JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 7, pp. 87-96, July 2025
https://doi.org/10.9708/jksci.2025.30.07.087

Design and Evaluation of LLM-Based Conversational Agents for Diagnosing Science Learning Difficulties

Tae-Ho Min*

*Teacher, Bora Middle School, Yongin, Korea

[Abstract]

This study explores the use of large language model-based conversational agents to diagnose science learning difficulties. Four prototype models were developed by varying the classification schemes of science learning difficulties and the inclusion of emotional support, and simulation-based experiments were conducted by using ChatGPT-4o in the roles of students. The results showed that models using detailed classifications of environmental factors achieved higher diagnostic accuracy than those using detailed classifications of individual factors, while the inclusion of emotional support did not have a statistically significant effect. Based on these findings, an improved model was designed and applied to twelve students in the ninth grade. Analysis of the interactions revealed that certain areas, including community-related factors, were sometimes omitted from the diagnosis. In addition, the diagnostic process was prematurely terminated when students deviated from the topic or refused to continue the conversation. Drawing on these results, the study proposes design recommendations for conversational agents aimed at diagnosing science learning difficulties.

▸ **Key words**: Large language model, Generative AI, Conversational agent, Prompt engineering, Science education, Learner diagnosis

[요 약]

본 연구는 대형 언어 모델 기반 대화형 에이전트를 활용해 과학 학습의 어려움을 진단하는 방안을 탐색했다. 과학 학습의 어려움 분류 방식과 정서적 지지 여부를 달리한 4개 프로토타입 모델을 개발하고, ChatGPT 4o에 학생 역할을 부여해 시뮬레이션 실험을 수행했다. 실험 결과, 환경적 요인을 세분화한 모델이 개인 요인을 세분화한 모델보다 더 높은 진단 정확도를 보였으며, 정서적 지지의 포함 여부는 유의미한 영향을 미치지 않았다. 이를 바탕으로 설계한 개선 모델을 중학교 3학년 학생 12명에게 적용하고 대화 내용을 분석한 결과, 지역사회 요인을 포함한 일부 영역의 진단이 누락됨을 확인했다. 또한 학생이 주제를 이탈하거나 대화를 회피할 경우 진단을 조기에 종료하는 문제가 나타났다. 본 연구의 결과로부터 과학 학습의 어려움을 진단하기 위한 대화형 에이전트의 설계 방향을 제안했다.

▸ **주제어**: 대형 언어 모델, 생성형 AI, 대화형 에이전트, 프롬프트 엔지니어링, 과학교육, 학습자 진단

• First Author: Tae-Ho Min, Corresponding Author: Tae-Ho Min
*Tae-Ho Min (ty3370@dankook.ac.kr), Bora Middle School
• Received: 2025. 05. 22, Revised: 2025. 07. 02, Accepted: 2025. 07. 02.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

우리나라 교육 환경은 입시 중심의 구조와 다인수 학급 운영이라는 특성을 보이며, 이러한 맥락은 개별 학습자의 수준과 필요를 충분히 반영하지 못하는 한계를 가진다[1-2]. 학생 개개인이 학습 과정에서 겪는 어려움을 교사가 세부적으로 파악하고 지원하는 것이 불가능에 가깝기 때문이다. 특히 많은 인원이 함께 수업을 듣는 환경에서는 개별 피드백 및 평가의 어려움, 교사의 지도 효과 감소, 학생 참여 부족 등의 문제가 발생할 수 있다[3]. 이는 교육 불평등을 심화하고 소외된 학생 집단의 성취도에 부정적 영향을 미친다[4].

이러한 문제는 과학교육에서 더욱 두드러지게 나타날 수 있다. 과학은 위계성이 강한 교과로, 하위 개념의 이해 여부가 상위 개념의 이해에 영향을 미치기 때문이다[5]. 예를 들어 물리학은 단순 암기보다 개념 이해와 문제 해결 전략이 필요하므로, 학습 결손이 누적되기 쉬운 구조를 가진다[6-7]. 위계적 구조가 뚜렷한 과목에서는 학습에서 경험하는 어려움이 지속될 가능성이 크기 때문에 더 세밀한 관찰과 진단이 요구된다[8]. 선행 연구에 따르면 과학 성취 기준 도달 실패는 위계적인 개념 구조와 더불어 가정, 학교, 지역사회 등의 다양한 요인이 복합적으로 작용한 결과이다[9-10]. 따라서 학생이 과학 학습에서 겪는 어려움을 다각도에서 진단하고 이해하는 노력이 필요하다.

이 같은 배경에서 과학 학습의 어려움을 파악하고 분류하기 위한 연구가 수행되었다. Amaliyah et al.[11]은 7학년 학생 대상으로 과학 학습에서 겪는 어려움을 설문과 인터뷰, 문서 자료로 분석하고, 이를 내부 요인과 외부 요인으로 분류했다. 내부 요인은 학생 개인의 심리적·생리적 특성과 관련되며, 흥미와 재능의 불일치, 낮은 학습 동기, 감정 조절의 어려움, 건강 문제 등이 포함되었다. 외부 요인은 학생을 둘러싼 환경과 관련되며, 가정의 관심 부족이나 경제적 제약, 수업 자료와 교수법의 한계, 열악한 교실 환경, 소셜미디어의 과다 사용 등이 학습에 부정적 영향을 미쳤다. Novianti et al.[12]도 이와 유사하게 학생의 건강 상태, 학습 동기 등의 내부 요인과, 가정 및 지역사회, 수업 방식 등 외부 요인이 복합적으로 작용해 과학 학습에 어려움을 초래함을 확인했다.

일부 연구에서는 과학 학습에서 겪는 내적·외적 어려움을 진단하기 위한 검사 도구를 개발하기도 했다[9, 13]. 이러한 도구들은 표준화된 문항을 통해 일정한 신뢰도와 타당도를 확보하려는 목적에서 유용하게 활용되고 있으나, 학습자의 정서적 상태나 학습 과정에서의 인지 전략 등 깊이 있는 정보를 파악하는 데는 한계가 있다. 이를 보완하기 위해 면담 방식의 진단이 병행되기도 하지만[14], 면담은 시간과 인력 측면에서 부담이 커 실제 교육 현장에서의 활용은 제한적이다. 최근 주목받고 있는 기술적 가능성 중 하나는 대형 언어 모델(Large Language Model, 이하 LLM)을 활용한 진단 방식이다. LLM은 자연어 처리 기반의 대화 기능으로 학습자의 반응을 유도하고, 이를 분석함으로써 정교하고 유연한 진단을 수행할 수 있는 잠재력을 가진다[15].

