

## GPTs 기반 문제해결 맞춤형 챗봇 제작 및 수학적 성능 분석

권미선(신봉초등학교, 교사)

본 연구는 폴리아의 문제해결 단계에 따라 풀이를 제공하는 GPTs 기반 맞춤형 챗봇을 제작하여 수학적 성능을 검증하였다. 우선 맞춤형 챗봇 베타 버전을 제작하여 수학적 성능을 검증한 후 대표적인 오류를 수정하여 최종 맞춤형 챗봇을 완성하였다. 완성된 맞춤형 챗봇은 초등 수학 6학년 교과서에 제시된 이미지 형태의 65개 문제 중 평균 약 57.8개를 옳게 해결하여 약 89.0%의 정답률을 보였으며, 베타 버전에 비해 약 4%p 높은 정답률을 나타냈다. 또한 그림이 문제를 해결하는 데 핵심적인 역할을 하지 않는 50개 문제의 경우 평균 45.5개를 옳게 해결하여 약 91.0%의 정답률을 보였다. 완성된 맞춤형 챗봇의 답변 중 대표적인 오류는 문제 인식 오류이며, 문제에 인식하기 어려운 그림이 사용되었거나 문제 구성이 복잡한 경우에 해당 오류가 나타났다. 다음으로 개념 혼동 오류, 문제 이해 오류 등이 나타났다. 본 연구에서 개발한 문제해결 맞춤형 챗봇은 범용적인 챗봇인 ChatGPT보다 우수한 수학적 성능을 보였다. 또한 학년 수준에 적절하도록 풀이 과정의 조정이 가능하여 학생 개별화 맞춤형 지도에 활용할 수 있으며, 누구나 제작이 가능하여 교사 개인별 수업 보조 등 수학교육에서의 다양한 활용 가능성을 엿볼 수 있다.

### I. 서론

생성형 AI의 등장과 함께 최근 몇 년간 인공지능과 관련된 기술은 급격히 발전하였으며, 제4차 산업 혁명 시대에 이러한 인공지능의 발전은 우리의 상상을 넘어 다양한 분야에서 혁신을 불러일으키고 있다. 특히 LLM(Large Language Model)을 기반으로 한 대화형 AI의 경우 전문 프로그래밍 언어를 사용하지 않고 인간이 사용하는 자연어로 작업하도록 명령을 내릴 수 있다(김덕진, 2023). 이러한 처리 방식의 발전으로 인해 교육 분야에서도 인공지능의 활용에 주목하고 있다.

\* 접수일(2024년 7월 4일), 심사(수정)일(2024년 7월 22일), 게재확정일(2024년 7월 26일)

\* MSC2020분류 : 97P80

\* 주제어 : GPTs, GPT, 맞춤형 챗봇, 문제해결, 인공지능, 대화형 튜터링 시스템

특히 자연어 처리 기술(Natural Language Processing: NLP)을 활용한 교육 방법은 누구나 쉽게 접근이 가능해 학생들에게도 새로운 경험을 제공할 수 있다. 또한 생성형 AI라는 점에서 다양한 학생들에게 맞춤형 학습을 제공할 수 있어 학생 각각의 개별적인 능력을 향상시킬 수 있는 잠재력을 가지고 있다.

현재 LLM을 기반으로 한 생성형 AI는 멀티모달(multimodal) 기술이 적용되어 이전에는 불가능했던 다양한 도구와의 연결이 가능해졌다. 이러한 기술의 발전으로 교육 분야에서도 생성형 AI를 활용할 수 있는 폭이 매우 넓어졌다. 멀티모달이란 이미지, 소리, 텍스트 등과 같이 서로 다른 형태의 데이터를 모두 학습할 수 있는 기술을 의미한다(김진중, 2024). 이러한 기술의 발전으로 사람이 음성 정보를 제공하면 이를 생성형 AI가 인식하여 텍스트로 답변하며, 원하는 경우 생성형 AI가 답한 텍스트를 음성으로 변환하여 들을 수 있다. 또한 사진을 업로드하고 생성형 AI와 사진에 관련된 대화를 통해 이를 변경할 수 있다. 이러한 발전은 자연어로 처리하기 불가능했던 미술 분야를 포함한 다양한 분야에서 큰 발전을 가져오고 있다(김덕진, 서승완, 2024).

수학교육에서도 이러한 생성형 AI를 이용하여 학생들에게 맞춤형 교육을 제공할 수 있을 것이다. 특히 선생님이 지도할 수 없는 시간에, 지도하기 어려운 부분에 관한 내용을 학생들이 학습하고자 할 때 자연어 처리가 가능한 생성형 AI 모델을 사용한다면 맞춤형 챗봇과의 상호작용을 통해 개인별 학습을 제공할 수 있다. 예를 들어 수학 교과서의 경우, 학생들이 학교에서 수업을 통해 학습하지만 수학 익힘의 경우 자학자습용으로 사용하며 시간의 한계상 학교에서 선생님과 함께 모두 학습하는 것은 어려울 수 있다. 이에 수학 익힘에 해설과 답안이 함께 제시되어 있으나 문제해결 단계에 따른 해설을 제공하는 등 구체적인 정보를 제공하기에는 지면의 한계가 있다.

특히 수학 익힘에서 문제해결 역량을 신장하기 위

해 제공된 문제의 경우 폴리아의 문제해결 단계에 따라 풀이 과정을 경험해 보는 것이 필요하나 학생 혼자 이러한 과정에 따라 문제를 해결하지 못할 수 있으며, 해설 역시 지면의 한계상 이에 대한 설명이 부족할 수 있다. 이러한 부분에 생성형 AI를 활용한다면 학생들의 학습에 도움을 줄 수 있을 것이다.

또한 2015 개정 수학과 교육과정에 이어 2022 개정 수학과 교육과정에서도 문제해결을 강조하고 있으며, 특히 교수·학습 방법에서 “문제에 주어진 조건과 정보를 분석하고 적절한 문제해결 계획을 수립하고 실행하며 문제해결 과정을 반성하도록 구체적인 발문과 권고를 제시한다(교육부, 2022, p. 44)”라는 내용을 제시하여 학생들이 문제해결 단계에 따라 문제를 해결해 볼 수 있도록 권장하고 있다. 이에 학교 수업뿐만 아니라 학생들이 수학 익힘을 해결할 때 또는 다른 상황에서도 문제해결 단계에 따라 문제를 해결하는 경험을 제공하는 것은 학생들의 수학 학습에 매우 효과적일 수 있다.

한편 생성형 AI는 다양한 위험 요소로 인해 미국 등 여러 나라에서 13세 미만의 어린이들이 이를 직접적으로 사용하지 못하게 제한하고 있으며, 우리나라의 서울특별시교육청에서 제작한 학교급별 생성형 AI 활용 지침에서도 초등학생의 경우 교사 주도로 교육적 의도에 따라 활용이 가능하며 교사 시연 중심으로 사용하도록 제시하고 있다. 만약 학생이 직접 체험하고자 할 때는 해당 연령에서 사용 가능한 서비스인 경우 또는 교사의 추가 작업을 통해 생성형 AI 산출물의 안전성을 확보할 수 있는 경우에만 사용하도록 제한하고 있다(김귀훈 외, 2023). 이에 생성형 AI를 이용할 경우 교사의 추가 작업을 통해 산출물의 안전성을 확보하는 것이 우선되어야 할 것이다.

