컴퓨터교육학회 논문지 2025년 제28권 제4호
https://doi.org/10.32431/kace.2025.28.4.004

CSCL에서의 인지적 참여 수준 분석: 생성형 인공지능의 교육적 활용 가능성 탐색*

Cognitive Engagement Analysis in CSCL: Exploring Educational Potential of Generative AI

진명화† · 박은주†† · 임규연†††
Myunghwa Jin† · Eun Joo Park†† · Kyu Yon Lim†††

요약

컴퓨터기반 협력학습 (Computer-Supported Collaborative Learning, CSCL)에서 학습자의 인지적 참여는 협력학습의 질을 좌우한다. 특히 동시적 CSCL 환경에서 교수자가 적절한 지원을 제공하기 위하여 인지적 참여 수준을 실시간으로 분석하는 것이 필요하다. 본 연구는 생성형 인공지능이 텍스트 데이터를 인지적 참여 수준에 따라 정확하게 분류할 수 있는지 평가하는 데 목적이 있다. 이를 위하여 채팅 기반의 토론을 통해 텍스트 데이터를 수집하였다. 이후 GPT-3.5-turbo와 GPT-4o에 인지적 참여 수준별 예시를 제시한 퓨샷러닝 모델과 1,080개의 텍스트 데이터를 훈련한 파인튜닝 모델을 개발하고, 인지적 참여 수준별로 GPT 모델의 성능을 평가하였다. 분석 결과, 토론 주제와 무관한 내용인 인지적 참여 수준 0을 분류하는 모델의 성능은 GPT 모델의 종류나 데이터 훈련 정도와 관련 없이 높은 것으로 나타났다. 단편적 의견을 제시하는 인지적 참여 수준 1과 정교화된 설명을 나타내는 수준 2를 분류하는 경우, 퓨샷러닝을 실시한 GPT-3.5-turbo는 보통 수준의 성능을 나타냈으며 나머지 모델은 높은 성능을 보였다. 다양한 아이디어를 통합하는 내용인 인지적 참여 수준 3을 분류하는 성능은 파인튜닝 한 모델이 퓨샷러닝 모델보다 월등하게 높게 나타났다. 연구 결과를 토대로 생성형 인공지능을 활용하여 텍스트의 인지적 참여 수준을 분석하는 데 실질적으로 적용할 수 있는 LLM 모델 선택 및 데이터의 훈련 정도, 프롬프트 설계와 관련된 시사점을 제시하였다.

주제어 컴퓨터기반 협력학습, CSCL, 인지적 참여, 생성형 AI, 텍스트 분류

ABSTRACT

Cognitive engagement is crucial for determining the quality of Computer-Supported Collaborative Learning (CSCL). Real-time analysis of group cognitive engagement levels is essential for instructors to provide appropriate support in synchronous CSCL environments. This study evaluates how effectively GPT models can classify text data by cognitive engagement levels. The research employed two technical approaches using chat-based discussion data: few-shot learning models with GPT-3.5-turbo and GPT-4o using examples per engagement level and fine-tuned models trained on 1,080 text examples. Performance of the GPT models was evaluated and compared across different cognitive engagement levels. Results indicated consistently high performance in classifying cognitive engagement level 0 across all model variations and training conditions. For classifying cognitive engagement level 1 and level 2, the few-shot learning GPT-3.5-turbo model showed moderate performance, while other models demonstrated high performance. For classifying cognitive engagement level 3 , fine-tuned models demonstrated significantly superior performance to few-shot learning models. Based on these findings, implications were presented regarding the practical application of generative AI for analyzing cognitive engagement levels in chat messages, specifically addressing the selection of LLM models, degree of data training, and prompt design.

Keywords Computer-Supported Collaborative Learning, CSCL, cognitive engagement, generative AI, text classification

†정회원 이화여자대학교 교육공학과 박사후연구원
††정회원 이화여자대학교 교육공학과 석사
†††정회원 이화여자대학교 교육공학과 교수(교신저자)

논문투고 2025년 02월 18일
심사완료 2025년 04월 04일
게재확정 2025년 04월 09일
발행일자 2025년 04월 16일

* 본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2023R1A2C1004108).

본 논문은 제1저자의 이화여자대학교 박사학위논문 일부를 수정 및 확장한 것임.

1. 서론

컴퓨터기반 협력학습 (Computer-Supported Collaborative Learning, 이하 CSCL)은 테크놀로지를 활용하여 학습자들이 협력적으로 공동의 지식을 창출하는 학습 과정이다[1]. CSCL에서 학습자의 참여(engagement)는 협력학습의 질을 좌우한다[2, 3]. 특히 높은 품질의 협력학습이 이루어지려면 공동으로 지식을 구축하고 협력하여 문제를 해결하는 과정에서 학습자가 인지적으로 얼마나 심도 있게 참여했는지가 중요하다. 선행연구에서는 협력 과정에서 나타난 학습자의 인지적 참여 수준을 점검하기 위해 주로 학습자가 작성한 메시지의 글자 수나 게시물 수, 발화 횟수나 지속 시간 등을 활용해 왔다[4, 5, 6]. 그러나 이러한 데이터는 협력 과정을 단편적으로 반영하기 때문에 협력의 내용을 종합적으로 파악하는 데 한계가 있다. 따라서 학습자의 인지적 참여 수준을 보다 심층적으로 파악하기 위해서는 학습자가 협력 과정에서 작성한 텍스트나 발화의 내용을 분석하는 접근이 필요하다. 특히 학습자 간 상호작용이 즉각적으로 발생하는 동시적 CSCL 환경에서는 교수자가 협력 품질을 점검하고 적절한 지원을 제공하기 위해 학습자의 인지적 참여 수준을 실시간으로 분석하고 협력 과정을 점검할 수 있어야 한다.

협력 과정에서 생성된 내용을 자동으로 분석하려면 실시간으로 텍스트의 의미를 파악하고 이를 분류하는 고도의 기술이 필요하다. 그동안 기술적인 한계로 인해 이러한 기능의 구현이 현실적으로 어려웠지만 최근 생성형 인공지능 기술(예: GPT-4o, BERT 등)의 발전으로 그 제약을 극복할 기반이 마련되고 있으며, 이에 따라 동시적 CSCL에서 학습자의 인지적 참여 수준을 실시간으로 분석할 수 있는 가능성이 높아지고 있다. 기존 연구에서도 인공지능을 활용하여 협력 과정에서 생성된 텍스트를 분석하고 이를 점검 및 지원에 활용하려는 다양한 시도가 있었다. 예를 들어 Sullivan과 Keith[7]의 연구에서는 협력적 대화의 구문을 자연어 처리(Natural Language Processing, NLP) 기술을 활용하여 품사 단위로 자동으로 태그한 뒤 분석하는 시도를 하였다. 또한 Cai 외[8]의 연구에서는 협력학습에서 작성된 채팅 데이터를 인식적 네트워크 분석 및 토픽모델링을 활용하여 분석하고자 하였다. 그러나 이러한 접근은 텍스트의 맥락을 이해하지 못해 정확한 의미를 파악하는 데 어려움을 겪거나 실시간 자동 분석이 이루어지지 않아 협력학습이 종료된 후 교수자가 분석 결과를 확인하고 지원 여부 및 수준을 결정할 수 있다는 한계가 존재했다.

반면, 생성형 인공지능(generative artificial intelligence, 이하 생성형 AI)은 여러 측면에서 기존 한계를 극복할 수 있는 가능성을 보여준다. 우선 생성형 인공지능은 입력받은 텍스트를 단어나 숫자, 구두점 등의 작은 단위로 나누는 토큰화를 거친 다음, 토큰화된 단어 간의 벡터 값을 산출하여 텍스트의 문맥을 인식한다. 이에 이전 기술에 비해 텍스트의 맥락을 보다 정확하게 인식할 수 있다[9]. 이는 학습자가 작성한 텍스트에 담긴 내용을 분석하여 인지적 참여 수준을 보다 정확하게 분석하는 데 도움이 된다. 또한 생성형 AI는 협력학습이 발생하는 동안 실시간으로 텍스트를 처리하고 분류할 수 있다[10]. 이러한 실시간성은 동시적 CSCL에서 보다 효과적으로 학습을 지원할 수 있게 한다. 종합하면, 기술적 발전은 협력학습 과정에서 생성된 메시지를 단순히 통계적으로 분석하는 것을 넘어 보다 심층적이고 종합적인 접근을 가능하게 하며, 학습자의 인지적 참여를 더 잘 이해하고 지원할 수 있는 기반을 제공한다.