본 연구에서는 이러한 기술적 가능성을 바탕으로, 과학 학습에서 겪는 어려움을 진단하는 도구로써 LLM의 활용 가능성을 확인하고자 하였다. 이를 위해 먼저 사용자와의 대화를 통해 과학 학습의 어려움을 진단하는 4개의 LLM 기반 대화형 에이전트를 개발했다. 각 에이전트의 초기 프롬프트는 진단 및 상호작용 방식이 서로 다르도록 작성되었다. 4개의 에이전트를 학생 역할이 부여된 ChatGPT 4o 세션과 대화하게 하여 진단 정확도가 높은 방식을 확인했다. 이 결과를 바탕으로 실제 학생에게 투입할 대화형 에이전트를 설계하고, 중학생 12명에게 적용해 대화 내용을 수집·분석했다. 본 연구의 연구 문제는 다음과 같다.

1. 초기 프롬프트 설계 방식은 LLM의 과학 학습 어려움 진단 정확도에 어떤 영향을 미치는가?

2. 실제 학생과 LLM의 상호작용에서 관찰되는 과학 학습 어려움 진단의 개선점은 무엇인가?

II. Preliminaries

1. Factors in science learning difficulties

학습자는 인지적·정의적 요인 등의 내적 요인, 가정·학교·지역사회 요인 등의 외적 요인으로 인해 과학 학습에 어려움을 겪는다[11-12]. 선행 연구에서는 과학 학습에서 어려움을 겪는 원인을 몇 가지로 범주화해 제시했다. 예를 들어 이상일[9], 이민애와 박윤배[16], 김상윤 등[17]은 과학 학습에서 겪는 문제를 인지적 요인, 정의적 요인, 환경적 요인으로 구분했다. 인지적 요인은 학습자가 과학 개념을 이해하고 문제를 해결하는 데 필요한 인지적 능력의 결핍에서 비롯된다. 선행 학습의 누적 결손, 과학 개념에 대한 오개념, 언어적 표현 능력 부족, 수리력과 사고력 저하 등으로 나타나며, 과학 용어의 의미를 정확히 이해하지 못하거나, 실험 결과를 논리적으로 해석하지 못하는 경향이 있었다. 정의적 요인은 학습에 대한 태도, 정서, 동기와 관련된 요인으로, 학습자가 학업에 대해 느끼는 정서적 반응

과 자기 인식에서 비롯된다. 학습에 대한 흥미 부족, 낮은 자아개념, 성취동기의 결여, 주의 집중력의 부족, 불안과 열등감 등의 형태로 나타나며, 학습 실패의 경험이 반복됨에 따라 자포자기적 태도를 보이기도 한다. 환경적 요인은 학습자가 처한 외적 환경이 학업 성취에 영향을 미치는 경우다. 가정환경, 교사와의 관계, 또래와의 상호작용, 학습 환경의 질 등이 포함된다. 다문화가정 학생의 언어적 소통 문제, 부모의 무관심 혹은 지나친 기대, 불안정한 가정 구조, 또래 관계에서의 갈등, 교사와의 부정적 상호작용 등이 환경적 요인으로 확인되었다.

일부 연구는 인지적 요인에 초점을 두고 과학 학습의 어려움을 분석했다. 임준홍과 이봉우[18-19]는 물리의 역학 학습에서 겪는 어려움을 개념 지식과 과정 지식의 측면으로 분류했다. 개념 지식과 관련해서는 선수 지식의 부족, 개념에 대한 불완전한 이해, 공식의 의미를 모르는 상태에서 기계적으로 적용하려는 경향 등이 주요 원인으로 확인되었다. 과정 지식 측면에서는 상호작용 법칙의 적용, 변수 간 관계 분석, 조건 해석, 시각적 정보(그래프, 그림)의 이해와 변환 능력 등의 부족이 주요 문제로 나타났다. 이경희 등[13]은 과학 저성취의 원인을 인지적 요인과 정의적 요인 중심으로 분석하여 과학 탐구 능력 부족, 학습전략 부족, 학습 동기 부족의 세 가지 요인으로 구분했다.

한편, 해외 연구는 환경적 요인을 더 세분화하는 경향이 있었다. Banerjee[10]는 사회적 약자가 과학 및 수학 학습에서 겪는 문제를 개인, 가정, 학교, 지역사회 수준으로 분류했다. 개인 요인의 경우 빈곤층 가정에서 자란 아동은 부정적 정서, 낮은 자존감, 낮은 학습 동기를 보이는 경향이 있으며, 이는 과학 학습 과정에서 지속적인 어려움으로 이어졌다. 가정 요인으로는 부모의 낮은 학력, 교육 참여 부족, 권위 있는 양육 태도의 부재, 십대 부모, 부모 수감 경험 등이 학생의 학습에 부정적인 영향을 미쳤다. 학교와 교사 요인에 있어선 교사의 낮은 기대, 부정적 피드백, 차별적 대우, 학교 내 부정적 분위기 등이 과학 학습에 문제를 유발했다. 지역사회 수준에서는 저소득 지역의 열악한 교육 환경, 학습 롤모델의 부재, 높은 범죄율 및 고위험 행동 노출 등이 학생들에게 안정적인 학습 기회를 제공하지 못하여 학습의 어려움을 초래했다. Chere & Hlalele[20]도 이와 유사하게 개인, 가족, 학교라는 세 가지 요인에 초점을 두고 학습 문제를 설명했다.

선행 연구들은 인지적, 정의적, 환경적 요인들의 복합적인 상호작용으로 과학 학습의 어려움이 발생한다는 점에 주목하며, 개인, 가정, 학교의 측면을 포괄적으로 다루었다. 그러나 세부적인 요인이나 강조점에는 차이가 있었다. 국내 연구들은 주로 학생 개개인의 학습 행동에 초점을 두었으나, 해외 연구들은 사회·경제적 환경과 같은 거시적 요인까지 폭넓게 고려했다. 이는 국내외의 교육 환경과 사회적 맥락 차이에서 비롯된 것으로 보인다. 한국은 상대적으로 균질한 교육 환경을 전제로 개인 및 학교 요인에 집중한 반면, 해외는 지역 간 교육 여건의 차이나 계층적 격차 등이 비교적 커 구조적 요인까지 함께 고려했을 수 있다.