이에 본 연구는 폴리아의 문제해결 4단계에 따라 문제 풀이를 제공해주는 GPTs(Generative pretrained transformers) 기반 맞춤형 챗봇을 개발하고 그 수학적 성능을 평가하여, 맞춤형 챗봇이 학생들의 문제해결 역량 향상에 도움을 줄 수 있는 가능성을 탐색하였다. 맞춤형 챗봇에서 생성된 산출물은 제공된 지식 안에서 답변을 하기 때문에 안정성이 확보되며, 모든 이들이 맞춤형 챗봇에 접근할 수 있어 많은 학생들이 문제해결 역량을 향상시키는 데 도움을 줄 수 있을 것으로 보인다.

## II. 이론적 배경

### 1. 교육에서의 인공지능(AI in Education: AIED)

최근 인공지능은 다양한 분야에서 혁신을 일으키고 있으며 교육 분야에서도 그 중요성이 강조되고 있다. 특히, 자연어 처리 기술은 교육에서 인공지능을 적용하는 데 핵심적인 역할을 하고 있다. 자연어 처리 기술은 인공지능이 인간의 언어를 이해하고 생성하는 기술로, 프로그래밍을 전문적으로 학습하지 않고 자연어를 이용하여 인공지능을 이용할 수 있다. 따라서 현재 사용되고 있는 다양한 인공지능은 자연어로 학생들과 쉽게 상호작용할 수 있어 교육 분야에서 맞춤형 학습을 지원하는 등 다양한 측면에 활용될 수 있다.

구체적으로 인공지능으로 단계별 맞춤형 교육 시스템을 통해 학생들의 학습을 지원하거나(예, 이지혜, 허난, 2020; Doc et al., 2023 등), 학생들의 학습한 내용을 인공지능으로 분석하고 피드백하거나(예, Afzaal et al., 2021 등), 게임 기반 환경을 사용하여 지능형 학습을 시키는(예, Yeh et al., 2019 등) 등 다양한 측면에서 활용할 수 있다. 이러한 인공지능의 교육적 활용은 행동주의의 아버지로 불리는 스키너(Skinner)의 연구에서 그 기원을 찾아볼 수 있다. 스키너가 고안한 ‘스키너의 티칭머신’은 개인 교사처럼 모든 답에 즉시 피드백하는 역할을 하였으며, AIED 분야에서 지능형 튜터링 시스템의 전조로 여겨지고 있다(Holmes et al., 2019)

이러한 교육적인 시도는 점차 컴퓨터가 발전하면서 컴퓨터 보조 수업(Computer-Assisted Learning)으로 연결되었다(Holmes et al., 2019). 대표적으로 일리노이 대학에서 PLATO라는 컴퓨터 보조 수업 시스템을 개발하여 중앙 컴퓨터에서 원격을 통해 약 1,000명의 학생들이 이를 동시에 이용할 수 있었다. 하지만 이러한 수업은 메인 프레임의 비용과 접근성으로 인하여 극소수만이 사용할 수 있었다. 또한 학습의 순서, 학생들의 행동에 대한 반응 등이 모두 이미 정해진 틀에 의해 작동하기 때문에 학생들 개인의 특성을 반영하지 못해 부작용을 일으키기도 하였다. 이러한 단점을 보완하여 제이미 카르보넬(Jaime Carbonell)의 경우 객체, 개념, 사건을 노드로, 노드 사이의 관계를 선으로 표현하는

네트워크 형태의 의미망을 활용한 SCHOLAR라는 시스템을 개발하여 활용하였다. 이 시스템을 이용할 경우, 학생들이 관련 네트워크 안에서 질문을 하면 답할 수 있어 서로 간단한 대화가 가능했다. 이러한 시스템은 당시로서는 독창적이었으며, 현재는 지능형 튜터링 시스템(Intelligent Tutoring System)의 원조로 여겨지고 있다.

현재 인공지능은 교육에 다양한 방식으로 사용되고 있으며, 대표적으로 지능형 튜터링 시스템, 대화형 튜터링 시스템, 탐구학습 시스템 등 다양한 방식으로 활용되고 있다(Holmes et al., 2019). 이중 지능형 튜터링 시스템(Intelligent Tutoring System; ITS)은 교육 분야에서 가장 오래전부터 활용해 왔으며 현재도 널리 연구되는 분야이다(손태권, 2024). 지능형 튜터링 시스템은 학습해야 할 내용적 지식에 기반하여 개발된 '도메인 모델', 가르치는 데 효과적인 접근법에 대한 지식에 기반하여 개발된 '교수 모델', 학생에 대한 지식에 기반하여 개발된 '학습자 모델'이 있으며, 이중 도메인 모델과 교수 모델은 컴퓨터 보조 수업에서도 효과적으로 구현되었으나 학습자 모델은 지능형 튜터링 시스템에서만 구현되며, 컴퓨터 보조 수업과 지능형 튜터링 시스템을 구별하는 차이점이기도 하다(Holmes et al., 2019). 구체적으로 학습자 모델은 학생들의 다양한 지식(학생들의 오개념, 학생들이 시도한 자료 등)을 기반으로 개별 학생들에게 어떤 맞춤형 학습 활동을 제시할지 결정하고 이를 제공하는 모델이다.

대화형 튜터링 시스템의 경우 최근에 더욱 활발하게 연구되는 분야로 학생들을 학습 주제에 관한 대화에 참여시킨다는 점에서 기존의 지능형 튜터링 시스템

과 차이를 보인다(선주오, 2024). 대화형 튜터링 시스템은 아직 지능형 튜터링 시스템과 같이 체계가 규정되지는 않았으나, LLM을 기반으로 한 생성형 AI의 등장으로 활용 가능성이 매우 커질 것으로 예상된다. 탐구학습 시스템의 경우, 학생이 정해져 있는 단계별 순서를 따르기보다 학습 환경의 요소를 스스로 탐구하여 지식을 적극적으로 구성하도록 지원하는 시스템을 말한다(Holmes et al., 2019).

이처럼 인공지능은 교육의 다양한 측면에서 활용되고 있으며 지속적으로 발전해가고 있다. 현재 생성형 AI 기술의 발전으로 인해 그 속도가 매우 빨라지고 있으며, '스키너의 티칭머신'부터 '탐구학습 시스템'까지 교육에서의 인공지능의 발전 경로를 살펴봄으로써 앞으로 나아가야 할 방향을 모색해 볼 수 있을 것이다.