다양한 대규모 언어 모델(Large Language Model, 이하 LLM)이 제시되면서 여러 분야에 빠르게 적용되고 있지만 생성형 AI의 교육적 활용은 아직 초기 단계에 있으며, 그 타당성과 신뢰성에 대한 충분한 검토가 이루어지지 않은 상태이다. 일부 선행연구에서 생성형 AI 기술을 활용한 텍스트 분류를 CSCL 맥락에서 시도하였으나[11, 12] 학습자의 인지적 참여 수준을 자동으로 분석하는 연구는 극히 미비하다. 그러나 생성형 AI가 인지적 참여 수준을 분석하는 성능에 대한 타당성을 확보하고, 교육 현장에서 효과적으로 생성형 AI를 활용하기 위한 시사점을 도출하기 위해서는 여러 LLM 모델의 성능을 비교, 분석하는 과정이 선행되어야 한다. 이를 통해 인지적 참여 수준을 분석하기 위하여 어떠한 LLM 모델을 선택하고, 어떤 프롬프트로 데이터를 어떻게 훈련해야 보다 나은 성능을 나타내는가를 확인할 수 있다. 특히 인지적 참여는 학습자가 정신적으로 노력한 정도에 따라 피상적 수준인 낮은 단계부터 심층적 수준인 높은 단계까지 여러 단계로 나눌 수 있다[13, 14, 15]. LLM 모델이 단순한 내용을 분류하는 성능은 높지만 복잡한 내용을 분류하는 성능은 낮다는 연구 결과를 고려할 때[16], 협력학습에서 생성된 텍스트를 분석하는 LLM의 성능 또한 인지적 참여 수준에 따라 다를 것으로 예상할 수 있다. 이에 인지적 참여 수준별로 텍스트를 분석하는 LLM 모델들의 성능을 확인하여 생성형 AI의 타당도를 확보하고, 효과적인 LLM 활용을 위한 시사점을 도출하는 것이 필요하다.

이에 본 연구는 생성형 AI인 GPT 모델이 협력 과정에서 생성된 텍스트 데이터의 인지적 참여 수준을 정확히 분석할 수 있는지를 평가하고자 한다. 구체적으로 GPT-3.5-turbo와 GPT-4o 모델을 각각 적용하여 CSCL 과정에서 생성된 채팅 메시지를 인지적 참여 수준에 따라 분석하는 성능을 점검하고자 한다. 이를 통해 CSCL 맥락에서 협력학습 과정을 자동으로 분석하는 생성형 AI의 활용 가능성을 탐색하고, 향후 연구와 교육 현장에서 효과적으로 생성형 AI를 활용하기 위한 실질적인 시사점을 제시하고자 한다. 본 연구의 연구 문제는 'CSCL에서 메시지의 인지적 참여 수준을 분석하는 GPT 모델의 성능과 활용 가능성은 어떠한가?'이다.

2. 이론적 배경

2.1 컴퓨터기반 협력학습과 인지적 참여

CSCL은 네트워크나 컴퓨터로 매개되거나 촉진되는 협력학습으로, 여러 학습자가 테크놀로지를 활용하여 함께

학습하는 형태를 의미한다[1]. CSCL에서는 학습자들이 테크놀로지를 기반으로 상호작용하고 공동의 지식을 구축하는 과정에서 학습이 발생한다. 특히 학습자들이 협력 과정에서 다양한 의견을 통합하고 새로운 지식을 창출하는 높은 수준의 인지적 참여를 할 때 더 높은 학업성과가 나타난다[14, 17]. 이에 학습자들의 인지적 참여 정도는 성공적인 협력학습에서 주요한 요인으로 고려되어 왔다.

인지적 참여(cognitive engagement)는 학습자가 비판적으로 사고하며 심층적이고 적극적인 학습을 위해 정신적으로 관여(involvement)하는 것으로, 복잡한 문제를 해결하거나 새로운 학습 내용을 이해하기 위해 심리적으로 투자하는 상태로 정의된다[18, 19]. 일반적인 학습 맥락에서 인지적 참여 수준은 학습자의 사용 전략에 따라 재연이나 단순한 암기와 같은 피상적 수준과 비교, 비판적 사고, 정보 통합과 같은 심층적 수준으로 구분할 수 있다[20]. CSCL 맥락에서 인지적 참여는 협력학습 과제를 해결하기 위해 정보를 탐색, 분석, 해석하며, 여러 의견과 정보에 대해 비판적으로 사고하고, 나아가 다양한 정보와 의견을 종합하는 정신적 노력으로 정의할 수 있다[21]. CSCL에서 인지적 참여는 그룹이 지식을 구축한 정도를 반영하기 때문에 협력학습 과정에서 어떠한 인지적 수준으로 협력학습을 실시하였는지 확인하는 지표로 기능한다[2, 22]. 협력학습 과정에서 학습자들이 다양한 의견과 정보를 재구조화하거나 통합하고, 조직화하는 등 높은 수준으로 인지적 참여를 할 때 양질의 협력학습이 이루어지기 때문에[3, 15] 협력학습에서 나타나는 인지적 참여 수준을 분석하고 점검하는 것이 필요하다.

인지적 참여는 협력 과정에서 학습자가 생성한 메시지나 게시글, 발화 내용 등의 내용을 통해 확인할 수 있다[23]. 이에 여러 선행연구에서 학습자가 작성한 메시지나 게시글, 발화 내용을 기반으로 협력학습의 인지적 참여 수준을 분석하는 프레임워크를 제시하였다[13, 14]. 구체적으로, Ouyang과 Chang[14]은 온라인 토론에서 발생하는 집단 차원의 인지적 참여 수준을 세 가지 수준으로 구분하여 제시하였다. 인지적 참여 수준이 가장 낮은 피상적 수준(superficial-level)은 학습자가 구체적인 설명이나 이유를 제시하지 않고 단순히 동의·반대 의견이나 개인의 관점을 제시하거나, 상대방이 제시한 내용에 대해 질문하거나, 사실적 정보를 확인하고 질문하는 수준을 의미한다. 중간 수준(medium-level)은 다른 구성원의 의견을 바탕으로 확장된 아이디어나 관점을 제시하는 수준을 의미한다. 심층적 수준(deep-level)은 구체적인 설명이나 정보를 바탕으로 여러 구성원 간 의견을 연결, 통합하여 아이디어나 논점을 심화하는 수준이다. 이러한 분류와 유사하게 Guo 외[13]는 온라인 토론에서 나타나는 인지적 참여 수준을 낮은 수준과 높은 수준으로 구분하였다. 낮은 수준은 주제와 관련되지 않은 진술이나 정서적 표현, 주제와 관련 있지만 구체적 설명 없이 사실적인 정보만 제시하는 진술을 포함한다. 높은 수준은 개인의 견해에 대해 추가적인 설명이나 이유를 제시하는 진술과 다양한 견해나 자료를 종합하는 진술, 그리고 토론 내용에 대해 평가하거나 판단하는 의견을 포함한다.

종합하면, 학습자와 그룹의 인지적 참여 정도는 성공적인 협력학습에 영향을 미치는 주요 요인으로, 보다 심층적 수준의 인지적 참여가 나타날수록 그룹이 다양한 의견을 통합하고 아이디어를 심화한다고 볼 수 있다. 이에 효과적인 협력학습이 이루어지도록 지원하기 위하여 협력 과정에서 나타나는 인지적 참여 수준을 분석하고 점검하는 것이 필요하다.

2.2 생성형 AI를 활용한 협력 텍스트 자동 분석

협력학습에서 나타난 인지적 참여 수준을 확인하기 위해 종래의 연구에서는 주로 인간 평가자가 수기로 협력 과정에서 생성된 내용을 분석해 왔다. 하지만 이러한 분석 방식은 노동집약적이며, 교수자가 협력학습이 진행되는 동안에 실시간으로 인지적 참여 수준을 분석하는 것이 현실적으로 어렵기 때문에 학습자에게 그룹의 인지적 참여 수준에 대한 정보를 즉각적으로 제공할 수 없다는 한계가 있다.

최근에는 인공지능 기술이 고도로 발전함에 따라 인간 평가자가 실시하였던 내용분석을 자동으로 분석하는 연구가 시도되고 있다. 로지스틱 회귀나 랜덤 포레스트 등 머신러닝 기법을 통해 협력 내용을 분석한 선행연구들은[24, 25] 내용 측면에 대한 자동 분석의 가능성을 시사한 바 있다. 나아가 최근에는 생성형 AI가 보편화되면서 자동화된 내용분석을 가능하게 하는 새로운 장이 마련되었다.