2. LLM-based advisor for learner diagnosis and counseling

2022년 11월 OpenAI가 ChatGPT를 발표한 이후, LLM 기반의 대화형 에이전트를 활용하여 학습자의 개별적 특성과 학습 관련 요구를 진단하려는 시도가 꾸준히 이루어지고 있다. Lekan & Pardos[21]는 GPT-4를 전공 상담에 적용해 학생의 관심사, 성향, 진로 목표 등을 바탕으로 개인화된 전공 추천을 생성하고, 이를 인간 상담자와 비교·평가하는 방식을 통해 학습자 진단의 실효성을 탐색했다. 연구 결과 GPT-4는 학생이 제공한 정보를 충분히 반영하여 논리적이고 설득력 있는 피드백을 생성함을 확인했다. 그러나 일부 상담자들은 GPT-4의 진단이 맥락적 정교함이나 상호작용 면에서 한계를 가지며, 일방적 제안 중심이라는 점을 지적했다. 이러한 결과는 LLM 기반 진단 도구가 학습자의 개별 상황과 정서적 반응을 충분히 고려하지 못했음을 시사한다.

Huang et al.[22]은 LLM을 포함한 챗봇 상담자가 학습자의 동기를 어떻게 촉진하는지에 대해 43편의 선행 연구를 검토했다. 연구 결과는 챗봇 활용이 개별화된 피드백을 제공하고, 학습자의 자율성과 참여를 유도함으로써 몰입을 높이며, 최근의 LLM 기반 챗봇은 자연스러운 언어 이해와 감정적 반응까지 가능하다는 장점을 보여주었다. 그러나 챗봇이 맥락에 대한 정교한 이해가 부족하다는 점, 인간 상담자에 비해 정서적 반응의 깊이가 떨어진다는 점, 과도한 자율성 부여가 오히려 학습자의 방향 상실로 이어질 수 있다는 점 등의 한계도 지적되었다.

이와 유사하게 Park et al.[23]은 LLM 기반의 청소년의 정신 건강 상담 대화형 에이전트가 정서적 공감과 맞춤형 피드백 제공이 가능함을 확인했으나, 맥락에 대한 이해 부족, 정서적 깊이의 한계, 안전성 검증 미비 등의 문제를 안고 있어 교육적 활용에는 신중한 접근이 필요하다고 주장했다. Ramandanis & Xinogalos[24] 또한 학습자 진단 및 지원을 위한 챗봇이 맞춤형 피드백 제공, 학습 동기 촉진, 반복적 질의응답을 통한 심화된 자기 인식 유도 등의 장점이 있으나, 학습자의 정서 상태나 맥락적 상황을 충분

히 고려하지 못하는 한계, 정보 정확성의 불확실성, 인간 상담자에 비해 낮은 상호작용의 깊이 등의 문제점도 있음을 확인했다.

Meyer & Elsweiler[25]는 LLM 기반 상담 시스템의 실제 효과와 한계를 알아보고자 동기 강화 상담(Motivational Interviewing) 기반의 프롬프트를 적용한 GPT-4 대화형 에이전트 MIcha를 개발하고, 사용자의 학습적 자기성찰을 유도하는 대조실험을 수행했다. 연구 결과 LLM은 자유로운 언어 표현과 텍스트 이해 능력을 활용해 사용자가 자신의 학습 동기나 어려움을 구체적 언어로 풀어내도록 유도할 수 있었다. 그러나 동시에 LLM이 상담 맥락에서 감정 표현이나 조언, 사실 제공과 같은 발화를 생성할 경우 사용자에게 혼란을 주거나 잘못된 인식을 심어줄 수 있음을 지적했다.

이상의 선행 연구들은 LLM 기반 학습 상담 및 진단 시스템이 개별화된 피드백 제공, 학습 동기 촉진, 자기성찰 유도 등에서 가능성을 보여준다고 평가한다. 그러나 동시에 정서적 반응의 깊이 부족, 맥락에 대한 피상적 이해, 상호작용의 일방향성 등의 한계가 언급되었다. 특히 대부분 연구에서 LLM 기반 대화형 에이전트가 정서적 공감과 감정적 대화 자체는 수행함에도 그 깊이가 제한적이라고 지적한 점은 주목할 필요가 있다. LLM은 사용자의 감정을 인식하고 공감하는 언어적 반응을 생성할 수 있지만, 이는 실제 감정 경험 없이 패턴 기반으로 구성된 인지적 공감 수준에 국한된다[26]. 따라서 LLM 기반의 학습자 진단 대화형 에이전트 개발에서는 공감의 한계를 보완하는 방안에 대한 고려가 필요하다.

III. The Proposed Scheme

본 연구의 목적은 과학 학습의 어려움을 진단하는 LLM 기반 대화형 에이전트를 개발하고 평가하는 것이다. 이를 위해 프로토타입 모델 개발 및 시뮬레이션 실험, 개선 모델 개발 및 학생 대상 투입, 학생과 개선 모델의 대화 분석 과정으로 연구를 진행했다.

1. Design of the prototype models and simulation-based experiments

실제 학생 대상 연구에 앞서 대화형 에이전트의 설계 방향을 도출하고자 시뮬레이션 실험을 수행했다. 구체적으로는 선행 문헌을 바탕으로 서로 다른 초기 프롬프트가 입력된 4개의 프로토타입 모델을 개발하고, 학생 역할이 부여된 ChatGPT 4o 세션과의 대화를 통해 각 모델의 진단 정확도를 비교했다.

4개의 프로토타입 모델은 두 가지 측면에서 차이가 있었다. 첫 번째 측면은 과학 학습에서 겪는 어려움을 분류하는 방식이다. 국내 문헌에서는 주로 인지적, 정의적, 환경적 요인으로 분류했고[9, 16-17], 해외 연구는 개인, 가족, 학교 등의 요인으로 분류했다[10, 20]. 국내에서는 해외의 개인 요인을 세분화하고, 해외에서는 국내의 환경적 요인을 세분화했다고 볼 수 있다. 이러한 분류 방식에 따라 진단 정확도 차이가 발생하는지 확인하기 위해, 과학 학습에서 겪는 어려움을 인지적, 정의적, 환경적 요인으로 분류해 초기 프롬프트를 작성한 모델과, 개인, 가족, 학교 요인으로 분류해 초기 프롬프트를 작성한 모델을 비교했다.

초기 프롬프트 작성의 또 다른 측면은 정서적 지지 여부였다. LLM이 생성하는 정서적 반응은 피상적 수준에 머물고 있음이 여러 선행 연구에서 지적되었다[21-26]. 초기 프롬프트에 명시적으로 정서적 지지를 요구하는 내용을 포함할 때 사용자와의 상호작용 및 진단 정확도에 변화가 있는지 확인했다.

이상의 두 가지 측면을 기준으로 시뮬레이션 실험을 위한 4개의 프로토타입 모델 M1-M4를 개발했다(Table 1 참조). M1과 M2 모델은 과학 학습의 어려움을 인지적, 정의적, 환경적 요인으로 나누어 진단하고, M3와 M4 모델은 개인, 가족, 학교 요인으로 나누어 진단했다. M1, M3 모델은 정서적 공감과 지지를 포함해 응답하라는 지시가 초기 프롬프트에 명시적으로 포함되었고, M2, M4 모델의 초기 프롬프트에는 이러한 내용이 없었다.