## 2. 생성형 AI를 이용하기 위한 기술

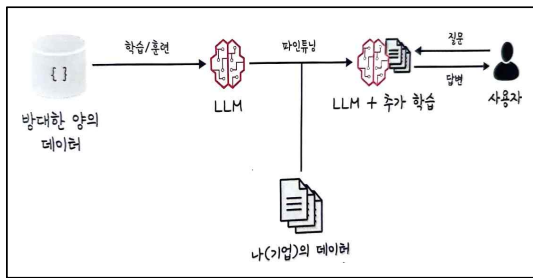
본 연구에서는 대화형 튜터링 시스템의 일종인 GPTs를 이용한 맞춤형 챗봇을 개발하여 학생들의 문제해결을 돕고자 하였다. 이를 위해서 GPTs와 같은 생성형 AI가 기반으로 하는 LLM을 활용하기 위한 기술 3가지를 집중적으로 살펴보았다([표 1] 참조).

LLM을 활용하는 방법으로는 파인튜닝(Fine-tuning), 검색 증강 생성(Retrieval-Augmented Generation: RAG), 퓨샷 러닝(Few-shot Learning) 등이 있다(김진중, 2024; 이규남 외, 2023). 구체적으로 살펴보면, 파인튜닝은 사전 훈련된 모델을 특정 작업에 맞게 조정하는 과정이다(서지영, 2024). 사전 훈련된 모델은 대규모 데이터셋을 사용하여 일반적인 패턴을 학습한 후

[표 1] LLM을 활용하기 위한 3가지 기술의 비교 (파인튜닝, 검색 증강 생성, 퓨샷 러닝)

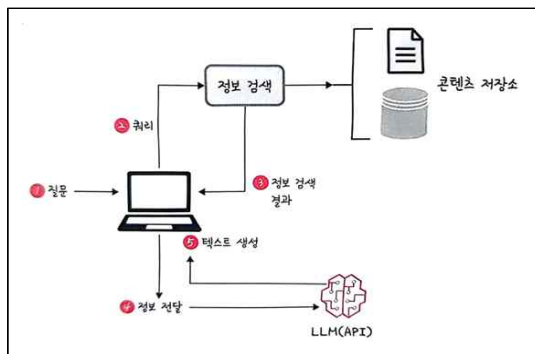
	개념	목적	장점	단점	응용 분야
파인 튜닝	사전 훈련된 모델을 특정 작업에 맞게 조정하는 과정	이미 배운 모델을 특별한 작업에 맞추기	적은 데이터와 자원으로 좋은 성능을 보임	새로운 데이터를 잘 처리하도록 조정이 필요함	질문 답변, 감정 분석
검색 증강 생성	주어진 질문에 대해 관련 문서를 검색한 후, 이러한 문서를 기반으로 응답을 생성하는 기술	정보를 찾아서 답을 만들기	최신 정보와 특정 주제에 대한 지식 사용 가능	정보 찾기와 답 만들기를 잘 연결해야 함	정보 검색, 질문 답변 시스템, 대화형 AI
퓨샷 러닝	극히 적은 수의 학습 샘플만으로 모델을 훈련시키는 기법	적은 예시로 모델 학습하기	데이터가 적은 상황에서 유용	일반화 능력 확보 어려움	의료 진단, 희귀 언어 처리, 맞춤형 추천 시스템

특정 작업에 맞게 더 작은 데이터셋으로 재훈련하며, 이 과정을 통해 사전 훈련에서 학습한 정보를 유지하면서 새로운 데이터에 적응하게 된다([그림 1] 참조). 예를 들어 교육 분야에 전문적으로 사용하기 위해서 교육 관련 문서를 추가 학습시키는 것을 파인튜닝이라고 하며, 이미 배운 기본 지식을 유지하면서 교육 분야에 더 정확하고 효과적으로 반응할 수 있다. 하지만 재훈련 비용이 많이 들고, 다른 분야에서 활용하기 위해서는 새롭게 데이터셋을 조정해야 한다는 단점이 있다.



[그림 1] 파인튜닝의 과정 (서지영, 2024, p. 45)

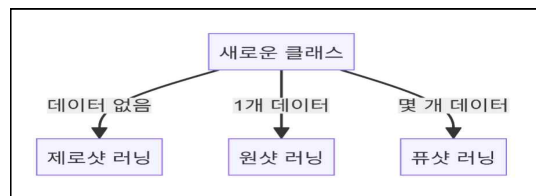
다음으로 활용할 수 있는 기술은 검색 증강 생성이다. 검색 증강 생성은 [그림 2]와 같이 정보 검색과 생성을 결합한 인공지능 기술로 먼저 주어진 질문에 대해 관련 문서를 검색한 후, 이러한 문서를 기반으로 응답을 생성한다(김진중, 2024; 허민, 2023). 따라서 검색 증강 생성은 대규모 언어 모델의 생성 능력과 정보 검색의 정확성을 결합하여, 보다 정교하고 유용한 답변을 제공하는 기술이라고 할 수 있다. 검색 증강 생성은 최신 정보나 특정 도메인 지식에 접근할 수 있다



[그림 2] 검색 증강 생성의 과정 (서지영, 2024, p. 48)

는 장점이 있으며, 이는 사전 훈련된 모델이 최신 정보를 포함하지 않을 때나 내가 원하는 지식 안에서 답변을 원할 때 유용하게 사용할 수 있다. 또한 검색 증강 생성은 검색된 문서를 기반으로 하기 때문에 모델의 응답 정확성을 높일 수 있다. 이러한 특징으로 인하여 다양한 분야에 많이 활용되고 있다. 최근 ChatGPT에서 제공하는 GPTs 역시 이 기술을 사용할 수 있어 답변의 정확도가 범용적인 챗봇보다 높다(유호석, 2024; Almeida, 2024)

마지막으로 퓨샷 러닝(Few shot Learning)은 극히 적은 수의 학습 샘플만으로 모델을 훈련 시키는 기법이다. 퓨샷 러닝과 함께 사용되는 용어는 제로샷 러닝(Zero-shot Learning), 원샷 러닝(One-shot Learning)이 있으며, 데이터의 양에 따라 구분한다([그림 3] 참조). 모델이 학습하는 과정에서 전혀 보지 않은 데이터에 대해 예측할 수 있는 것을 제로샷 러닝이라고 하며, 하나를 학습하고 예측하는 것을 원샷 러닝, 몇 개를 학습하고 예측하는 것을 퓨샷 러닝이라고 한다. 퓨샷 러닝은 특히 데이터 수집이 어려운 분야, 예를 들어 의료 진단, 희귀 언어 처리, 맞춤형 추천 시스템 등에서 큰 잠재력을 가지고 있다. 이를 통해 모델은 빠르게 새로운 작업에 적응할 수 있으며, 데이터가 제한된 상황에서도 유용하게 활용할 수 있다. 다만, 새롭고 다양한 데이터를 일반화하는 데에는 어려움을 겪을 수 있다.



[그림 3] 데이터의 양에 따른 모델 훈련 기법

본 논문에서 개발한 맞춤형 챗봇은 GPT-4o 버전에서 3가지 기술 중 검색 증강 생성과 퓨샷 러닝을 사용하여 LLM을 기반으로 한 생성형 AI를 효과적으로 조정하고자 하였으며(Chelli et., 2024), 이를 통해 초등학교 생들에게 적합한 수준의 답변을 제공하며 실제 존재하지 않는 정보를 생성하는 환각 현상(Hallucination)이 적은 챗봇을 개발하였다.

