상담 교사가 다수 학생의 응답을 효율적으로 검토하고 집중 상담 대상을 선별하는 1차 스크리닝 도구로 활용되어야 한다. 학생에게는 자동 판정 결과가 직접 제공되기보다는, 반드시 전문가의 해석과 맥락화 과정을 거친 개별화된 피드백으로 제공되어야 하며, 특히 '중' 등급과 같은 경계적 사례에 대해서는 전문가의 추가 검토가 필수적으로 병행되어야 할 것이다.

본 연구는 다음과 같은 제한점을 가진다. 첫째, 단일 영역('자기이해 및 긍정적 자아상')과 단일 모델(GPT-4o-mini)만을 분석 대상으로 하였다. 후속 연구에서는 다양한 진로성숙도 영역으로 분석을 확장하고, 다양한 LLM 모델 간 비교를 통해 본 연구 결과의 일반화 가능성을 검증할 필요가 있다. 상용 API 모델은 인프라 부담 없이 최신 성능을 활용할 수 있으나 대규모 적용 시 지속적인 비용 발생과 학생 데이터의 외부 서버 전송에 따른 개인정보 보호 상의 제약이 있다. 반면, 오픈소스 모델은 로컬 환경 구축이 가능하다는 장점이 있으나, 높은 초기 구축 비용과 모델별로 프롬프트 민감도나 한국어 처리 능력이 상이할 수 있다. 따라서 LLM 기반 시스템의 교육 현장 도입을 위해서는, 본 연구의 프롬프트 전략이 다양한 모델에서 일관된 효과를 보이는지 검증하고, 예산 및 인프라 여건에 적합한 실용적 모델 선택 기준을 제시하는 연구가 필요할 것이다.

둘째, 본 연구는 등급 판정의 정확도에 초점을 맞추었으나, 실제 교육 현장에서는 판정 결과에 대한 설명 가능성과 피드백 생성 기능도 중요하므로, 이에 대한 후속 연구가 필요하다. 또한 본 연구는 조건 간 성능 차이에 대한 통계적 유의성 검증(예: 부트스트래핑 기반 신뢰구간, 반복 샘플링 등)을 포함하지 못한 한계가 있다. 후속 연구에서는 이러한 방법론적 검증을 통해 결과의 신뢰성을 제고할 필요가 있다.

셋째, 본 연구에서 사용한 AI-Hub 데이터셋에는 초등학생 응답도 포함되어 있으나, 초등학생의 경우 중·고등학생과 평가 기준 및 문항 구성이 상이하여 본 연구의 분석 대상에서 제외하였다. 초등학생은 발달 단계상 언어 표현력이 제한적이고 진로 인식이 형성 초기 단계에 있어, 동일한 평가 기준을 적용하기 어려운 특성이 있다. 그러나 초등학생 대상 진로교육의 중요성이 강조되는 현 시점에서, 초등학생의 발달 특성을 고려한 별도의 평가 기준과 프롬프트 전략을 개발하여 LLM 기반 자동 판정의 적용 가능성을 탐색하는 후속 연구가 필요하다.

부 록

A-1. Baseline

```
## [System Instruction]
당신은 진로상담 전문가입니다.
학생의 진로문장완성검사 응답을 분석하여 진로성숙도 등급을 판정합니다.

## [Input]
문항: {question}
응답: {answer}

## [Output Format]
반드시 다음 JSON 형식으로만 응답하십시오:
{"grade": "상"} 또는 {"grade": "중"} 또는 {"grade": "하"}

## [Hard Constraints]
1. 등급은 반드시 {상, 중, 하} 중 하나만 선택한다.
2. JSON 형식 외의 텍스트를 출력하지 않는다.
```

A-2. +Rubric (additional component)

```
## [Scoring Criteria]
상: 자신의 특성, 강점, 가치관을 구체적으로 표현함.
    긍정적 자아상이 명확히 드러남.
중: 자신의 특성을 표현하나 추상적이거나 단순함.
    중립적 표현.
하: 자신의 특성을 인식하지 못하거나 부정적 자아상.
    무응답/회피.

***짧은 응답이라도 긍정적이고 구체적이면 '상'으로 판정***
***단순하지만 적절한 응답은 '중'으로 판정***
```

A-3. +Few-shot (additional component)

```
## [Few-shot Examples]
문항: 부모님은 나를 | 응답: 사랑한다 | 등급: 상
문항: 내가 생각하는 나의 능력은 | 응답: 끈기가 있다 | 등급: 상
문항: 내가 가장 좋아하는 활동은 | 응답: 게임이다 | 등급: 중
문항: 내가 삶에서 가장 중요한 것은 | 응답: 건강 | 등급: 중
문항: 내가 가장 잘하는 것은 | 응답: 모르겠다 | 등급: 하
문항: 내가 생각하기에 나는 | 응답: 게으르다 | 등급: 하
```

A-4. +CoT (additional component)

```
## [Task]
다음 단계에 따라 판정하세요:
1. 응답에서 자기인식의 구체성과 긍정성을 분석한다.
2. Scoring Criteria와 비교하여 가장 적합한 등급을 선택한다.
3. 판정 근거와 등급을 출력한다.

## [Output Format]
{"reason": "판정 근거 1-2문장", "grade": "상/중/하"}
```

* The prompt is shown in Korean to preserve linguistic and contextual fidelity, as the experiment evaluates a Korean-language model using native Korean counseling data.

그림 4. 실험 조건별 프롬프트 구조

Fig. 4. Prompt structures across experimental conditions

감사의 글

본 연구는 제1저자의 석사학위논문을 수정·보완한 것임.