생성형 AI는 텍스트나 이미지 등 다양한 유형의 데이터를 사전 학습하고, 프롬프트에 대응하여 새로운 콘텐츠를 생성하는 기술이다[26]. 특히 대규모 언어 모델(Large Language Model, 이하 LLM) 기반의 생성형 AI는 입력값에 대한 답변을 인간과 유사한 수준으로 생성하며[27] 기존의 NLP나 머신러닝과 비교하여 적은 양의 데이터로 보다 용이하게 모델을 훈련할 수 있는 특징을 가진다. LLM 기반 생성형 AI는 텍스트 분류, 요약, 번역, 문서나 이미지 생성 등을 수행할 수 있으며, 대표적 예시로 OpenAI의 ChatGPT, 구글의 Gemini, 메타의 LLaMA가 있다. 교육 분야에서도 이들을 평가, 외국어 학습, 작문, 교수설계 등 다양하게 활용할 수 있는 방안이 제시되고 있다[28, 29].

CSCL 분야에서도 LLM 기반 생성형 AI를 활용하여 협력 텍스트를 자동으로 분류하는 연구를 실시하고 있다. 구체적으로, Gunawardena 외[10]는 딥러닝 기법 중 하나인 인공신경망(neural network)과 LLM 기반 생성형 AI를 활용하여 Gunawardena 외[30]가 제시한 상호작용 분석 모델(Interaction Analysis Model, IAM)에 따라 토론 텍스트를 5개 단계로 분류할 수 있는지 탐구하였다. 그 결과, 인공신경망 기법에 비해 LLM 기반 모델이 더 나은 성능을 나타냈으며, 특히 훈련 데이터를 추가로 투입하여 분류 목적에 맞게 파인튜닝(fine-tuning) 한 LLM 모델이

각 단계를 설명하는 프롬프트를 제시하고 토론 텍스트를 분류한 LLM보다 높은 정확도를 나타냈다. 다만, 상호작용 분석 모델의 5단계 중에서 높은 수준에 속하는 4, 5단계의 분류 결과는 나머지 단계에 비해 분류 정확도가 낮은 것으로 제시되었다. 이러한 결과는 LLM 기반 생성형 AI가 협력 텍스트를 자동으로 분류할 수 있다는 가능성을 제시하며, 동시에 충분한 데이터의 투입을 통한 훈련과 각 단계를 구분하는 명확한 기준이 필요하다는 시사점을 제시한다.

Suraworachet 외[12]는 협력 담화에서 나타나는 학습자의 어려움을 포착하고 이를 인지적, 메타인지적, 정의적, 기술적 어려움으로 자동 분류하기 위해 다양한 자연어 처리 기법의 모델을 개발하고 성능을 비교하였다. 그 결과, 규칙기반분석(rule-based approach) 기법에 비해 지도학습(supervised learning) 기법과 LLM 기반 생성형 AI의 분류 결과가 월등히 높은 것으로 나타났다. 다만, 연구진들은 규칙기반분석과 지도학습은 키워드나 특성 중요도(feature importance) 등을 함께 제공하여 분류 결과에 대한 증거를 제시할 수 있기 때문에 결과를 신뢰할 수 있으며 교수자가 보다 구체적인 교수적 처방을 제공하도록 돕지만, LLM 기반 생성형 AI는 분석 결과에 대한 간략한 요약만을 제시하기 때문에 의사결정에 충분한 증거를 제시하지 못한다는 점을 제시하였다.

마지막으로 Hou 외[11]는 협력적 주석 활동에서 학습자가 작성한 주석을 자동으로 분류하는 연구를 수행하였다. 이들은 프롬프트 기반의 GPT-3.5-turbo 성능과 파인튜닝을 실시한 GPT-3.5-turbo 성능을 비교하였다. 분석 결과, 프롬프트 기반의 GPT-3.5-turbo는 수용가능한 수준의 정확도를 나타냈으나, 파인튜닝 모델은 프롬프트 기반 모델보다 월등히 높은 정확도를 나타냈다. 특히 성찰이나 아이디어 통합과 같이 맥락 의존적 분류의 경우, 정확도의 차이가 큰 폭으로 나타나 파인튜닝이 필수적이라고 제안하였다.

종합하면, 인간 평가자가 수기로 협력 텍스트를 분류하였던 과정을 고도의 인공지능 기술이 수행할 수 있으며, 특히 텍스트 분류에 대한 LLM 기반 생성형 AI의 정확도는 상당히 높은 것으로 이해할 수 있다. 그러나 생성형 AI의 성능은 사용 목적이나 분류 기준뿐만 아니라 어떤 프롬프트를 사용하고 어떠한 기법을 사용하는가에 따라 달라질 수 있다. 이에 CSCL 맥락에서 인지적 참여 수준을 분석하는 LLM 모델의 성능을 확인할 필요가 있다.

3. 연구 방법

본 연구에서는 GPT-3.5-turbo와 GPT-4o를 특정 데이터 세트로 추가 훈련한 파인튜닝 기법과 GPT-3.5-turbo와 GPT-4o에 인지적 참여 수준별 예시를 포함한 프롬프트를 제공한 퓨샷러닝 기법이 인지적 참여 수준에 따라 채팅 메시지를 분류하는 성능을 평가, 비교하였다. 2025년 1월을 기준으로 무료 버전의 ChatGPT에 임베드된 GPT-3.5-turbo 모델과 유료 버전에 임베드된 GPT-4o 모델을 선정하였다. 또한 GPT 모델을 특정 목적에 맞게 사용하기 위해 적용할 수 있는 파인튜닝 기법과 퓨샷러닝 기법을 실시하였다. 이를 위해 Ahmad 외[31]가 제시한 자동적 내용분석을 위한 모델 개발 절차에 따라 1) 데이터 세트 생성, 2) 생성형 AI 기반 모델 개발, 3) 모델의 성능 평가 과정을 거쳤다. 구체적인 과정은 다음과 같다.

3.1 데이터 세트 생성

3.1.1 데이터 수집

본 연구에서 수집하는 데이터는 텍스트 기반의 채팅으로 이루어진 토론에서 연구 참여자가 작성한 채팅 메시지이다. 채팅 메시지는 연구 참여에 동의한 사범대생 및 교직이수자 60명이 협력학습을 실시하는 과정에서 생성되었다. 사범대생 및 교직 이수자를 연구 대상으로 선정한 이유는 이들이 토론 주제인 '청소년 학습에서의 생성형 AI 사용'과 관련성이 높은 예비 교사로, 주제를 유의미하게 인식할 것으로 판단하였기 때문이다. 연구 참여자는 채팅 기반의 토론에 1회 참여하였으며, 토론은 4인 1조로 실시되었다. 수집한 메시지는 의미 단위로 구분하였다. 예를 들어, '기초적인 지식이 우선이 되어야 하는데', ' 챗지피티로 모두 대체할 시에는 기초가 제대로 쌓이지 않겠네요'와 같이 하나의 문장이나 의견이 여러 개의 채팅 메시지로 구분된 경우에는 하나의 메시지로 산정하였다. 이러한 과정을 통해 수집된 메시지의 의미 단위 개수는 총 1,350개였으며, 수집한 채팅 메시지는 익명화 처리하였다.

3.1.2 인지적 참여 수준 분석

채팅 메시지의 인지적 참여 수준은 교육공학 박사과정 1인과 교육공학 석사 1인이 분석하였다. 인지적 참여 수준의 분석은 Ouyang과 Chang[14]이 제시한 인지적 참여 수준 프레임워크를 본 연구의 맥락인 동시적 CSCL에 적합하게 수정하여 실시하였다. 해당 프레임워크는 온라인 토론에서 나타나는 인지적 참여 수준의 위계를 명시적으로 제시하였다는 측면에서 인지적 참여 수준을 구분하는 본 연구의 목적에 부합한다고 판단하였다.

Table 1을 살펴보면, 수준 1은 피상적인 인지적 참여 수준으로 참여자의 관점이나 생각이 제시되지 않은 단순 동의나 반대, 부연 설명 요청, 질문, 단순한 경험이나 정보를 공유하는 메시지를 포함하였다. 수준 2는 중간 수준의 인지적 참여로 의견에 대한 설명이나 경험, 근거 등을 포함하여 구체적으로 설명하거나 상세한 아이디어를 제시하는 의견을 포함하였다. 수준 3은 심층적인 인지적 참여 수준으로, 여러 그룹원이 제안한 아이디어를 통합하고 의견을 연결, 확장하는 메시지를 포함하였다. 마지막으로 수준 0은 주제에서 벗어난 메시지를 포함하였다.

Table 1. Framework for cognitive engagement levels

Level	Definition	Explanation
Level 0	Off-topic	Messages presenting content unrelated to the discussion topic
Level 1	Simple question or simple opinion	Messages that ask for definitions of terms or concepts, questions about information, express agreement/disagreement without specific reasons, share simple experiences or information, or present fragmented opinions
Level 2	Detailed explanation	A message that includes further explanation or expansion of other group members' opinions, explanations of terms or phenomena, their relationships, reasons, mechanisms, and ideas that suggest specific practical measures.
Level 3	Integration of the ideas	A message that integrates various ideas or connects and organizes the presented ideas at a higher level.