### 3. 선행 연구 분석

수학교육에서도 인공지능을 활용한 다양한 연구들이 진행되고 있으나 본 논문에서는 문제해결 맞춤형 챗봇을 제작하고 그 수학적 성능을 탐색하였으므로, 지금까지 실시된 범용적인 챗봇인 ChatGPT의 수학적 성능 및 오류 유형에 관한 연구를 구체적으로 살펴보고자 한다.

대표적인 연구로 권오남 외(2023)는 국가수준 학업성취도 평가(중·고등학생 용) 및 대학수학능력시험에 제시된 수학 문제를 활용하여 ChatGPT의 수학적 성능을 분석하였다. 이때 GPT-3.5 Turbo 버전을 사용하였으며, 그림이나 표가 필수 정보인 문제와 텍스트 기반이 아닌 문제는 제외하였으며, 선다형 문제일 경우에는 서술형 문제로 변경하여 분석하였다. 연구 결과 국가수준 학업성취도 평가 수학 문제의 정답률은 37.1%, 풀이 과정의 정확도는 5점 만점에 3.44점으로 나타났다. 분석 기준에서 3점이 'ChatGPT가 프롬프트의 맥락에 맞게 대답하였으나 답이 부분적으로 맞음', 4점이 'ChatGPT가 프롬프트의 맥락에 맞게 대답하였으나 답이 거의 맞음'임을 고려할 때 문제의 정답률에 비해 풀이 과정이 상대적으로 더 정확한 것을 알 수 있다. 오류 유형으로는 식 오류가 46.38%, 계산 오류가 24.63%, 인식 오류가 14.49%, 판단 오류가 7.25%, 수 출력 오류(단순 오타 등의 수의 오류)가 7.25%로 나타났다.

대학수학능력시험 수학 문제의 경우 정답률은 15.97%이며, ChatGPT의 대학수학능력시험 등급은 6~9등급으로 나타났다. 풀이 과정의 정확도의 경우 5점 만점에 2.49점으로 3점이 'ChatGPT가 프롬프트의 맥락에 맞게 대답하였으나 답이 부분적으로 맞음', 2점이 'ChatGPT가 프롬프트의 맥락에 맞게 대답하였으나 답이 완전히 틀림'임을 고려할 때, 어느 정도 ChatGPT가 프롬프트의 맥락에는 맞게 답함을 알 수 있다. 오류 유형으로는 식 오류가 13.82%, 계산 오류가 33.33%, 인식 오류가 15.44%, 판단 오류가 34.14%, 출력 오류(풀이 과정 출력 등의 오류)가 3.25%로 나타났다. 이는 국가수준 학업성취도 평가 수학 문제의 오류 유형별 비중과는 차이가 있으며, 대학수학능력시험 수학 문제의 난도가 더 높아 단순한 식 오류나 계산 오류보다 수학 개념 사용 및 풀이 접근 방식과 관련된 판단 오류 등의 비중이 높게 나타난 것으로 보인다.

두 시험에서 ChatGPT의 정답률은 상대적으로 낮은 것으로 보이나 풀이 과정의 정확도에서는 일정 수준의 성과를 보였다. 이를 통해 이 논문에서는 ChatGPT의 성능을 판단할 때 풀이 과정의 정확도와 오류 유형도 고려해야 함을 강조하고 있다.

한편 강윤지(2024)의 경우 초등 수학 교과서에 제시된 문장제에 대한 ChatGPT의 수학적 성능을 분석하였다. 비와 비율, 비례식과 비례배분과 관련된 2015 수학 6학년 국정 및 검정 교과서 총 11종에 대해 GPT-3.5 버전을 이용하여 총 37개 문제를 분석하였다. 이때 사용된 문제는 모두 그림이나 표에 중요한 정보가 포함되어 있지 않은 문제들이며, 문제는 연구자에 의해 직접 ChatGPT에 텍스트로 입력되었다. 분석 결과 전체 정답률은 81.08%로, 비와 비율 단원의 정답률은 90%, 비례식과 비례배분 단원의 정답률은 74.29%로 나타났다. 각 문제에 대해 5회씩 ChatGPT에게 해결을 요청하였으며, 5회 정답을 제시한 비율은 54.04%, 4회는 21.62%, 3회는 8.11%, 2회는 10.81%, 1회는 5.41%로 나타났다.

구체적으로 답이 옳은 경우 중 풀이가 옳지만 교육 과정을 벗어난 경우는 10%, 답이 옳지 않으며 풀이 과정이 교육과정을 벗어난 경우는 0%로 나타났다. 오류 유형으로는 문제 이해 오류가 6.84%, 식 세우기 오류가 9.5%, 계산 오류가 4.74%로 나타났다. 문제 이해 오류의 경우, 남학생과 여학생 수의 비와 전체 학생 수를 제시하고 여학생은 남학생보다 몇 명 더 많은지 구하는 문제에서 ChatGPT는 여학생 수와 남학생 수를 각각 구하기만 하는 오류를 보였다. 식 세우기 오류에서는 비례식을 잘못 세우거나, 기준량과 비교하는 양을 잘못 파악하여 분자와 분모를 반대로 대입하기도 하였다. 계산 오류에서는 식은 옳게 세웠으나 계산 실수로 인하여 최종 답이 틀리는 경우가 있었다.

두 논문을 비교하여 살펴보면 초등 수학 교과서에 제시된 문장제에 대한 ChatGPT의 정답률이 국가수준 학업성취도 평가나 대학수학능력시험의 수학 문제의 정답률보다 매우 높게 나타났다. 이는 초등 수학 교과서 문제에 요구되는 수학적 개념이나 절차적 지식의 난도가 낮기도 하지만 두 연구의 오류 유형을 비교해 볼 때 초등 수학 교과서에 제시된 문제의 어휘 수준이나 문장의 복잡도가 낮고 수식이 적기 때문으로 유추할 수 있을 것이다.

### III. 문제해결 맞춤형 챗봇의 개발

#### 1. 맞춤형 챗봇 베타 버전의 전반적인 개발 절차

본 연구에서는 학생의 문제해결 역량을 신장시키기 위해 OpenAI사에서 제공하는 ChatGPT의 GPTs를 이용하여 맞춤형 챗봇을 제작하였다. GPTs는 개인이 맞춤형 챗봇을 제작하고 배포할 수 있는 서비스로 맞춤형 챗봇 각각을 GPT라고 명명하며, 그러한 맞춤형 챗봇들의 모임을 GPTs라고 한다(유효석, 2024). 지금까지 OpenAI사에서는 발전된 GPT 버전을 계속하여 발표하고 있다. 권오남 외(2023)와 강윤지(2024)에서 ChatGPT를 활용한 당시에는 GPT-3.5 버전은 무료로, GPT-4 버전은 유료로 사용이 가능하였다. 하지만 2024년 5월 발표한 GPT-4o(GPT-4omni) 버전의 경우 GPTs의 대중화를 통해 일정한 한도 내에서 누구나 무료로 사용할 수 있다(김덕진, 서승완, 2024). 이에 본 논문은 GPT-4o 버전을 이용하여 GPTs 기반 맞춤형 챗봇을 제작하였으며, 따라서 본 연구에서 제작한 챗봇 역시 누구나 일정한 한도 내에서 무료로 이용할 수 있다.