Table 1. Initial Prompts for the Four Prototype Models

Models	Factors of Learning Difficulties	Emotional Support
M1	• Cognitive • Affective • Environmental	Yes
M2		No
M3	• Individual • Family • School	Yes
M4		No

프로토타입 모델은 gpt-4.1을 기반으로 Streamlit 환경에서 작동하는 웹 애플리케이션 형태로 개발됐다. gpt 계열 LLM은 Gemini, Llama, Mixtral 등 다른 모델들에 비해 텍스트에 표현된 감정의 방향성과 강도 분석 능력이 우수했기에 본 연구의 맥락에 적합하다고 판단했다[27]. 웹 애플리케이션은 총 3개의 페이지로 구성되었다. 첫 페이지에서 학번, 이름 등 사용자의 기초 정보를 입력한 뒤 [다

음] 버튼을 누르면 Fig. 1과 같이 에이전트와 대화하며 과학 학습의 어려움을 진단하는 두 번째 페이지가 제시되었다. 사용자 편의를 위해 최근 대화는 상단에 배치했다.

Fig. 1. Web Interface Featuring the Conversational Agent

두 번째 페이지에서 에이전트는 사용자가 과학 학습에 어떤 어려움을 겪는지 대화하며 그 원인을 진단하고, 진단을 마치면 [다음] 버튼을 누르라고 안내했다. 사용자가 [다음] 버튼을 누르면 대화 내용을 바탕으로 세 번째 페이지에 진단 결과를 출력했다. [다음] 버튼을 누를 때 모든 대화 기록이 MySQL 데이터베이스에 저장되었으며, 저장된 대화는 연구자가 별도의 웹에서 확인할 수 있도록 구현했다. 이 외에도 초기 프롬프트에는 대화를 시작할 때 과학에 관한 전반적인 인식을 파악할 것, 과학 학습에서 겪는 어려움을 단계적으로 심화하여 질문할 것, 한 번에 하나의 사항만 질문할 것 등의 내용이 포함되었다.

M1-M4 모델의 진단 정확도를 비교하기 위해 학생 역할을 부여한 ChatGPT 4o 세션과 대화하는 시뮬레이션 실험을 수행했다. LLM에 사용자 역할을 부여해 시뮬레이션하는 방법은 대량의 데이터나 복잡한 규칙 설계 없이도 다양한 상황에 대응하는 일반화된 대화 시스템을 효과적으로 평가할 수 있어 새로운 대화 시스템 평가 방법으로 주목받고 있다[28]. ChatGPT 4o 세션에는 과학 학습의 어려움을 유발하는 5개의 요인을 가진 학생 역할이 부여되어, M1-M4의 프로토타입 모델과 각각 대화했다. 설정된 과학 학습의 어려움 요인은 개인, 가족, 학교 또는 인지적, 정의적, 환경적 요인 등 다양한 유형을 포함했다. 여러 측면에서의 비교를 위해, 서로 다른 과학 학습의 어려움 요인을 가진 다섯 종류의 학생 역할 세션을 구성했다. 또한 ChatGPT 4o는 대화를 지나치게 길게 생성하는 경향이 있어, 실제 학생처럼 대화하도록 유도하기 위해 모든 대화를 한 줄 이내로 생성하라는 지시를 포함했다. Table 2는 ChatGPT 4o에 부여한 학생 역할 세션 S1-S5를 요약한 표이다.

M1-M4의 프로토타입 모델은 학생 역할 세션 S1-S5의 과학 학습 어려움을 모두 진단하기까지 대화를 수행했다. 결과의 신뢰도 향상을 위해, 모든 프로토타입 모델은 모든 학생 역할 세션에 대해 각각 2번씩 진단했다. 즉 하나의 프로토타입 모델은 학생 역할 세션 S1-S5를 2번씩 총 10회 진단했으며, 하나의 학생 역할 세션은 과학 학습의 어려움 요인을 5개씩 가지고 있으므로, 하나의 프로토타입 모델이 진단해야 하는 요인은 50개였다.

진단 정확도 산출을 위해, 각 프로토타입 모델의 진단 결과가 사전에 설정된 과학 학습의 어려움 요인 중 몇 개를 포함하는지 확인했다. 예를 들어 S1 세션에는 “교사가 빠른 속도로 수업을 진행하고, 질문하기 어려운 분위기임.”이라는 과학 학습의 어려움 요인이 있었는데, M1 모델의 진단 결과에는 “학교는 설명이 빠르고 질문하기 어려운 분위기”라는 내용이 포함되어 있었다. 이와 같이 의미상으로 사전 설정된 어려움 요인과 일치하는 진단 결과가 출력된 경우, 해당 요인을 정확히 진단한 것으로 간주했다.

Table 2. Five Simulated Student Roles Assigned to ChatGPT 4o

Session	Role Type	Science Learning Difficulties
S1	Underserved Region	Lacks Conceptual Foundation; Holds Fixed Mindset about Science; Discouraged from Asking Questions; Receives Little Parental Support; Lacks Access to Resources
S2	School Maladjustment	Unstable Home Life; Social Alienation at School; Disengaged Behavior in Class; Missing Basic Knowledge; Minimal Parental Involvement
S3	Academic Stress	Poor Language Comprehension; Fear of Failure; Low Engagement due to Rote-focused Lessons; Excessive Parental Pressure; Reduced Focus from Stress
S4	Arts-oriented Learner	Stronger Interest in the Arts; Difficulty with Abstract Concepts; Arts-focused Family Support; Perceived Teacher Neglect; Low Confidence in Science
S5	Poor Learning Strategies	Low Motivation from Poor Performance; Lacks Summarizing Skills; Passive Learning Attitude; Relies on Memorization; No Academic Support at Home

진단 정확도는 각 프로토타입 모델이 진단해야 할 50개의 요인 중 정확히 진단한 요인의 비율로 계산되었다. 또한 사전 설정된 요인들을 인지적·정의적·환경적 요인과 개인·가족·학교 요인으로 분류해 별도로 진단 정확도를 산출했다.

Table 3는 M1-M4의 진단 정확도를 비교한 표이다. 진단 정확도 차이가 통계적으로 유의한지 확인하기 위해 ANOVA 분석을 수행했다. 전체 진단 정확도는 M3가 .62로 가장 높았고, .60의 M4 모델이 뒤를 이었다. M1, M2 모델의 진단 정확도는 각각 .48과 .40으로 비교적 낮았다. 그러나 그 차이는 통계적으로 유의하지 않았다.