채팅 메시지에 대한 인지적 참여 수준의 코딩은 교육공학 박사 1인과 교육공학 석사 1인이 실시하였다. 두 평정자는 코딩 스킴의 내용을 숙지한 다음, 채팅 메시지의 20%를 함께 분석하며 공동의 이해를 가지도록 하였다. 그다음, 평정자는 각자 나머지 메시지에 대한 코딩을 실시하고, 의견이 불일치 하는 문장에 대하여 합의에 도달할 때까지 논의와 검토의 과정을 거쳤다. 평정자 간 신뢰도를 확인한 결과, Cohen의 Kappa는 .86으로 높은 수준의 신뢰도를 나타내[32] 평정자 간 유사성은 신뢰할 수 있는 수준인 것으로 판단하였다.

3.1.3 데이터 세트 구성

GPT-3.5-turbo와 GPT-4o의 파인튜닝을 실시하기 위하여 데이터 세트를 구성하였다. 데이터 세트는 8회에 걸친 실험에서 수집한 채팅 메시지로 구성하였으며, 인지적 참여 수준 0은 28개, 수준 1은 598개, 수준 2는 656개, 수준 3은 68개의 메시지로 구성되었다. 수집 데이터를 8대 2의 비율로 훈련(training) 및 검증(validation) 데이터 세트(N = 1,080)와 테스트(test) 데이터 세트(N = 270)로 나누었으며, 훈련 및 검증 데이터는 다시 8대 2의 비율로 데이터 세트를 구분하였다($N_{training}$ = 900, $N_{validation}$ = 180, N_{test} = 270). 훈련 데이터 세트에서 인지적 참여 수준 0은 10개, 수준 1은 413개, 수준 2는 435개, 수준 3은 42개, 테스트 데이터 세트에서 인지적 참여 수준 0은 11개, 수준 1은 102개, 수준 2는 139개, 수준 3은 18개로 이루어졌다.

3.2 생성형 AI 기반 모델 개발

3.2.1 파인튜닝 기반 GPT-3.5-turbo와 GPT-4o

GPT-3.5-turbo와 GPT-4o의 파인튜닝은 OpenAI에서 제시한 가이드라인(OpenAI, n.d.)에 따라 실시하였다. 먼저, 생성한 데이터 세트를 파인튜닝을 위한 데이터 형식으로 변환하였다. 코드에 포함된 역할 객체를 살펴보면, system은 GPT-3.5-turbo, GPT-4o가 수행하는 맥락과 역할을 지정하는 것으로, 청소년의 생성형 AI 사용에 대한 주제로 토론이 이루어지는 맥락과, 채팅 메시지를 인지적 참여 수준으로 구분하는 역할이 부여된다는 내용을 포함하였다. 인지적 참여 수준별 정의와 예시 문장도 함께 제시하였다. 지시 사항을 제공하는 객체인 user의 경우, GPT-3.5-turbo, GPT-4o가 채팅 메시지를 인지적 참여 수준에 따라 분류하도록 지시하였다. 원하는 답변 형식에 대한 정보를 제공하는 객체인 assistant에는 분류 결과를 수준 0~3으로 제공하도록 하였다. 모델의 하이퍼파라미터 설정에서 에포크(epoch) 수는 3으로, 배치 사이즈(batch size)는 8로 설정하였으며, 학습률 배수(learning rate multiplier)는 2를 적용하였다. 출력값의 무작위성을 결정하는 온도(temperature)의 경우에는 정확한 답변을 출력하기 위해서는 낮은 온도를 권장하므로[33] 0으로 설정하였다.

다음으로, 변환된 형식의 훈련 데이터세트와 검증 데이터세트를 OpenAI에서 제공하는 파인튜닝 인터페이스(fine-tuning UI)에 투입하여 GPT-3.5-turbo와 GPT-4o 각각에 대한 파인튜닝을 실시하였다. 파인튜닝이 제대로 수행되었음을 확인하기 위하여 인터페이스에서 제공하는 훈련 손실(training loss)과 검증 손실(validation loss)이 점진적으로 줄어드는 양상을 확인하였다.

3.2.2 퓨샷러닝 기반 GPT-3.5-turbo와 GPT-4o

GPT-3.5-turbo와 GPT-4o의 퓨샷러닝(few-shot learning)을 실시하기 위해 OpenAI에서 제공하는 프롬프트 엔지니어링 가이드라인에 따라 프롬프트를 작성하였다. 언어 모델의 수행 내용에 대한 몇 가지 사례를 제시할 때 성능이 높아진다는 선행연구[34, 35]에 따라 프롬프트에 인지적 참여 수준별 메시지 사례를 함께 제시하는 퓨샷러닝을 실시하였다. 퓨샷러닝을 위한 프롬프트는 시스템의 수행 역할(role), 수행 맥락(context), 수행 작업(assistant need to do), 평가 기준(rubric)으로 구성하였으며, 인지적 참여 수준별로 4~5개의 문장을 포함하였다.

프롬프트에 제시된 문장의 예시는 다음과 같다. 수준 0은‘오늘 날씨가 어때요?’, 수준 1에는 ‘원인이 뭐라고 생각하시나요? ’,‘저는 학생들이 생성형 인공지능을 사용하는 데 회의적이에요’와 같은 문장이, 수준 2에는‘생성형 인공지능은 잘못된 정보나 가짜 정보를 진짜처럼 제시할 수 있기 때문에 학생들이 잘못된 지식을 습득할 수 있습니다’와 같은 예시가 포함되었다. 또한 수준 3에는‘의견을 종합 해보면, ChatGPT의 보편화로 인한 학습자의 무분별한 학습 가능성으로 인하여 교육계에서 발생하는 문제라는 걸 도출할 수 있습니다’와 같은 문장을 포함하였다.

프롬프트는 ChatGPT에서 여러 프롬프트를 반복적으로 점검하고 개선하는 순환적 과정을 거쳐 최적화되었다. 초기에는 프롬프트를 서술형으로 작성하였으나, 명확성 향상을 위해 역할, 맥락, 수준별 정의 및 예시를 구분

하여 구조화하였다. 또한 프롬프트의 효과성을 평가하기 위해 응답의 일관성과 정확성을 주요 평가 기준으로 설정하였다. 프롬프트 점검을 위해 반복적인 테스트를 수행하고, 개선 방향을 도출하였다. 특히 수준 1과 수준 2를 혼동하는 경향이 나타나, 수준 1 내에서도 단순 동의, 이유 없이 의견만 제시하는 경우 등 다양한 사례를 명확히 구분하고 이에 대한 설명과 예시를 추가하였다. 수준 3의 경우, 초기에는 수준 3에 해당하는 문장만을 제시하였으나 여러 의견을 통합하는 과정이 보다 명확히 드러날 수 있도록 이전의 대화 흐름을 포함한 예시를 함께 제공하였다. 이러한 반복적 과정을 통해 최종적으로 작성된 프롬프트의 예시는 Figure 1과 같다.

[Role]

Your role is to classify the cognitive engagement levels of chat-based discussion content. [...]

[Context]

Four students had a chat-based discussion on the topic [...]

[Task]

Classify each message according to cognitive engagement levels 0, 1, 2, or 3 based on the [Rubric and Examples].

[Rubric and Examples]

- Definition and example sentences for cognitive engagement levels

Figure 1. Example of the prompt

3.3 모델의 성능 평가

앞서 생성한 테스트 데이터 세트를 활용하여 파인튜닝을 실시한 GPT-3.5-turbo와 GPT-4o, 퓨샷러닝을 실시한 GPT-3.5-turbo와 GPT-4o의 성능 평가를 실시하였다. 본 연구의 연구진이 분류한 채팅 메시지별 인지적 참여 수준 분류 결과와 언어 모델의 분석 결과를 기반으로 각 언어 모델의 수행을 종합적으로 평가하기 위하여 정확도(accuracy), 정밀도(precision), 민감도(sensitivity), 정밀도와 민감도의 조화평균인 F1 점수(F1 score)를 산출한 다음, 4개 지표를 고려하여 각 언어 모델의 성능을 평가하였다[12, 34].