맞춤형 챗봇은 우선, 문제해결 4단계에 따라 문제 풀이를 제공하도록 제작하였다. 이처럼 일정한 단계에 따라 풀이를 제공하는 것은 학생들의 학습에 도움이 될 수 있을 뿐만 아니라 프롬프트 엔지니어링 기법 중 생각의 사슬(Chain of Thought; CoT)을 사용하게 하여 생성형 AI에게 좀더 정확한 답변을 얻을 수 있다. 둘째, 학생들이 쉽게 이용할 수 있는 방식을 채택하여 문제를 텍스트가 아닌 이미지 형태로 올려도 문제에 대한 풀이를 바로 제공하도록 구성하였다. 셋째, 학생 수준에 적합한 답변을 얻기 위해 교육과정 안에서 답하도록 조정하였으며, 2015 개정 수학과 교육과정에 따른 교과서 중 문제해결 차시를 매 단원마다 명시적으로 제시한 3종의 지도서를 맞춤형 챗봇에 탑재하여 해당 자료를 이용하여 답변할 수 있도록 구성하였다.

본 논문에서는 맞춤형 챗봇을 제작할 때, 이론적 배경에서 제시한 LLM을 활용하기 위한 다양한 기법 중 검색 증강 생성을 집중적으로 이용하여 문제해결에 관해 검색할 수 있는 자료를 제공하고 그 안에서 답변할 수 있도록 구성하였다. 이를 통해 답변의 신뢰도를 높이고 환각 현상을 줄이도록 하였다. 제작된 맞춤형 챗봇은 프로그래밍 전공의 GPTs 제작 관련 전문가의 검토를 통해 수정·보완하였다. 이를 통해 맞춤형 챗봇 베타 버전을 개발하였으며, 맞춤형 챗봇의 수학적 성능을 검증하기 위해 65개 문제를 추출하여 그에 대한 정답률과 풀이 과정을 확인하였다. 이때 나타난 주요 오류를 바탕으로 맞춤형 챗봇을 수정하였다.

#### 2. 맞춤형 챗봇 베타 버전에 대한 예비 실험

맞춤형 챗봇 베타 버전의 수학적 성능을 점검하고 수정하기 위하여 2015 개정 수학과 교육과정이 반영된 국경 1종, 김정 10종, 총 11종의 수학 교과서에서 비와 비율, 비례식과 비례배분 단원의 마지막에 제시된 평가 차시 중 문장제를 추출하여 정답률과 풀이 과정을 분석하였다. 비와 비율, 비례식과 비례배분 단원의 경우, 범용적인 챗봇인 ChatGPT의 수학적 성능을 분석한 선행 연구가 있어 맞춤형 챗봇의 성능을 제시하는 것을 넘어 범용적인 챗봇과의 수학적 성능과 비교하고자 동일한 단원으로 선정하였다. 구체적으로 본 연구에서는 강윤지(2024)에서 분석한 텍스트만으로 해결이 가능한 37개 문제(예, [그림 4] 참조)와 선행 연구에서 제외되었던 그림, 표 등이 문제를 해결하는 데 필수적인 28개 문제(예, [그림 5] 참조)를 포함한 총 65문제를 분석하였다. 구체적인 문제 수는 [표 2]와 같다.

본 연구에서 개발하는 맞춤형 챗봇은 선행 연구에서 사용된 챗봇과는 달리 멀티모달 기술을 사용하여 이전에는 인식하지 못했던 그림이나 표가 포함된 문제를 해결할 수 있다. 이에 그림, 표 등이 포함된 문제

[표 2] 맞춤형 챗봇의 베타 버전과 최종 버전의 검증에 사용된 문제 수

	국경	A	B	C	D	E	F	G	H	I	J	합계
비와 비율	6	4	1	3	3	3	7	4	3	1	3	38
비례식과 비례배분	2	2	4	2	2	2	2	2	2	4	3	27
전체 단원	8	6	5	5	5	5	9	6	5	5	6	65

지 함께 분석하였다. 또한 모든 문제를 맞춤형 챗봇에 이미지 형태로 입력하였다. 학생들이 문제를 일일이 텍스트로 입력하는 것은 시간이 오래 걸리고 어려울 수 있으나 사진을 찍어 바로 업로드할 경우 문제를 직접 입력하지 않고도 빠르고 쉽게 해결이 가능하다. 이에 본 논문은 학생들이 맞춤형 챗봇을 활용한다는 것을 가정하여 수학적 성능을 살펴보고자 이미지 형태로 입력하는 방식을 선택하였다. 다만 이미지로 캡처하여 맞춤형 챗봇에 입력했을 때 화면이 깨지는 경우, 이미지의 화질에 따라 맞춤형 챗봇의 인식률이 달라질 수 있어 해당 문제는 이미지 파일로 저장하여 맞춤형 챗봇에 입력하였다.

학교 앞길을 청소하는 자원봉사자 20명 중 여자는 8명입니다. 전체 자원봉사자 수에 대한 남자 자원봉사자 수의 비율을 써 보세요.

[그림 4] 텍스트만으로 해결이 가능한 문제의 예 (교육부, 2021, p. 88)



[그림 5] 텍스트만으로 해결할 수 없는 문제의 예 (교육부, 2021, p. 89)

맞춤형 챗봇 베타 버전에 문제를 입력한 후 맞춤형 챗봇이 제시한 답변에 대해 정답률을 분석하였다. 또한 맞춤형 챗봇 베타 버전이 제시한 풀이 과정에 어떠한 오류가 있는지 [표 3]을 기준으로 분석하였다. 정답과 오답은 각각 풀이 과정이 옳은 경우와 옳지 않은 경우로 분류하였는데, 오답의 경우에는 풀이 과정이 옳은 경우가 없다고 생각할 수 있으나 맞춤형 챗봇에서 풀이 과정을 옳게 제시하고 마지막 정답만 옳지 않게 제시하거나, 문제 자체를 다르게 인식하여 오답을 제시하였으나 풀이 과정을 옳게 제시한 경우가 있어 해당 분석 기준을 사용하였다.

맞춤형 챗봇 베타 버전이 제시한 정답과 풀이 과정은 초등수학교육 박사 과정 이상의 2인이 분석하였으

며, 정답에 대한 채점은 100% 일치하였다. 풀이 과정의 분류에서는 일부 일치하지 않는 부분이 있어 이를 조정하여 일관되게 수정하였다. 예를 들어 문제에서 비를 구하라고 요구하였으나 맞춤형 챗봇은 비를 비율로 잘못 인식하였다. 이때 비율을 옳게 구하기도 하고 옳지 않게 구하기도 하였다. 이에 대해 분석자 1인은 두 경우 모두 오답으로 처리하고 문제 인식 오류로 분류하였으나 다른 1인은 비율을 옳지 않게 구한 경우는 문제 인식 오류 0.5문제, 개념 혼동 오류 0.5문제로 분류하여 논의를 통해 문제 인식 오류 0.5문제, 개념 혼동 오류 0.5문제로 조정하였다. 검증에 사용된 65개 문제 중 일부 문제는 두 가지의 답을 요구하고 있어 해당 문제에 대해서는 각각을 0.5문제로 정답률과 풀이 과정을 분석하였으며, 풀이 과정에서 두 가지 이상의 오류를 보인 경우에는  $1/(\text{오류 유형의 수})$ 로 분석하였다. 또한 분석 결과는 모두 소수 둘째 자리에서 반올림하여 제시하였다.