세부 요인별 진단 정확도를 살펴보면, 환경적 요인과 학교 요인에 대한 진단 정확도 차이가 통계적으로 유의했다 ($p < .05$). 두 요인 모두 M3와 M4 모델의 정확도가 높았다. M1과 M2 모델은 과학 학습에서 겪는 개인적인 어려움을 인지적, 정의적 요인으로 세분화해 진단한 모델이었고, M3와 M4 모델은 환경적인 어려움을 가족, 학교 요인으로 세분화해 진단한 모델이었다. M1과 M2 모델은 개인 요인들에 대해 전반적으로 진단 정확도가 높긴 했으나 통계적으로 유의한 수준은 아니었다. 반면 M3와 M4 모델은 환경적 요인들에 대해 통계적으로 유의하게 높은 진단 정확도를 보였다. 개인 요인의 세부 요인이라고 할 수 있는 인지적, 정의적 요인은 서로 밀접한 관련이 있어 모델별로 큰 차이를 만들지는 않았으나, 환경적 요인은 가족, 학교, 지역사회 등 다양하므로[10], 환경적 요인에 대한 차이가 더 컸을 것으로 해석된다.

정서적 공감과 지지를 포함한 응답을 생성했던 M1, M3 모델의 전체 진단 정확도는(각각 .48과 .62) 간결한 응답을 생성한 M2, M4 모델보다(각각 .40, .60) 약간 높았다. 그러나 과학 학습의 어려움 분류 방식에 따른 차이만큼 크지는 않았으며, 일부 요인에 대해서는 간결한 응답을 생성한 모델의 진단 정확도가 더 높은 사례도 있었다. 예컨대 가족 요인에 대한 진단 정확도는 M3 모델이 .57, M4 모델이 .79로 산출되어 정서적 지지를 포함하지 않은 M4 모델에서 더 높았다.

실제 대화 내용을 살펴보면 정서적 지지 없이 간결한 답변만 제공하도록 설계된 M2와 M4 모델에서도 정서적 공감과 지지를 제공하는 대화가 다수 발견되었다. S2 세션이 “기초부터 잘 몰라서 선생님 말이 무슨 소리인지 잘 모르겠어요.”라고 말했을 때, M4 모델은 “기초 개념을 잘 모른다는 게 불안할 수도 있겠다. 그래서 선생님 설명도 잘 이해가 안 되고, 자연스럽게 흥미나 집중이 떨어지는 것 같구나. 혹시 과학 공부를 할 때 집중이 잘 안 되거나, 학습 의욕이 잘 생기지 않는 편이야?”라고 응답하며 공감과 원인 탐색을 동시에 수행했다.

정서적 지지를 포함하도록 명시한 M1과 M3 모델은 M2, M4 모델보다 풍부한 양의 공감과 지지를 응답에 포함했다. 그러나 응답의 내용이 형식적이고 반복적인 경향을 보여 심화된 대화를 유도하는 데 한계가 있었다. 예를 들어 M3 모델은 S2 세션과의 대화에서 연속된 3개 응답의 첫 문장을 “네가 이렇게 솔직하게 말해줘서 정말 고마워.”, “네가 물질의 성질과 관련된 용어나 개념이 헷갈린다고 말해줘서 고마워.”, “네가 밀도에서 특히 어렵다고 느낀다는 점을 말해줘서 고마워.”라는 유사한 문장 구조로 생성했다.

다수의 선행 연구에서 언급한 것처럼 LLM은 자체적으로 공감과 정서적 지지를 수행할 수는 있으나, 그 깊이에 한계가 있다[22-23]. 초기 프롬프트에 정서적 지지를 포함하도록 명시하는 경우 더 깊이 있는 정서적 반응을 유도하는지 확인했으나, 실제 상호작용 내용 및 진단 정확도에 있어선 유의미한 차이가 발견되지 않았다.

2. Design and application of the improved model

시뮬레이션 실험 결과를 바탕으로, 프로토타입 모델의 초기 프롬프트에서 세 가지 사항을 수정해 개선 모델의 초기 프롬프트를 작성했다. 수정 내용은 다음과 같다. 첫째, 환경적 요인을 가족, 학교, 지역사회 요인으로 구분하고 그 예시를 포함해 상세하게 작성했다. 이는 환경적 요인을

Table 3. ANOVA of Diagnostic Accuracy for Science Learning Difficulties in Four Prototype Models

Classification of Science Learning Difficulties		Prototype Models				ANOVA	
		M1	M2	M3	M4	F-value	p-value
Total		.48	.40	.62	.60	2.19	.09
Classification Perspective (1)	Cognitive	.75	.75	.67	.50	.72	.54
	Affective	.50	.36	.43	.43	.18	.91
	Environmental	.33	.25	.71	.75	7.47*	.00
Classification Perspective (2)	Individual	.62	.54	.54	.46	.40	.75
	Family	.43	.36	.57	.79	2.09	.11
	School	.20	.10	.90	.70	9.76*	.00

* $p < .05$

세부적으로 분류했을 때 진단 정확도가 높았던 시뮬레이션 실험 결과를 반영한 것이다. 둘째, 진단 내용을 특정 단원(중학교 3학년 운동과 에너지)으로 한정하고, 해당 단원의 인지적·정의적 어려움을 세부적으로 작성했다. 시뮬레이션 실험에서 개인 요인을 단순히 인지적, 정의적 요인으로만 분류하는 것은 진단 정확도에 유의한 영향을 미치지 못했다. 그러나 진단 대상을 과학의 특정 단원으로 한정해 개인 요인을 더 구체화한다면, 환경적 요인과 마찬가지로 진단 정확도가 향상할 것으로 판단했다. 선행 연구[13, 18-19]를 분석해 물리학의 역학 단원에서 학생들이 겪는 인지적·정의적 어려움 세분화하고, 이를 초기 프롬프트에 포함했다. 셋째, 정서적 지지 여부에 따른 큰 차이는 없었던 점을 고려해, 대화 전반에 따뜻하고 공감 어린 말투를 유지하는 수준으로 조정했다. 지나치게 형식적이고 반복적인 표현을 방지하고자 정서적 반응에 대한 세부 지시는 배제했다. 프로토타입 모델의 프롬프트에 포함되었던 내용인 “진단을 마치면 [다음] 버튼을 누르도록 안내할 것”, “대화 초기에 과학에 관한 인식을 파악하고 단계적으로 학습의 어려움을 심화해 질문할 것”, “한 번에 하나의 내용만 질문할 것” 등은 그대로 유지했다.