4. 연구 결과

4.1 인지적 참여 수준 0에 대한 성능

인지적 참여 수준이 0인 토론 메시지를 분류하는 GPT 모델의 성능 분석 결과는 Table 2와 같다. GPT-3.5-turbo에 대해 퓨샷러닝을 실시한 모델의 성능은 정확도와 정밀도가 0.952, 0.900으로 매우 높은 수준이었으며, 민감도와 F1 점수는 각각 0.818, 0.857로 양호한 수준으로 나타났다. GPT-3.5-turbo를 파인튜닝 한 모델의 경우, 정확도가 0.993, 정밀도와 민감도, F1 점수가 모두 0.909로 나타나 전반적으로 높은 성능을 나타냈다. GPT-4o는 파인튜닝 모델과 퓨샷러닝 모델이 동일한 결과를 나타냈으며, 정확도 0.993, 정밀도와 민감도, F1 점수는 0.917로 나타나 인지적 참여 수준 0을 분류하는 데 높은 성능을 보였다. 이러한 결과를 바탕으로 토론 주제와 관련 없는 인지적 참여 수준 0의 문장을 분류하는 GPT 성능은 GPT 모델의 종류나 데이터 훈련 정도와 관련 없이 전반적으로 뛰어나다고 볼 수 있다.

Table 2. Performance of GPT models classifying cognitive engagement level 0

GPT models	Accuracy	Precision	Sensitivity	F1 Score
GPT-3.5-turbo_few-shot	0.952	0.900	0.818	0.857
GPT-3.5-turbo_fine-tuning	0.993	0.909	0.909	0.909
GPT-4o _few-shot	0.993	0.917	0.917	0.917
GPT-4o _fine-tuning	0.993	0.917	0.917	0.917

4.2 인지적 참여 수준 1에 대한 성능

Table 3은 단편적인 의견을 제시하는 인지적 참여 수준 1의 토론 메시지를 분류하는 GPT 모델의 성능을 나타낸다. 퓨샷러닝을 실시한 GPT-3.5-turbo는 정확도가 0.770, 정밀도가 0.685, 민감도가 0.725, F1 점수가 0.705로 보통 수준의 성능임을 확인하였다. 퓨샷러닝을 실시한 GPT-4o의 경우, 정확도와 F1 점수가 각각 0.889, 0.840으로 양호한 수준이었으며 정밀도는 0.919로 높게 나타났지만, 민감도는 0.775의 보통 수준으로 확인되었다. 이에 반해 파인튜닝을 실시한 GPT 모델은 더 나은 성능을 나타냈다. 파인튜닝을 수행한 GPT-3.5-turbo는 모든 지표에서 0.942~0.959 사이의 수치를 나타냈으며, GPT-4o는 모든 지표에서 0.941~0.967 사이의 수치를 나타냈다. 이러한 결과를 토대로 GPT 모델에 대해 추가 학습을 수행한 파인튜닝 모델이 구체적인 설명없이 단편적인 의견을 제시하는 인지적 참여 수준 1의 메시지를 보다 정확하게 분류할 수 있음을 확인하였다.

Table 3. Performance of GPTs classifying cognitive engagement level 1

GPT models	Accuracy	Precision	Sensitivity	F1 Score
GPT-3.5-turbo_few-shot	0.770	0.685	0.725	0.705
GPT-3.5-turbo_fine-tuning	0.959	0.942	0.951	0.946
GPT-4o _few-shot	0.889	0.919	0.775	0.840
GPT-4o _fine-tuning	0.967	0.970	0.941	0.955

4.3 인지적 참여 수준 2에 대한 성능

인지적 참여 수준 2에 해당하는 메시지를 분류하는 모델의 성능 분석 결과는 Table 4와 같다. 퓨샷러닝을 실시한 GPT-3.5-turbo에서 정확도는 0.774, F1 점수는 0.751로 나타나 전반적인 성능이 보통 수준으로 나타났다. 퓨샷러닝을 수행한 GPT-4o의 경우, 정확도가 0.830, F1 점수가 0.838로 나타나 퓨샷러닝 기반의 GPT-3.5-turbo보다 나은 성능을 보였으며 전반적으로 높은 성능을 나타냈다. 파인튜닝을 실시한 GPT-3.5-turbo의 정확도는 0.948, F1 점수는 0.949였으며, GPT-4o의 정확도는 0.952, F1 점수는 0.953으로 매우 높은 것으로 나타났다. 이러한 결과는 추가적인 데이터 세트로 훈련한 GPT 모델이 프롬프트로 예시 데이터만 제공한 기본 GPT 모델과 비교하여 구체적인 의견을 제시하는 인지적 참여 수준을 높은 정확성으로 분류하고 있음을 나타낸다.

Table 4. Performance of GPTs classifying cognitive engagement level 2

GPT models	Accuracy	Precision	Sensitivity	F1 Score
GPT-3.5-turbo_few-shot	0.774	0.852	0.672	0.751
GPT-3.5-turbo_fine-tuning	0.948	0.942	0.956	0.949
GPT-4o _few-shot	0.830	0.821	0.856	0.838
GPT-4o _fine-tuning	0.952	0.950	0.957	0.953

4.4 인지적 참여 수준 3에 대한 성능

Table 5는 인지적 참여 수준 3에 해당하는 메시지를 분류하는 모델의 성능 분석 결과를 나타낸다. 모든 모델에서 정확도는 0.885~0.985 사이로 상당히 높은 수준을 나타냈다. 그러나 퓨샷러닝을 실시한 GPT-3.5-turbo와 GPT-4o는 각각 정밀도가 0.310, 0.393으로 미흡한 수준으로 나타났으며, 민감도도 0.450, 0.611로 낮은 수준으로 확인되었다. 파인튜닝을 실시한 모델은 퓨샷러닝 모델보다는 나은 성능을 나타냈지만, 다른 인지적 참여 수준에 비해 높은 성능에 도달하지 못하였다. 파인튜닝한 GPT-3.5-turbo의 정밀도와 민감도는 각각 0.810, 0.850으로 확인되었고, GPT-4o도 민감도가 0.833, F1 점수가 0.882로 양호한 수준에 그쳤다. 이러한 결과는 퓨샷러닝 모델에 비해 파인튜닝 모델이 인지적 참여 수준 3을 분류하는 성능이 더 높지만, 데이터로 추가 훈련 한 모델도 여러 아이디어를 통합하는 고차적 수준의 메시지를 분류하는 정확성이 매우 높은 수준은 아님을 나타낸다. 이와 관련하여 인지적 참여 수준 3이 주로 어떤 수준과 혼동되었는지 확인한 결과, 일부 메시지가 수준 2로 오분류되는 경향을 보였다. 구체적으로, 여러 메시지를 종합하는 '요약하면, 종합하면, 의견에 덧붙이자면, 의견을 정리하면'과 같이 통합의 의미를 가진 표현이 포함된 경우에는 수준 3으로 바르게 분류되었으나, 이러한 패턴 없이 앞선 토론 내용에 대해 확장된 의견을 제시한 경우에는 수준 2로 분류되는 한계를 보였다. 이러한 결과는 특정 패턴이 부재하거나 메시지 간 의미의 통합을 포착하는 데 생성형 AI가 한계를 가짐을 시사한다.

Table 5. Performance of GPTs classifying cognitive engagement level 3

GPT models	Accuracy	Precision	Sensitivity	F1 Score
GPT-3.5-turbo_few-shot	0.885	0.310	0.450	0.367
GPT-3.5-turbo_fine-tuning	0.974	0.810	0.850	0.829
GPT-4o _few-shot	0.911	0.393	0.611	0.478
GPT-4o _fine-tuning	0.985	0.938	0.833	0.882

4.5 인지적 참여 수준에 따른 성능 비교

다음으로, 메시지의 인지적 참여 수준에 따라 GPT 모델이 메시지를 분류하는 성능을 비교하였다. 특히 성능 확인을 위해 산출한 네 가지 지표 중에서 성능 비교에 일반적으로 사용하는 정확도와 정밀도 및 민감도의 조화평균으로 산출되는 F1 점수를 중심으로 인지적 참여 수준 간 GPT 모델의 분류 성능을 비교하였다. 인지적 참여 수준에 따른 GPT 모델의 정확도를 비교한 결과는 Figure 2, F1 점수를 비교한 결과는 Figure 3과 같다.

인지적 참여 수준 0의 경우, 모든 모델에서 0.952 이상의 높은 정확도가 나타났으며, F1 점수도 0.857 이상의 수치가 나타났다. 이러한 결과는 토론 주제와 무관한 메시지 내용이 제시되었을 때 GPT 모델이 가진 기본적인 성능이나 추가 데이터의 훈련 여부에 상관없이 토론 주제와의 연관성을 분석하는 GPT의 성능은 상당히 높은 수준인 것으로 해석할 수 있다.