[표 3] 분석 기준

		분석 기준	
정답	풀이 과정이 옳은 경우	학년 수준에 적합한 풀이	학년 수준에 적합하지 않은 풀이
	풀이 과정이 옳지 않은 경우	문제 이해 오류	개념 혼동 오류
오답	풀이 과정이 옳은 경우	학년 수준에 적합한 풀이	학년 수준에 적합하지 않은 풀이
	풀이 과정이 옳지 않은 경우	문제 인식 오류	문제 이해 오류
		개념 혼동 오류	단순 계산 오류
출력 오류			

### 3. 맞춤형 챗봇 베타 버전의 수학적 성능 및 수정

맞춤형 챗봇 베타 버전의 수학적 성능을 평가한 결과, 전체 문제에 대한 정답률은 80% 이상으로 나타났다([표 4] 참조). 구체적으로 총 65문제 중 55문제를 옳게 해결하여 84.6%의 정답률을 보였으며, 9문제를 옳지 않게 해결하여 15.4%의 오답률을 보였다. 기타의 경우 1문제(1.5%)로 나타났으며, 그림으로 제시된 도형의 수를 비율로 나타내는 문제에서 도형의 수를 인식하지 않고 맞춤형 챗봇이 스스로 예를 들어 도형의 수

를 제시하고 그 수에 맞는 비율을 제시하여 기타로 분류하였다.

[표 4] 맞춤형 챗봇 베타 버전의 정답률 (N=65)

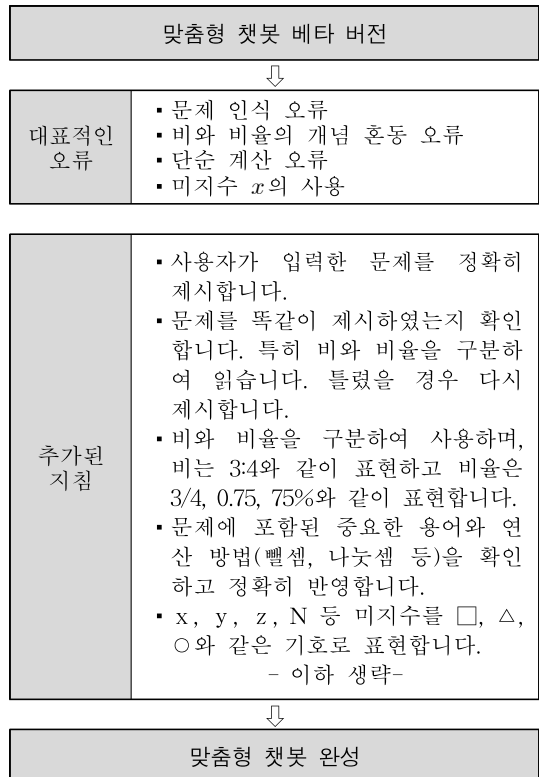
	정답	오답	기타	합계
문제 수	55 (84.6%)	9 (15.4%)	1 (1.5%)	65 (100%)

맞춤형 챗봇 베타 버전의 풀이 과정을 유형별로 분류한 결과는 [표 5]와 같다. 정답의 경우 대부분의 풀이 과정이 학년 수준에 적합한 것으로 나타났지만 일부 문제에서 미지수  $x$ 를 사용하는 등 학년 수준에 적합하지 않은 풀이를 제공하는 것으로 나타났다. 오답인 경우 대부분의 풀이 과정은 학년 수준에 적합한 것으로 나타났으나 1문제에서 미지수  $x$ 를 사용하여 풀이를 제시하였다. 풀이 과정에서 나타난 대표적인 오류로는 문제 인식 오류로 그림을 옳게 인식하지 못하거나 비를 비율로 인식하거나 그림으로 제시된 글과 수를 다르게 읽는 경우가 있었으며, 비율과 비의 개념을 혼동하는 오류나 단순 뺄셈에서 차를 잘못 구하는 계산 오류를 보이기도 하였다.

[표 5] 맞춤형 챗봇 베타 버전의 풀이 과정의 유형 (N=64, 기타는 제외)

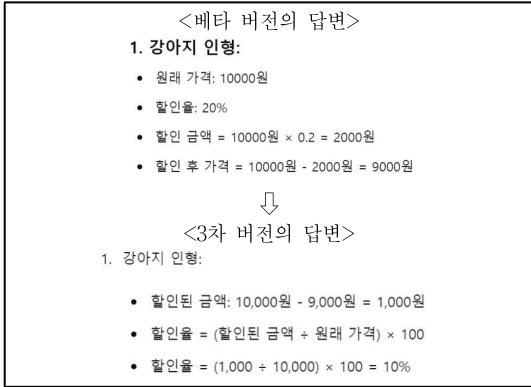
		분석 기준		문제수
정답	풀이 과정이 옳은 경우	학년 수준에 적합한 풀이	43 (66.2%)	
		학년 수준에 적합하지 않은 풀이	12 (18.5%)	
오답	풀이 과정이 옳지 않은 경우	문제 이해 오류	0	
		개념 혼동 오류	1 (1.5%)	
오답	풀이 과정이 옳은 경우	학년 수준에 적합한 풀이	8 (12.3%)	
		학년 수준에 적합하지 않은 풀이	1 (1.5%)	
	풀이 과정이 옳지 않은 경우	문제 인식 오류	5.5 (8.5%)	
		문제 이해 오류	1 (1.5%)	
		개념 혼동 오류	2 (3.1%)	
		단순 계산 오류	0.5 (0.8%)	
출력 오류	0			

이러한 오류 유형을 바탕으로 새롭게 지침을 추가하여 맞춤형 챗봇 베타 버전을 수정하였으며, 수정 절차는 [그림 6]과 같다. 구체적으로 문제 인식 오류를 해결하기 위하여 맞춤형 챗봇에게 입력한 문제를 정확히 제시하고 똑같이 제시하였는지 확인하도록 요구하였다. 또한 개념 혼동 오류를 수정하기 위하여 맞춤형 챗봇이 자주 혼동하는 용어인 비와 비율을 구분하도록 요구하였으며, 비와 비율의 의미를 제공하였다. 단순 계산 오류를 수정하기 위하여 문제에 포함된 중요한 용어와 연산 방법을 확인하고 정확히 반영하도록 요구하였다. 풀이 과정의 경우 학년 수준에 적합한 풀이를 할 수 있도록 미지수를 □, △, ○ 등으로 표현하도록 요구하였다. 이후 맞춤형 챗봇의 정답률과 풀이 과정에서의 학년 수준의 적합성이 높아졌으며 단순 계산 오류의 경우 [그림 7]과 같이 수정되어 최종 개발된 맞춤형 챗봇에서는 3회 동안 해당 오류는 발견되지 않았다.