참고문헌

[1] Electronic Times. With High School Credit System Implementation, Only One Career Counseling Teacher? Cases Where One Teacher Handles Career Counseling for 200 Students [Internet]. Available: https://m.etnews.com/20250520000205.

[2] J. B. Rotter and J. E. Rafferty, *Manual for the Rotter Incomplete Sentences Blank*, New York, NY: Psychological Corporation, 1950.

[3] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, ... and X. Xie, "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology,* Vol. 15, No. 3, March 2024. https://doi.org/10.1145/3641289

[4] B. H. Kim and S. A. Lee, "Automatic Analysis and Feedback Generation of Career Maturity Test Open-Ended Responses Using Large Language Models," in *Proceedings of the Korea Information Processing Society Conference*, Seoul, pp. 2242-2245, 2025.

[5] L. Jiang and N. Bosch, "Short Answer Scoring with GPT-4," in *Proceedings of the 11th ACM Conference on Learning @ Scale (L@S '24)*, Atlanta: GA, pp. 438-442, July 2024. https://doi.org/10.1145/3657604.3664685

[6] H. R. Kang and J. P. Baek, "Automatic Scoring of Korean Speaking Accuracy Using Generative AI: Focusing on Vocabulary and Grammar Accuracy," *Urimalgeul,* Vol. 107, pp. 311-332, December 2025. https://doi.org/10.18628/urimal.107..202512.311

[7] E. Mayfield and A. W. Black, "Should You Fine-Tune BERT for Automated Essay Scoring?," in *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications,* Seattle: WA, pp. 151-162, 2020. https://doi.org/10.18653/v1/2020.bea-1.15

[8] T. A. N. F. Eneye, C. F. Ijezue, M. Amjad, A. I. Amjad, S. Butt, and G. Castañeda-Garza, "Advances in Auto-Grading with Large Language Models: A Cross-Disciplinary Survey," in *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications,* Vienna, Austria, pp. 477-498, July 2025. http://doi.org/10.18653/v1/2025.bea-1.35

[9] D. Carpenter, W. Min, S. Lee, G. Ozogul, X. Zheng, and J. Lester, "Assessing Student Explanations with Large Language Models Using Fine-Tuning and Few-Shot Learning," in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications,* Mexico City, Mexico, pp. 403-413, 2024.

[10] F. Yavuz, Ö. Çelik, and G. Y. Çelik, "Utilizing Large Language Models for EFL Essay Grading: An Examination of Reliability and Validity in Rubric-Based Assessments," *British Journal of Educational Technology,* Vol. 56, No. 1, pp. 150-166, 2025. https://doi.org/10.1111/bjet.13494

[11] X. Bronlet, "Leveraging on Large Language Model to

Classify Sentences: A Case Study Applying STAGES Scoring Methodology for Sentence Completion Test on Ego Development," *Frontiers in Psychology,* Vol. 16, 1488102, 2025. https://doi.org/10.3389/fpsyg.2025.1488102

[12] Y. Huang, M. Lei, H. Zhang, L. Zong, B. Zhu, and H. Luo, "An Intelligent Agent for Sentence Completion Test: Creation and Application in Depression Assessment," *Frontiers in Psychology,* Vol. 16, 1649905, 2025. https://doi.org/10.3389/fpsyg.2025.1649905

[13] S. Lee and K.-S. Min, "Automated Quantification of Career-Maturity Sentence-Completion Test Responses Using GPT-3.5 and Application of Item Response Theory," *Journal of Vocational Education and Training,* Vol. 28, No. 2, pp. 63-81, 2025. https://doi.org/10.36907/krivet.2025.28.2.63

[14] W. T. Lee and S. Y. Yoon, "A Study on LLM-Based Chatbot Design for Adolescent Career Counseling," in *Proceedings of the 2025 Summer Conference of the Korean Institute of Communications and Information Sciences,* pp. 638-639, 2025.

[15] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large Language Models Are Zero-Shot Reasoners," in *Advances in Neural Information Processing Systems*, Vol. 35, pp. 22199-22213, 2022. https://doi.org/10.48550/arXiv.2205.11916

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, ... and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24824-24837, 2022. https://doi.org/10.48550/arXiv.2201.11903

[17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, ... and D. Amodei, "Language Models Are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877-1901, 2020. https://doi.org/10.48550/arXiv.2005.14165

[18] S. W. Park, Y. H. Kim, J. R. Moon, and W. S. Lim, "Performance Analysis of Zero-Shot Prompting Strategies for Korean Language Inference," in *Proceedings of the 6th Korea Artificial Intelligence Conference*, Seoul, pp. 10-15, 2025.

[19] M. D. Shermis, "State-of-the-Art Automated Essay Scoring: Competition, Results, and Future Directions from a United States Demonstration," *Assessing Writing*, Vol. 20, pp. 53-76, 2014. https://doi.org/10.1016/j.asw.2013.04.001

최예슬(Yeseul Choi)

2009년 : 서울대학교(교육학 학사)
2011년 : 서울대학교(HRD 석사)
2019년 : University of Wisconsin-Madison (교육정책 박사)
2026년 : 서울과학종합대학원대학교 (AI·빅데이터 공학석사)
2022년～2024년: 한국교육개발원
2024년～현 재: 전북대학교 교육학과 조교수
※관심분야 : 인공지능(AI)과 교육

오태연(Taeyeon Oh)

2008년 : 서강대학교(수학사)
2012년 : 서강대학교(경제학석사)
2016년 : 서울대학교(스포츠경영학박사)
2018년～2022년: University of Mississippi
2022년～현 재: 서울과학종합대학원대학교 AI첨단학과 조교수
※관심분야 : 인공지능 데이터 분석