개선 모델은 2025년 5월 26-28일의 기간에 수도권 소재 중학교 3학년 학생 12명을 대상으로 투입되었다. 학생들은 15-20분 정도의 시간 동안 학교에서 제공하는 태블릿 PC를 사용해 운동과 에너지 단원에서 경험한 어려움에 관해 개선 모델과 대화했다. 진단을 마치고 개선 모델의 안내에 따라 [다음] 버튼을 클릭하면 해당 학생에 대한 과학 학습의 어려움 요인이 출력되었다. 진단 시기는 운동과 에너지 단원의 수업이 끝나고 새로운 단원이 시작된 직후였다.

학생과 개선 모델의 상호작용을 분석하기 위해, 개선 모델이 학생에게 제시한 질문과 이에 대한 학생의 응답을 짝지어 ‘대화’로 정의했다. 대화의 개수는 총 222개로, 학생 1인당 평균 18.5개의 대화가 이루어졌다. 이후 개선 모델의 초기 프롬프트에 기반한 대화 분석틀을 작성했다(Table 4 참조). 분석틀은 크게 개인 요인, 환경적 요인, 기타의 세 범주로 구성되었다. 개인 요인은 인지적 요인과 정의적 요인의 하위 범주를 포함했고, 환경적 요인은 가족 요인, 학교 요인, 지역사회 요인으로 구성됐다. 하나의 대화가 여러 개의 요인을 포함한다면(예: 집이나 학교에서 과학을 공부할 때 방해가 되는 환경이 있나요?) 그 대화는 복수의 요인들에 해당하는 것으로 판단했다.

과학 학습의 어려움을 직접적으로 진단하지 않는 대화는 기타로 분류했다. 이 범주에는 운영적 대화와 무관한 대화가 포함됐다. 예를 들어 개선 모델이 “[다음] 버튼을 눌러주세요.”와 같이 상호작용의 진행을 위한 대화를 제시한 경우 운영적 대화로 분류했다. 때로 학생이 과학 학습의 어려움 진단과 직접적으로 연관되지 않은 응답(예: 고마워, 그만 말해 등)을 제시한 사례가 발견되었는데, 이는 무관한 대화로 분류했다.

Fig. 2는 분석틀의 범주별 대화 개수를 나타낸 그래프이다. 빨간색 막대그래프는 개인 요인, 초록색 막대그래프는 환경적 요인이며, 파란색 막대그래프는 기타 대화다. 인지적 요인과 관련된 대화가 114개로 가장 많았으며, 정의적 요인 관련 대화는 38개로 그 뒤를 이었다. 환경적 요인에 해당하는 대화는 전반적으로 적었다. 가족 요인, 학교 요인 관련 대화는 각각 20개와 22개였고, 특히 지역사회 요인과 관련된 대화는 7개만 관찰되어 학생 수(12명)보다도 적었다. 이는 개선 모델이 지역사회 요인에 대해 전혀 언급하지 않은 학생이 존재한다는 의미다. LLM은 프롬프트의 토큰 수가 많아질수록 중요 정보에 대한 집중도가 희석되어 성능이 하락한다[29]. 개선 모델의 초기 프롬프트에는 인지적 요인과 정의적 요인, 가족, 학교, 지역사회 요인을 모두 확인하라는 내용이 포함되었으나, 초기 프롬프트가 길어지면서 일부 지시를 실행하지 못했을 가능성이 있다.

Table 4. Conversation Coding scheme

Category	Subcategory	Description
Individual Factors	Cognitive	Difficulties in understanding concepts, formulas, symbols, or problem-solving
	Affective	Low motivation, anxiety, low self-esteem related to science learning
Environmental Factors	Family	Lack of parental support, emotional neglect, or family conflicts
	School	Mismatch with teaching style, peer/teacher conflicts, or negative school climate
	Community	Poor learning environment due to socioeconomic conditions or local risks
Others	Operational Turn	System-guided turns for proceeding with the interaction
	Unrelated Turn	Student responses irrelevant to learning difficulties

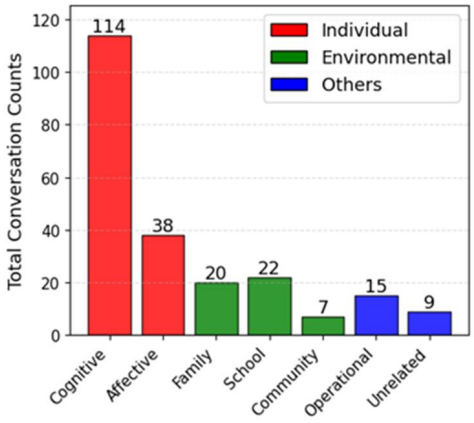

Fig. 2. Total Conversation Counts by the Coding Scheme

Table 5는 12명 학생별 각 범주의 대화 개수를 나타낸 표이다. 표를 보면 개선 모델이 지역사회 요인뿐 아니라 다른 요인에 대해서도 진단을 수행하지 않은 사례가 다수 있었음을 알 수 있다. 인지적 요인은 모든 학생에게 진단을 수행했으나, 그 외의 요인에서는 진단을 수행하지 않은 사례가 존재했다. 인지적 요인, 정의적 요인, 가족 요인, 학교 요인, 지역사회 요인 순으로 진단이 진행되는 모델의 특성상, 나중에 진단하는 요인에서는 긴 초기 프롬프트에 대화 맥락까지 더해져 처리해야 할 토큰이 늘어나면서 지시에 따라 진단을 수행하지 못한 것으로 보인다.

개선 모델과 학생 C의 상호작용에서는 진단과 관련된 대화가 4개밖에 이루어지지 않았다. 실제 대화 내용을 보면 초기에는 인지적 어려움에 관한 모델의 질문에 성실히 답했으나, 중간부터 학생 C가 자신의 일상생활과 관련된 대화를 시작하자 개선 모델은 이 맥락을 수용해 대화를 이어갔다. 6번의 일상 대화가 진행된 후 학생 C는 대화해줘서 고맙다고 말했고, 개선 모델은 [다음] 버튼을 누르라고 안내했다. 이러한 경향은 학생 H와의 대화에서도 발견되었는데, 인지적 요인과 관련된 13개의 대화를 진행한 후 학생이 과학 잘하는 방법을 질문하자 개선 모델은 이에 대한 답변과 함께 [다음] 버튼을 누르라는 안내를 제공했다. 진단 도중 맥락에서 벗어난 대화를 하게 되면 원래 맥락으로 돌아오지 못하고 그대로 상호작용이 종료되는 양상을 보였다. 이는 LLM을 활용한 상담이 과도한 자율성으로 인해 대화 흐름의 방향성을 상실할 수 있다는 선행 연구의 결과와 일맥상통한다[22].