[그림 6] 맞춤형 챗봇 베타 버전의 수정 절차





[그림 7] 맞춤형 챗봇 베타 버전의 수정 후 답변

최종 완성된 맞춤형 챗봇은 [그림 8]과 같으며, 학생들이 대화 스타터에 제시된 문구인 ‘문제해결 단계별로 설명해주세요’, ‘문제를 입력해 주세요.’ 등을 선택해도 작동되며, 이미지 형태나 텍스트 형태로 문제를 입력해도 문제 풀이를 제공하도록 구성하였다.



[그림 8] 완성된 맞춤형 챗봇의 시작 화면

#### IV. 문제해결 맞춤형 챗봇의 수학적 성능

##### 1. 맞춤형 챗봇의 전반적인 수학적 성능

맞춤형 챗봇 최종 버전의 수학적 성능을 검증하기 위해 맞춤형 챗봇 베타 버전을 검증할 때 사용한 [표

2]의 문제를 사용하였다. 완성된 맞춤형 챗봇에 총 3회 동일한 문제를 입력한 후 맞춤형 챗봇이 제시한 답변에 대해 정답률과 풀이 과정을 분석하였다. 맞춤형 챗봇의 오류 유형도 베타 버전과 동일하게 [표 3]을 기준으로 하였다. 맞춤형 챗봇 최종 버전이 제시한 1회~3회까지의 정답 분석은 베타 버전에서 채점자 간 일치도가 높아 초등수학교육 박사 학위 소지자 1인이 단독으로 하였으며, 1회~3회까지의 풀이 과정은 베타 버전 분석에 참여하지 않은 초등수학교육 박사 학위 소지자 1인과 함께 논의를 통해 분류하였다.

맞춤형 챗봇의 수학적 성능을 총 3회 평가한 결과, 1회에서는 총 65문제 중 56.5문제(86.9%), 2회에서는 57문제(87.7%), 3회에서는 60문제(92.3%)로 평균 약 57.8문제를 맞추었으며 약 89.0%의 평균 정답률을 보였다([표 6] 참조). 완성된 맞춤형 챗봇의 정답률은 수정 전인 베타 버전보다 약 4.4%p 향상된 것으로 나타났다. 완성된 맞춤형 챗봇의 3회 동안 정답률은 유사하였으나 회차가 지날수록 정답률이 약간 상승하는 경향을 보였다. 또한 회차가 지날수록 베타 버전에서 추가한 지침을 더욱 잘 반영하는 양상을 보였다. 이를 통해 본 연구에서 제작된 맞춤형 챗봇이 대체적으로 일관성 있게 답하고 있음을 알 수 있다.

[표 6] 맞춤형 챗봇의 정답률 (N=65)

	문제 수			
	1회	2회	3회	평균
정답	56.5 (86.9%)	57 (87.7%)	60 (92.3%)	57.8 (89.0%)
오답	8.5 (13.1%)	8 (12.3%)	5 (7.7%)	7.2 (11.0%)

맞춤형 챗봇에서 정답을 보인 답변을 살펴보면, [그림 9]와 같이 대부분의 문제를 정확하게 인식하였으며 문제에 제시된 표의 각 항에 있는 항목 및 수를 정확하게 인식하였다.

구체적으로 ‘부자 은행과 미래 은행에 같은 기간 동안 예금한 돈과 받은 이자를 나타낸 표입니다. 예금한 돈에 대한 이자의 비율을 비교하여 어느 은행에 예금하는 것이 더 좋은지 쓰고 그 이유를 이야기해 보세요’라는 비와 비율에 관련된 문제를 이미지 형태로 제공하였을 때, 맞춤형 챗봇은 이를 텍스트로 변환하였다. 이때 문제뿐만 아니라 표에 제시된 항목과 단위까지

정확하게 인식하였다. 또한 문제에서 중요하지 않은 그림에 대한 정보는 인식하지 않아 이를 통해 인공지능의 성능을 가늠해 볼 수 있다. 문제에서는 어느 은행에 예금하는 것이 좋을지 쓰고 그 이유를 답하는 것까지 두 가지 사항을 요구하였으며, 맞춤형 챗봇은 미래 은행이라고 답하고, 그 이유까지 정확하게 제시하였다.

<챗봇에 입력된 문제>

문제 배경 | 유사성도

부자 은행과 미래 은행에 같은 기간 동안 예금한 돈과 받은 이자를 나타낸 표입니다. 예금한 돈에 대한 이자의 비율을 비교하여 어느 은행에 예금하는 것이 좋을지 쓰고, 그 이유를 이야기해 보세요.

은행	부자	미래
예금한 돈(원)	80000	50000
이자(원)	2000	1500

( )은행

이유: \_\_\_\_\_

<맞춤형 챗봇의 답변>

문제: "부자 은행과 미래 은행에 같은 기간 동안 예금한 돈과 받은 이자를 나타낸 표입니다. 예금한 돈에 대한 이자의 비율을 비교하여 어느 은행에 예금하는 것이 좋을지 쓰고, 그 이유를 이야기해 보세요."

은행	부자	미래
예금한 돈(원)	80000	50000
이자(원)	2000	1500

-중략-

비율 구하기

- 부자 은행:  $\frac{2000}{80000} = 0.025$
- 미래 은행:  $\frac{1500}{50000} = 0.03$

비교

- 부자 은행의 이자 비율: 0.025
- 미래 은행의 이자 비율: 0.03

답: 미래 은행

이유: 미래 은행의 이자 비율이 부자 은행보다 높아서 같은 돈을 예금했을 때 더 많은 이자를 받을 수 있습니다.

[그림 9] 문제를 옳게 해결한 예 (김성여 외, 2023, p.99)

반면, 오답을 보인 답변을 살펴보면 [그림 10]과 같이 비율의 개념을 정확히 인식하지 못하여 A에 대한 B의 비율에서 어떠한 양이 기준량인지를 혼동하는 특징을 보였다. 구체적으로 이 문제는 ‘지후네 가족은 넓이가 25㎡인 밭에 상추와 딸기를 심었습니다. 상추밭과 딸기밭에 대해 잘못 설명한 것을 찾아 △표 하고, 바르게 고쳐 보세요.’로 첫 번째 설명에 대해 상추밭 넓이에 대한 전체 밭 넓이의 백분율을 (상추밭 넓이 / 전체 밭 넓이) × 100’으로 옳지 않게 제시하고 틀린 설명이 없다고 답변하였다. 반면 아래의 전체 밭 넓이

에 대한 딸기밭 넓이의 비율은 ‘딸기밭 넓이 / 전체 밭 넓이’로 옳게 제시하고 있다. 이처럼 맞춤형 챗봇은 일부 개념에 대한 오류를 보였다.

<챗봇에 입력된 문제>

지후네 가족은 넓이가 25㎡인 밭에 상추와 딸기를 심었습니다. 상추밭과 딸기밭에 대해 잘못 설명한 것을 찾아 △표 하고, 바르게 고쳐 보세요.