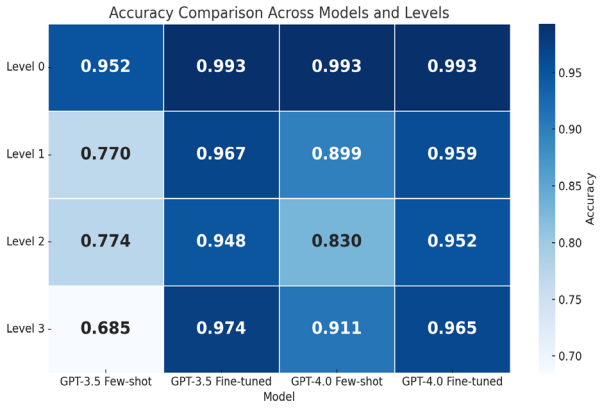
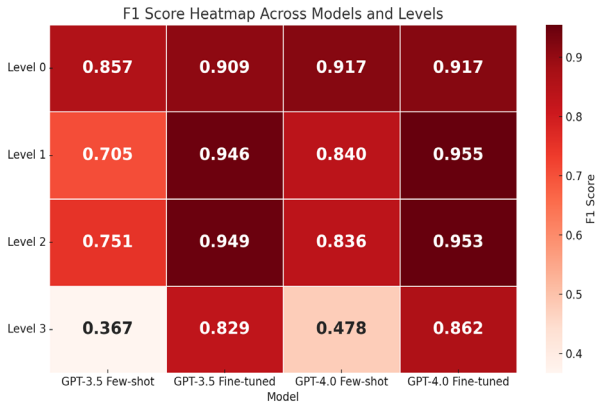


Figure 2. Accuracy comparison across models and levels



Figuer 3. F1 score comparison across models and levels

반면, 인지적 참여 수준 1, 2, 3에서는 공통적으로 토론 데이터로 추가 훈련을 실시한 파인튜닝 모델이 기본 모델에 프롬프트로 4~5개의 예시만 제시한 퓨샷러닝 모델보다 높은 성능을 나타냈다. 구체적으로, 인지적 참여 수준 1과 2에서는 파인튜닝 모델과 퓨샷러닝 모델 간의 정확도 및 F1 점수의 차이가 0.060~0.241 사이로 나타났다. 또한 퓨샷러닝을 실시한 GPT-3.5-turbo를 제외한 나머지 모델은 0.830 이상의 정확도와 0.836 이상의 F1 점수를 나타내 인지적 참여 수준 0을 분류한 퓨샷러닝 모델의 성능보다는 낮지만, 양호한 수준 이상으로 인지적 참여 수준 1, 2를 분류할 수 있음을 확인하였다. 특히 파인튜닝을 실시하면 그 성능은 매우 높아짐을 확인할 수 있다.

이와 다르게 아이디어의 통합과 고차적 수준의 메시지를 나타내는 인지적 참여 수준 3을 분류한 모델의 성능은 다른 인지적 참여 수준과 상이한 성능을 나타냈다. 정확도 측면에서는 퓨샷러닝을 실시한 GPT-3.5-turbo가 보통 수준으로 나타난 것 이외에 다른 모델은 높은 수준으로 나타나, 인지적 참여 수준 3인 메시지를 정확하게 분류할 수 있는 것으로 이해할 수 있다. 하지만 F1 점수를 고려할 때, 퓨샷러닝 모델은 매우 미흡한 수치를 나타내 기본 모델에 소수의 사례를 제시하는 방식으로는 인지적 참여 수준 3을 제대로 분류하지 못한다고 볼 수 있다. 파인튜닝 모델의 F1 점수는 0.800 이상의 수치를 나타냈지만, 다른 인지적 참여 수준의 F1 점수와 비교할 때 낮은 수치임을 확인할 수 있다. 이러한 결과는 GPT 모델을 추가 데이터로 학습시키더라도 인지적 참여 수준 3에 속하는 고차적 수준의 메시지를 분류하는 성능은 다른 인지적 참여 수준보다 미흡할 수 있음을 나타낸다. 다만, 파인튜닝 과정에서 수준 3의 데이터 수가 다른 수준에 비해 상대적으로 적었다는 점을 감안하면 데이터의 불균형으로 인해 성능이 낮게 나타났을 가능성도 고려할 필요가 있다.

5. 결론 및 논의

본 연구는 GPT-3.5-turbo와 GPT-4o가 인지적 참여 수준에 따라 협력 과정에서 생성된 텍스트 데이터를 분류하는 성능을 확인하고자 하였다. 이를 위해 기본 모델에 여러 예시를 제공하는 퓨샷러닝 기법과 데이터 세트로 모델을 추가 훈련하는 파인튜닝 기법을 적용하여 각 인지적 참여 수준을 분류하는 모델의 성능을 비교하였다. 주요 결과는 다음과 같다.

토론 주제와 무관한 인지적 참여 수준 0의 메시지를 분류하는 GPT 성능은 GPT 모델의 종류나 데이터 추가 학습 여부와 관련 없이 전반적으로 높은 것으로 나타났다. 이러한 결과는 게임 기반 협력학습에서 오프 태스크(off-task) 메시지를 분류하는 LLM의 성능이 보통 수준에 그쳤다는 Carpenter 외[36]의 결과와 상반된다. 이러한 차이는 해당 연구에서 사용한 word2vec, BERT와 같은 전통적인 임베딩 기반 모델에 비해 본 연구에서 사용한 LLM 모델이 문맥을 이해하고 자연어 처리 작업에 더 높은 성능을 보이는 특성에 기인한 것으로 해석할 수 있다. 또한 주제와 무관한 메시지는 일반적으로 일상적 대화나 감정 표현과 같은 언어적 특징을 가지고 있기 때문에 LLM이 쉽게 식별할 수 있었다고 추측할 수 있다. 본 결과를 바탕으로 협력학습 맥락에서 토론 주제와 관련 없는 낮은 수준의 인지적 참여 메시지를 분류하는 생성형 AI의 성능은 LLM 모델이나 데이터 추가 훈련 여부가 큰 영향을 미치지 않는다는 것을 확인하였다.

단편적인 의견을 제시하는 인지적 참여 수준 1과 구체적으로 설명을 포함하는 인지적 참여 수준 2를 분류하는 GPT 성능의 경우, 퓨샷러닝을 실시한 GPT-4o와 파인튜닝을 실시한 GPT-3.5-turbo, GPT-4o의 성능은 상당히 높은 것으로 나타났으나 퓨샷러닝을 실시한 GPT-3.5-turbo 모델은 보통 수준의 성능을 나타냈다. 이러한 결과는 GPT 모델 간의 차이에 기인한 것으로 볼 수 있다. GPT-4o는 GPT-3.5-turbo에 비해 한 번에 처리할 수 있는 토큰 한도가 높기 때문에 복잡한 언어 구조와 맥락을 이해하고 처리할 수 있다는 특징을 가진다[37]. 이에 GPT-4o에 소수의 사례만 제공한 경우에도 인지적 참여 수준 1, 2를 분류하는 데 높은 수준의 정확도와 F1 점수가 나타난 것으로 해석할 수 있다.

인지적 참여 수준 3에 대한 GPT 성능을 검증한 결과, 파인튜닝을 실시한 모델이 퓨샷러닝을 실시한 모델보다 월등하게 높은 성능을 보였으며, GPT-3.5-turbo와 GPT-4o 간의 성능 차이는 미미한 것으로 나타났다. 이러한 결과는 교실 담화의 유형을 분류하기 위해 파인튜닝을 실시한 LLM 모델이 퓨샷러닝 기반 모델보다 높은 성능을 보였다는 선행연구 결과를 지지한다[38]. 인지적 참여 수준 1, 2와 다르게, 인지적 참여 수준 3에서 모델의 종류와 관계없이 파인튜닝 여부에 따라 성능에 차이가 나타난 이유는 인지적 참여 수준 3에 해당하는 메시지의 내용에 기인하는 것으로 볼 수 있다. 인지적 참여 수준 3에 해당하는 메시지

는 여러 그룹원이 제시한 아이디어를 통합하고 의견을 연결, 확장하는 메시지이기 때문에 인지적 참여 수준 3을 분류하기 위해서는 토론 메시지 하나에 포함된 의미뿐만 아니라 앞서 제시된 메시지들의 맥락과 의미를 파악하는 것이 필요하다. 이에 인지적 참여 수준 3의 경우에는 프롬프트로 몇 개의 사례를 제시하는 것에는 한계가 있으며, 충분한 추가 데이터로 학습해야만 수준 3에 속하는 메시지를 정확하게 분류할 수 있다고 해석할 수 있다.

본 연구 결과를 토대로 CSCL에서 나타나는 인지적 참여 수준을 분석하는 데 효과적으로 생성형 AI를 활용하기 위한 시사점을 도출할 수 있다. 첫째, 분류하고자 하는 인지적 참여 수준에 따라 적합한 LLM 모델을 선정할 필요가 있다. 본 연구에서 생성형 AI를 활용하여 분류하고자 하는 인지적 참여의 수준에 따라 필요한 LLM 모델이나 데이터 훈련 정도가 다르다는 것을 확인하였다. 물론 모든 인지적 참여 수준에서 파인튜닝을 실시한 GPT-4o의 성능이 가장 높았지만, 현재(2025년 2월 기준) GPT-4o는 유료 서비스로 제공되고 있으며 GPT-3.5-turbo는 ChatGPT를 통해 무료로 사용할 수 있다. 이에 비용적 측면을 고려할 때 무조건적으로 상위 버전의 모델을 선택하기 보다는 분류 목적에 따라 비용효과적인 접근을 취할 필요가 있다.