학생 E와 F의 대화에서도 환경적 요인에 관한 언급이 전혀 없었다. 대화 내용을 보면 해당 학생들이 상호작용 도중 대화를 거부해 이러한 현상이 나타났음을 알 수 있다. 개선 모델이 정의적 요인에 관한 진단을 진행하고 있을 때, 학생 E는 “내가 알아서 할게.”라는 응답을 작성했고, 학생 F는 “이제 그만 말해.”라는 응답을 제시했다. 개선 모델은 곧바로 진단을 중단하고 [다음] 버튼을 누르라고 안내했다. 선행 연구[30]는 학습 맥락에서 LLM과의 대화가 길어지면 학생이 피로감을 느낄 수 있음을 지적한 바 있다. 진단 과정이 길어지면 학생이 상호작용을 중단하거나 회피하는 반응을 보일 수 있으며, 이로 인해 일부 요인에 대한 진단이 누락될 가능성이 있다.

IV. Conclusions

본 연구는 과학 학습의 어려움을 진단하는 LLM 기반 대화형 에이전트를 설계하고, 시뮬레이션 실험과 실제 중학생을 대상으로 한 적용을 통해 설계 방향을 도출했다. 선행 연구에 기반하여 설계된 4개의 프로토타입 모델로 시뮬레이션 실험을 진행한 결과, 과학 학습의 어려움을 환경적 측면에서 세분화한 프롬프트가 더 높은 진단 정확도를 보였으며, 정서적 지지의 포함 여부는 진단 정확도에 큰

Table 5. Conversation Counts by the Coding Scheme for Each Student

Student Label	Cognitive	Affective	Family	School	Community	Operational Turn	Unrelated Turn
A	1	2	9	3	1	1	1
B	8	1	2	2	2	1	0
C	4	0	0	0	0	1	6
D	9	3	0	6	0	1	0
E	8	5	0	0	0	1	1
F	10	1	0	0	0	1	1
G	6	7	2	3	0	2	0
H	13	0	0	0	0	2	0
I	10	10	2	1	0	2	0
J	15	6	2	1	3	1	0
K	17	1	1	3	1	1	0
L	13	2	2	3	0	1	0
Total	114	38	20	22	7	15	9

영향을 미치지 않았다. 이러한 결과를 바탕으로 개선된 모델을 설계하고 12명의 학생에게 투입했다. 학생과 개선 모델의 상호작용을 분석한 결과, 대화가 길어지면 일부 진단이 누락되는 문제가 나타났으며, 대화 도중 학생이 주제를 이탈할 경우 모델이 진단 흐름으로 복귀하지 못하는 한계도 드러났다. 또한 학생이 대화를 그만하라고 요구하면 상호작용이 조기에 종료되어 진단이 충분히 이루어지지 않는 사례가 있었다.

본 연구의 결과를 바탕으로, 과학 학습의 어려움을 진단하기 위한 LLM 기반 대화형 에이전트의 설계 방향을 다음과 같이 제안한다. 첫째, 학습의 어려움 요인들을 몇 개의 에이전트가 나누어 진단하는 방식이 필요하다. 이는 진단 항목이 많아질수록 초기 프롬프트와 대화가 길어져 일부 요인의 진단이 누락되는 문제를 예방할 방안이 된다. 예를 들어 하나의 에이전트가 인지적·정의적 요인을 진단하고, 다른 에이전트가 가족·학교·지역사회 요인을 진단하는 방식의 분산 설계를 고려할 수 있다. 둘째, 학습의 어려움 진단이라는 맥락에서 벗어나지 말라는 지시가 초기 프롬프트에 명시되어야 한다. 이것은 대화 중 학생이 진단 맥락에서 벗어날 경우에도 진단을 이어가기 위한 조치이다. 셋째, 학생의 부담을 줄이기 위한 대화 길이 조정이나 선택적 응답 구조 등의 방안이 탐색 되어야 한다. 진단의 정밀도는 대화의 양에 비례할 수 있지만, 이로 인해 학생이 피로감을 느끼지 않도록 적절한 타협점을 찾는 후속 연구가 요구된다.

본 연구는 LLM 기반 대화형 에이전트를 활용하여 과학 학습의 어려움을 진단하려는 초기 시도로서, 프롬프트 설계 방식이 진단 정확도에 미치는 영향을 실험적 방법으로 검증하고, 실제 학생과의 상호작용을 통해 현장 적용 가능성을 탐색했다는 점에서 의의가 있다. 그러나 적은 수의 시뮬레이션 세션과 제한된 교육적 맥락에서 수행되었기에 일반화에 한계가 있으며, 다양한 연령대와 학습 상황을 반영한 추가 연구가 필요하다. 또한 진단 이후의 구체적인 개입이나 피드백 전략이 포함되지 않았다는 점에서, 향후 진단을 넘어 학습 지원으로 확장되는 통합적 모델 개발이 요구된다. 본 연구가 학습의 어려움 진단 및 지원 체계 개발의 기초 자료로 활용되어, 학습 문제의 이해와 개선에 도움이 되기를 소망한다.

REFERENCES

[1] J. W. Song, “Constructivist science education and the map of students' physics misconceptions,” The Mathematical Education, Vol. 42, No. 2, pp. 87-109, May 2003.

[2] S. H. Jeon, and H. J. Lee, “Exploring high school science teachers’ perceptions of instructional changes due to achievement Standards-based assessment: Focusing on the impact of no longer indicating course ranking,” Journal of the Korean Association for Science Education, Vol. 44, No. 2, pp. 195-207, Apr. 2024. DOI: 10.14697/jkase.2024.44.2.195

[3] A. Roshan, M. Gurbaz, and S. Rahmani, “The effects of large classes on English language teaching,” Integrated Journal for Research in Arts and Humanities, Vol. 2, No. 2, pp. 38-41, March 2022. DOI: 10.55544/ijrah.2.2.20

[4] D. Zyngier, “Class size and academic results, with a focus on children from culturally, linguistically and economically disenfranchised communities,” Evidence Base: A Journal of Evidence Reviews in Key Policy Areas, No. 1, pp. 1-24, March 2014. DOI: 10.4225/50/5582118F8790B

[5] P. L. Morgan, G. Farkas, M. M. Hillemeier, and S. Maczuga, “Science achievement gaps begin very early, persist, and are largely explained by modifiable factors,” Educational researcher, Vol. 45, No. 1, pp. 18-35, Jan, 2016. DOI: 10.3102/0013189X16633182

[6] S. I. Hofer, and E. Stern, “Underachievement in physics: When intelligent girls fail,” Learning and Individual Differences, Vol. 51, pp. 119-131, Oct. 2016. DOI: 10.1016/j.lindif.2016.08.006

[7] K. F. Tsai, and G. Fu, “Underachievement in gifted students: A case study of three college physics students in Taiwan,” Universal Journal of Educational Research, Vol. 4, No. 4, pp. 688-695, Apr. 2016. DOI: 10.13189/ujer.2016.040405

[8] W. K. Jeong, and D. J. Shin, “A case study on the arithmetic thinking and the causes of learning difficulties in students struggling with mathematics: Focusing on first-year middle school students,” East Asian Mathematical Journal, Vol. 41, No. 2, pp. 165-183, DOI: 10.7858/eamj.2025.012

[9] S. I. Lee, “Effects of teaching-learning methods on the academic attitude and achievement toward science learning with underachievers,” [master’s thesis]. Korea National University of Education, pp. 5-61, 2004.