둘째, 분류하는 인지적 참여 수준에 따라 데이터의 훈련 정도를 달리할 필요가 있다. 본 연구에서 실시한 퓨샷러닝과 파인튜닝은 요구되는 데이터 양이나 실시 방법에 상당한 차이가 있다. 퓨샷러닝은 ChatGPT를 사용하여 프롬프팅을 통해 실시할 수 있으므로 접근성이 높으며, 소수의 사례를 제시하기 때문에 데이터 수집에 대한 어려움이 낮다고 볼 수 있다. 반면, 파인튜닝은 GPT의 API를 활용해야 하기 때문에 초기 접근성이 상대적으로 높으며, 상당한 양의 데이터를 수집해야 한다는 점에서 시간소모적이다. 따라서 모든 인지적 참여 수준을 분류하기 위해 파인튜닝을 실시하기보다는, 퓨샷러닝으로 텍스트 분류에 한계가 있는 고차적 인지적 참여 수준에 대해 파인튜닝을 수행할 수 있다. OpenAI에서도 ChatGPT를 통해 원샷러닝(one-shot learning), 퓨샷러닝 등을 실시한 다음, 원하는 결과가 나오지 않을 때 파인튜닝을 실시할 것을 권장하고 있다[37].

셋째, 생성형 AI가 높은 성능을 낼 수 있도록 효과적인 프롬프트를 설계할 필요가 있다. 생성형 AI는 순환적 과정을 통해 프롬프트를 설계해야 한다. 프롬프트가 구체적이고 명확할수록 사용자가 원하는 결과를 얻을 수 있으므로, 시스템이 수행해야 하는 역할과 토론 맥락, 인지적 참여 수준을 구분하는 명확한 기준과 적절 수의 예제, 수행해야 하는 과제를 명료하게 프롬프트에 포함할 필요가 있다. 더욱이 프롬프트를 설계하는 과정은 순환적 과정으로[39], 반복적인 점검과 개선의 과정을 통해 최적의 프롬프트를 작성하는 것이 요구된다.

넷째, 생성형 AI가 인지적 참여 수준을 자동으로 분석하는 결과를 바탕으로 현장 교사는 협력 그룹의 특성에 적합한 실시간 피드백을 제공할 수 있다. 예를 들어, 인지적 참여 수준 0이 발생한 그룹에는 토론 주제에 초점을 두도록 유도하는 피드백을, 인지적 참여 수준 1이 높은 그룹에는 정교화된 설명과 구체적인 이유를 제시하도록 촉진하는 피드백을 제공할 수 있다. 또한 인지적 참여 수준 2의 비율이 높은 그룹에는 학습자들이 서로 아이디어를 연결하고 통합하는 방향으로 논의를 이끌 수 있는 피드백을 제공할 수 있다. 이렇듯 생성형 AI가 자동 분석한 데이터를 활용한 맞춤형 피드백은 교수자가 즉각적으로 협력학습 과정을 확인하기 어려운 동시적 CSCL 환경에서 협력 과정을 점검하고 지원하는 데 효과적으로 기여할 수 있다.

본 연구는 인지적 참여 수준을 분류하는 LLM 모델을 개발하고 모델의 성능 평가 결과를 바탕으로 CSCL 맥락에서 생성형 AI를 활용할 수 있는 가능성을 제시하였다는 데 의의가 있다. 그럼에도 불구하고 본 연구는 다음과 같은 제한점을 가지며, 후속 연구에 대한 제언은 다음과 같다. 첫째, 본 연구는 인지적 참여 수준에 따른 LLM 모델의 성능을 확인한 연구로, CSCL에서 생성형 AI를 활용하기 위한 기초 연구로 볼 수 있다. 추후 연구에서는 LLM 모델이 분류한 결과를 학습자에게 제공하거나 해당 정보를 바탕으로 교수자가 실시간 피드백을 제공하는 등 협력학습 과정을 개선할 수 있는 정보를 제공하는 교수적 처방에 적용될 필요가 있다. 둘째, 본 연구는 채팅 기반의 동시적 CSCL이라는 특정 맥락에서 나타나는 인지적 참여 수준을 분류하는 LLM 모델의 성능을 비교하였기 때문에 본 연구 결과를 일반화하는데 한계가 있다. 추후 연구에서는 보다 다양한 토론 주제나 화상 토론 등 여러 학습 맥락에서 인지적 참여 수준을 분류하는 LLM 모델의 성능을 검증하는 것이 필요하다.

참고문헌

[1] Stahl, G., Koschman, T., & Suthers, D. (2014). Computer-supported collaborative learning. In K. Sawyer (Ed.), *The cambridge handbook of the Learning Sciences (2nd ed., pp. 479-500)*. Cambridge University Press. https://doi.org/10.1017/CBO9781139519526

[2] Liu, S., Liu, S., Liu, Z., Peng, X., & Yang, Z. (2022). Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement. *Computers & Education, 181*, 104461. https://doi.org/10.1016/j.compedu.2022.104461

[3] Zhu, G., Xing, W., Costa, S., Scardamalia, M., & Pei, B. (2019). Exploring emotional and cognitive dynamics of knowledge building in grades 1 and 2. *User Modeling and User-Adapted Interaction, 29*, 789-820. https://doi.org/10.1007/s11257-019-09241-8

[4] Strauß, S., & Rummel, N. (2021). Promoting regulation of equal participation in online collaboration by combining a group awareness tool and adaptive prompts. But does it even matter?. *International Journal of Computer-Supported Collaborative Learning, 16*, 67-104. https://doi.org/10.1007/s11412-021-09340-y

[5] Cornide-Reyes, H., Riquelme, F., Monsalves, D., Noel, R., Cechinel, C., Villarroel, R., Ponce, F., & Munoz, R. (2020). A multimodal real-time feedback platform based on spoken interactions for remote active learning support. *Sensors, 20*(21), 6337. https://doi.org/10.3390/s20216337

[6] Jin, S. H. & Yoo, M. (2020). The effects of feedback types for online participation information on online participation and learning outcomes. *Journal of Educational Technology, 36*(3), 561-588.

[7] Sullivan, F. R., & Keith, P. K. (2019). Exploring the potential of natural language processing to support microgenetic analysis of collaborative learning discussions. *British Journal of Educational Technology, 50*(6), 3047-3063. https://doi.org/10.1111/bjet.12875

[8] Cai, Z., Eagan, B., Dowell, N., Pennebaker, J., Shaffer, D., & Graesser, A. (2017). Epistemic network analysis and topic modeling for chat data from collaborative learning environment. *In Proceedings of the 10th international conference on educational data mining*. 104-111.

[9] Brantley, P., Chang, Y., Voicu, R., C. (2024). Its All About Context. *In Proceedings of SoutheastCon 2024, Atlanta, GA, USA*, 1382-1387. https://doi.org/10.1109/SoutheastCon52093.2024.10500277

[10] Gunawardena, C., Chen, Y., Flor, N., & Sánchez, D. (2023). Deep learning models for analyzing social construction of knowledge online. *Online Learning, 27*(4), 69-92. https://doi.org/10.24059/olj.v27i4.4055

[11] Hou, C., Zhu, G., Zheng, J., Zhang, L., Huang, X., Zhong, T., Li, S., Du, H., & Ker, C. L. (2024). Prompt-based and fine-tuned gpt models for context-dependent and-independent deductive coding in social annotation. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, Japan, 518-528. https://doi.org/10.1145/3636555.3636910

[12] Suraworachet, W., & Cukurova, M. (2024). Predicting challenge moments from students' discourse: A comparison of large language models to other natural language processing approaches. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, Japan, 473-485. https://doi.org/10.1145/3636555.3636905

[13] Guo, L., Du, J., & Zheng, Q. (2023). Understanding the evolution of cognitive engagement with interaction levels in online learning environments: Insights from learning analytics and epistemic network analysis. *Journal of Computer Assisted Learning, 39*(3), 984-1001. https://doi.org/10.1111/jcal.12781

[14] Ouyang, F., & Chang, Y. H. (2019). The relationships between social participatory roles and cognitive engagement levels in online discussions. *British Journal of Educational Technology, 50*(3), 1396-1414. https://doi.org/10.1111/bjet.12647