[10] P. A. Banerjee, “A systematic review of factors linked to poor academic performance of disadvantaged students in science and maths in schools,” Cogent Education, Vol. 3, No. 1, Article 1178441, May 2016. DOI: 10.1080/2331186X.2016.1178441

[11] S. Amaliyah, D. M. Fajar, and Aminulloh, “Analysis of factors causing students learning difficulties in learning science,” Science Education and Application Journal, Vol. 6, No. 1, pp. 61-74, March 2024. DOI: 10.30736/seaj.v6i1.1015

[12] S. Novianti, L. Y. Sari, and A. Afza, “Factors caused difficulty in learning science for students,” Journal of Biology Education Research (JBER), Vol. 3, No. 2, pp. 50-59, Nov. 2022. DOI: 10.55215/jber.v3i2.5949

[13] K. H. Lee, M. J. Han, M. J, Kim, and B. S. Choi, “Development and intervention effect of customized instructional program for underachievers in middle school science,” Journal of the Korean Association for Science Education, Vol. 34, No. 5, pp. 421-436, Aug. 2014. DOI: 10.14697/jkase.2014.34.5.0421

[14] O. T. Badmus, T. Jita, and L. C. Jita, “Exploring undergraduates' underachievement in science technology engineering and mathematics: Opportunity and access for sustainability,” European Journal of STEM Education, Vol. 9, No. 1, Article 10, June 2024. DOI: 10.20897/ejsteme/14741

[15] A. Hu, “Developing an AI-based psychometric system for assessing learning difficulties and adaptive system to overcome: A qualitative and conceptual framework,” arXiv, March 2024. DOI: 10.48550/arXiv.2305.14587

[16] M. A. Lee, and Y. B. Park, “Effects of a treatment program by types of underachiever on the science achievement and attitude toward science in junior high school students,” Journal of the Korean Association for Science Education, Vol. 22, No. 5, pp. 750-756, Dec. 2002.

[17] S. Y. Kim, K. R. Lee, N. G. Back, and J. H. Park, “Effects on individually tailored teaching according to types of under-achievement in science,” Journal of the Korean Association for Science Education, Vol. 35, No. 5, pp. 907-917, Oct. 2015. DOI: 10.14697/jkase.2015.35.5.0907

[18] J. H. Lim, and B. W. Lee, “Analysis of high-school students’ difficulty related to conceptual knowledge in solving problems in classical mechanics,” New Physics: Sae Mulli, Vol. 65, No. 4, pp. 333-342, Apr. 2015. DOI: 10.3938/NPSM.65.333

[19] J. H. Lim, and B. W. Lee, “Analysis of high-school students’ difficulty related to procedural knowledge in solving classical mechanics problems,” New Physics: Sae Mulli, Vol. 65, No. 9, pp. 888-899, Sept. 2015. DOI: 10.3938/NPSM.65.888

[20] N. E. Chere, and D. Hlalele, “Academic underachievement of learners at school: A literature review,” Mediterranean Journal of Social Sciences, Vol. 5, No. 23, pp. 827-839, Nov. 2014. DOI:10.5901/mjss.2014.v5n23p827

[21] K. Lekan, and Z. A. Pardos, “AI-augmented advising: A comparative study of GPT-4 and advisor-based major recommendations,” Journal of Learning Analytics, Vol. 12, No. 1, pp. 110-128, March 2025. DOI: 10.18608/jla.2025.8593

[22] W. Huang, J. Jiang, R. B. King, and L. K. Fryer, “Chatbots and student motivation: A scoping review,” International Journal of Educational Technology in Higher Education, Vol. 22, article 26, Apr. 2025. DOI: 10.1186/s41239-025-00524-2

[23] J. K. Park, V. K. Singh, P. Wisniewski, “Current landscape and future directions for mental health conversational agents for youth: Scoping review,” JMIR Medical Informatics, Vol. 13, article e62758, Feb. 2025. DOI: 10.2196/62758

[24] D. Ramandanis, and S. Xinogalos, “Investigating the support provided by chatbots to educational institutions and their students: A systematic literature review,” Multimodal Technologies and Interaction, Vol. 7, No. 11, article 103, Nov. 2023. DOI: 10.3390/mti7110103

[25] S. Meyer, and D. Elsweiler, “LLM-based conversational agents for behaviour change support: A randomised controlled trial examining efficacy, safety, and the role of user behaviour,” International Journal of Human-Computer Studies, Vol. 200, article 103514, May 2025. DOI: 10.1016/j.ijhcs.2025.103514

[26] V. Sorin, D. Brin, Y. Barash, E. Konen, A. Charney, G. Nadkarni, and E. Klang, “Large language models and empathy: Systematic review,” Journal of Medical Internet Research, Vol. 26, article e52597, Dec. 2024. DOI: 10.2196/52597

[27] L. Bojić, O. Zagovora, A. Zelenkauskaite, V. Vuković, M. Čabarkapa, S. Veseljević Jerković, and A. Jovančević, “Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm,” Scientific reports, Vol. 15, article 11477, Apr. 2025. DOI: 10.1038/s41598-025-96508-3

[28] A. Algherairy, and M. Ahmed, “Prompting large language models for user simulation in task-oriented dialogue systems,” Computer Speech & Language, Vol. 89, Article 101697, Jan. 2025. DOI: 10.1016/j.csl.2024.101697

[29] M. Levy, A. Jacoby, and Y. Goldberg, “Same task, more tokens: The impact of input length on the reasoning performance of large language models,” arXiv, July 2024. DOI: 10.48550/arXiv.2402.14848

[30] T. H. Min, and B. W. Lee, “Interaction between middle school students and generative AI in inquiry design,” New Physics: Sae Mulli, Vol. 75, No. 4, pp. 350-362, Apr. 2025. DOI: 10.3938/NPSM.75.350

Authors

Tae-Ho Min received the B.S degree in Science Education from Dankook University, Korea, in 2017, and the M.S. degree in Science Education from Seoul National University, Korea, in 2023.

He is currently a science teacher at Bora Middle School, Korea, and is pursuing the Ph.D. degree with the Department of Science Education at Dankook University. His current research interests are educational data analysis and AI for education.