[15] Sinha, S., Rogat, T. K., Adams-Wiggins, K. R., & Hmelo-Silver, C. E. (2015). Collaborative group engagement in a computer-supported inquiry learning environment. *International Journal of Computer Supported Collaborative Learning, 10*(3), 273–307. https://doi.org/10.1007/s11412-015-9218-y

[16] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez - Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90-112. https://doi.org/10.1111/bjet.13370

[17] Yang, Y., Yuan, K., Feng, X., Li, X., & van Aalst, J. (2022). Fostering low - achieving students' productive disciplinary engagement through knowledge - building inquiry and reflective assessment. *British Journal of Educational Technology, 53*(6), 1511-1529. https://doi.org/10.1111/bjet.13267

[18] Cooper, K. S. (2014). Eliciting engagement in the high school classroom: A mixed-methods examination of teaching practices. *American Educational Research Journal, 51*(2), 363-402. https://doi.org/10.3102/0002831213507973

[19] Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109. https://doi.org/10.3102/00346543074001059

[20] Greene, B. A., & Miller, R. B. (1996). Influences on achievement: Goals, perceived ability, and cognitive engagement. *Contemporary Educational Psychology, 21*(2), 181-192. https://doi.org/10.1006/ceps.1996.0015

[21] Zhu, E. (2006). Interaction and cognitive engagement: An analysis of four asynchronous online discussions. *Instructional Science, 34*, 451-480. https://doi.org/10.1007/s11251-006-0004-0

[22] Jin, M., Kim, H. J., Lim, J. Y., & Lim, K. Y. (2022). Relationships between cognitive engagement and academic achievement in online learning: Using penalized regression approach. *Journal of Korean Association for Educational Information and Media, 28*(3), 791-820. https://doi.org/10.15833/KAFEIAM.28.3.791

[23] Tan, S. C., So, H. J., & Chai, C. S. (2011). Methodological considerations for quantitative content analysis of online interactions. In B. K. Daniel (Ed.), *Handbook of research on methods and techniques for studying virtual communities: Paradigms and phenomena* vol 2 (pp. 611-630). IGI Global. https://doi.org/10.4018/978-1-60960-040-2.ch037

[24] Ferreira, M., Rolim, V., Mello, R. F., Lins, R. D., Chen, G., & Gašević, D. (2020). Towards automatic content analysis of social presence in transcripts of online discussions. Proceedings of the tenth International Conference on Learning Analytics & Knowledge, Germany, 141-150. https://doi.org/10.1145/3375462.3375495

[25] Yang, D., Wen, M., Howley, I., Kraut, R., & Rose, C. (2015). Exploring the effect of confusion in discussion forums of massive open online courses. Proceedings of the second (2015) *ACM conference on learning scale, Canada*, 121-130. https://doi.org/10.1145/2724660.2724677

[26] Van Dis, E., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. (2023). ChatGPT: Five priorities for research. *Nature, 614*(7947), 224-226. https://doi.org/10.1038/d41586-023-00288-7

[27] Alberts, I., Mercolli, L., Pyka, T., Prenosil, G., Shi, K., Rominger, A., & Afshar-Oromieh, A. (2023). Large language models (LLM) and ChatGPT: What will the impact on nuclear medicine be?. *European Journal of*

Nuclear Medicine and Molecular Imaging, 50(6), 1549-1552. https://doi.org/10.1007/s00259-023-06172-w

[28] Lim, C., Go, B., Jeong, Y., Lee, E., Jeon, M., & Kim, S. (2023). Exploring the use of a generative ai-based chatbot for instructional design: Focused on chatgpt and RPISD model. *Journal of Educational Technology, 36(special)*, 1213-1248. https://doi.org/10.17232/KSET.39.4.1213

[29] Kang, S., & Heo, H. (2023). Development and application of generative AI-based instructional design platform. *The Journal of Korean Association of Computer Education, 26*(6), 143-153. https://doi.org/10.32431/kace.2023.26.6.012

[30] Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of educational computing research, 17*(4), 397-431. https://doi.org/10.2190/7MQV-X9UJ-C7Q3-NRAG

[31] Ahmad, M., Junus, K., & Santoso, H. B. (2022). Automatic content analysis of asynchronous discussion forum transcripts: A systematic literature review. *Education and Information Technologies, 27*(8), 11355-11410. https://doi.org/10.1007/s10639-022-11065-w

[32] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychosocial Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

[33] OpenAI (2023). *GPT-4 technical report*. https://arxiv.org/abs/2303.08774

[34] Chiu, K. L., Collins, A., & Alexander, R. (2021). Detecting hate speech with gpt-3. arXiv preprint arXiv:2103.12407. https://doi.org/10.48550/arXiv.2103.12407

[35] Ren, Z., Caputo, A., & Jones, G. (2024). A Few-shot Learning Approach for Lexical Semantic Change Detection Using GPT-4. *In Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change.* 187-192. https://doi.org/10.18653/v1/2024.lchange-1.18

[36] Carpenter, D., Emerson, A., Mott, B. W., Saleh, A., Glazewski, K. D., Hmelo-Silver, C. E., & Lester, J. C. (2020). Detecting off-task behavior from student dialogue in game-based collaborative learning. *Proceedings in Artificial Intelligence in Education: 21st International Conference*. 55-66. Springer International Publishing. https://doi.org/10.1007/978-3-030-52237-7_5

[37] OpenAI (n.d.). https://openai.com/research/gpt-4

[38] Wang, H., Huang, T., Zhao, Y., & Hu, S. (2023). The impact of dashboard feedback type on learning effectiveness, focusing on learner differences. *Sustainability, 15*(5), 4474. https://doi.org/10.3390/su15054474

[39] Han, J., Koo, Y. L., & Kim, S. J. (2023). Personalized feedback with ChatGPT: A study on its generation and effectiveness. *The Journal of Educational Information and Media, 29*(4), 1123-1151. https://doi.org/10.15833/KAFEIAM.29.4.1123

진명화

· 2008 대구대학교 영어교육과(문학사)
· 2017 이화여자대학교 교육공학과 (석사)
· 2024 이화여자대학교 교육공학과 (박사)
· 2009 ~ 2019 경상북도교육청 소속 중등교사
· 2024 ~ 현재 이화여자대학교 교육공학과 박사후 연구원

관심분야 : CSCL, 협력학습분석학, 맞춤형 학습

mjin@ewha.ac.kr

박은주

· 2008 한국교원대학교 지리교육과/국어교육과(교육학사)
· 2024 이화여자대학교 교육공학과(석사)
· 2024 ~ 현재 천재교과서 재직

관심분야 : CSCL, 지능형 튜터링 시스템(ITS), 개별화 학습

joo1004@ewha.ac.kr

임규연

· 1997 이화여자대학교 교육공학과(문학사)
· 1999 이화여자대학교 교육공학과(석사)
· 2008 펜실베니아주립대 교육공학과(박사)
· 2012~현재 이화여자대학교 교육공학과 교수

관심분야 : 테크놀로지기반학습, 협력학습분석학, 학습경험설계(LXD)

klim@ewha.ac.kr

부 록

〈표 1〉 인지적 참여 수준 프레임워크

수준	정의	설명
수준 0	주제에서 벗어남	토론 주제와 관련 없는 내용을 제시하는 메시지
수준 1	단순한 질문, 단편적 의견 제시하기	용어나 개념의 정의에 대한 질문이나 사실적 정보에 대한 질문, 구체적인 설명을 요구하는 질문, 정교화나 정당화하는 내용없이 동의/반대 의견 제시, 구체적인 설명없이 아이디어를 제시하거나 단순한 경험이나 정보 공유, 단편적 의견을 제시하는 메시지
수준 2	정교화된 설명하기	다른 그룹원의 의견에 대한 부연 설명이나 아이디어 확장, 용어나 현상에 대한 설명이나 관계, 이유, 메커니즘 등에 대한 상세한 설명이나 구체적 실천 방안을 제안하는 아이디어를 포함하는 메시지
수준 3	통합하기	여러 그룹원이 제시한 다양한 아이디어를 통합하거나 제시된 아이디어를 고차적 수준으로 연결하고 조직화하는 메시지

[수행 역할]

당신은 채팅으로 토론한 내용을 인지적 참여 수준에 따라 분류하는 역할을 수행합니다. ...(중략)...

[수행 맥락]

4명의 학생이 '청소년이 생성형 인공지능을 사용할 때 발생하는 문제점과 해결방안'을 주제로 채팅으로 토론하였습니다. ...(중략)...

[수행 작업]

각 메시지를 [루브릭 및 예시]에 따라 인지적 참여 수준 0, 1, 2, 3 중 하나로 분류하세요.

[루브릭 및 예시]

* 인지적 참여 수준에 대한 정의 및 예시 문장 제시

[그림 1] 프롬프트 예시