상추밭 넓이 15㎡	딸기밭 넓이 10㎡
------------	------------

상추밭 넓이에 대한 전체 밭 넓이의 백분율은 60%입니다.	딸기밭 넓이에 대한 상추밭 넓이의 비는 15 : 10입니다.	전체 밭 넓이에 대한 딸기밭 넓이의 비율은 0.4입니다.
----------------------------------	-----------------------------------	---------------------------------

( ) ( ) ( )

바르게 고치기

<맞춤형 챗봇의 답변>

문제: "지후네 가족은 넓이가 25㎡인 밭에 상추와 딸기를 심었습니다. 상추밭과 딸기밭에 대해 잘못 설명한 것을 찾아서 △표 하고, 바르게 고쳐 보세요."

-중략-

설명이 맞는지 계산하기:

- 상추밭 넓이에 대한 전체 밭의 넓이 백분율:
  - 백분율 = (상추밭 넓이 / 전체 밭 넓이) × 100
  - 백분율 = (15㎡ / 25㎡) × 100 = 60%  (맞음)
- 딸기밭 넓이에 대한 상추밭 넓이 비율:
  - 비율 = 상추밭 넓이 / 딸기밭 넓이
  - 비율 = 15㎡ / 10㎡ = 1.5배 (15:10은 1.5:1과 동일하므로 맞음)
- 전체 밭 넓이에 대한 딸기밭 넓이 비율:
  - 비율 = 딸기밭 넓이 / 전체 밭 넓이
  - 비율 = 10㎡ / 25㎡ = 0.4 (0.4배)

잘못된 설명이 없습니다.

바르게 고치기:

- 모든 설명이 올바르게 수정할 필요가 없습니다.

[그림 10] 문제를 옳지 않게 해결한 예 (안병곤 외, 2023, p.99)

## 2. 학습 주제 및 문제 형태에 따른 맞춤형 챗봇의 수학적 성능

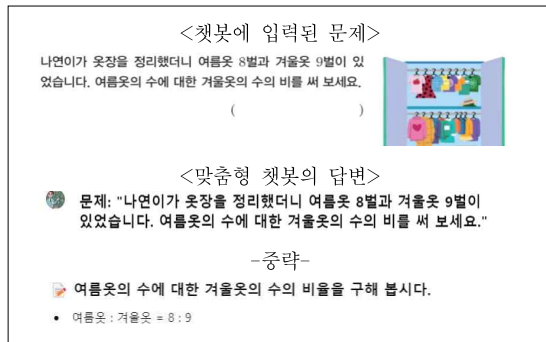
가. 학습 주제에 따른 맞춤형 챗봇의 수학적 성능  
 학습 주제에 따른 정답률을 살펴보면, 비와 비율 문제는 총 38개로 3회 평균 약 32.8문제를 맞췄으며 약

86.4%의 정답률을 보였다([표 7] 참조). 이는 전체 문제의 평균 정답률보다 약 2.63%p 낮은 것으로 나타나 비와 비율 문제에 대한 맞춤형 챗봇의 수학적 성능이 전체 문제보다 약간 떨어짐을 알 수 있다.

비와 비율 문제의 오답을 보면 대부분 비와 비율을 혼동하는 경향을 보였다([그림 11] 참조). 문제 자체에서 ‘비’를 제시하라고 요구하였으나 문제를 인식할 때 ‘비율’로 인식하거나, ‘비’로 인식하였다더라도 답은 ‘비율’로 제시하는 경우가 대부분이었다. 반대로 문제에서 ‘비율’을 제시하라고 하였을 경우 오답은 나타나지 않았다. 이 부분은 맞춤형 챗봇 베타 버전에서 수정하였으나 LLM을 기반으로 하기에 단번에 완벽한 수정은 어려운 것으로 보이며 추가적인 프롬프트 엔지니어링을 통해 지속적인 수정이 필요하다.

[표 7] 학습 주제에 따른 정답률

	문제 수		
	정답	오답	합계
비와 비율 (N=38)	32.8 (86.4%)	5.2 (13.6%)	38 (100%)
비례식과 비례배분 (N=27)	25.0 (92.6%)	2.0 (7.4%)	27 (100%)



[그림 11] 비와 비율을 혼용하여 사용한 예 (신항균 외, 2023, p. 90)

비례식과 비례배분 문제는 총 27개로 3회 평균 25 문제를 맞추어 92.6%의 정답률을 보였다. 이를 통해 맞춤형 챗봇은 비와 비율보다 비례식과 비례배분 문제를 더 잘 해결함을 알 수 있다. 비례식과 비례배분에서는 문제를 해결하는 과정이 아닌 대부분 문제의 형태가 복잡할 경우 문제를 인식하지 못하는 오류를 보

였다. 예를 들어, 문제 안에 세부 문항이 두 가지 이상 존재하거나 말풍선과 함께 그림을 인식해야 하는 경우에 문제 인식 오류가 많이 발생하였다.

이처럼 해당 주제에 따라 맞춤형 챗봇의 수학적 성능이 다르게 나타났으며, 각 주제별로 맞춤형 챗봇이 보인 오류 유형에도 차이가 나타났다. 맞춤형 챗봇의 수학적 성능을 높이기 위해서는 왜 이러한 차이를 보이는지 구체적으로 살펴볼 필요가 있다.

나. 문제 형태에 따른 맞춤형 챗봇의 정답률

GPTs의 경우 멀티모달이 가능하여 다양한 형태의 문제를 대체로 잘 인식하지만, 문제 구성에 따라 차이를 보일 수 있다. 이에 텍스트만으로 해결이 가능한 문제와 그림(표, 말주머니 등 제외)이 핵심적인 역할을 하는 문제를 구분하여 정답률을 살펴보았다.

우선 텍스트만으로 구성된 문제 경우 정답률은 [표 8]과 같다. 다만, 본 논문의 경우 모든 문제를 이미지 형태로 맞춤형 챗봇에 입력하였기 때문에 텍스트로 구성된 문제 또한 이미지 형태로 입력되었다. 따라서 맞춤형 챗봇에 직접 텍스트 형태로 제공한 것보다 문제에 대한 정답률이 낮게 나타날 수 있다.

[표 8] 텍스트로만 구성된 문제의 정답률 (N=37)

	문제 수		
	정답	오답	합계
텍스트만으로 구성된 문제	33.2 (89.6%)	3.8 (10.4%)	37 (100%)

텍스트만으로 구성된 문제의 경우 89.6%의 정답률을 보였다. 텍스트로만 구성된 문제는 강윤지(2024)에서 사용했던 문제와 동일하며, 해당 연구에서 범용적인 챗봇인 ChatGPT에게 문제를 풀게 했을 때 정답률은 81.08%였다. 이를 고려해 볼 때 맞춤형 챗봇의 수학적 성능이 좀 더 우수한 것을 알 수 있다. 만약 본 연구에서도 문제를 이미지 형태가 아닌 텍스트 형태로 맞춤형 챗봇에 입력했다면 정답률이 더 큰 차이가 있었을 것으로 보인다.

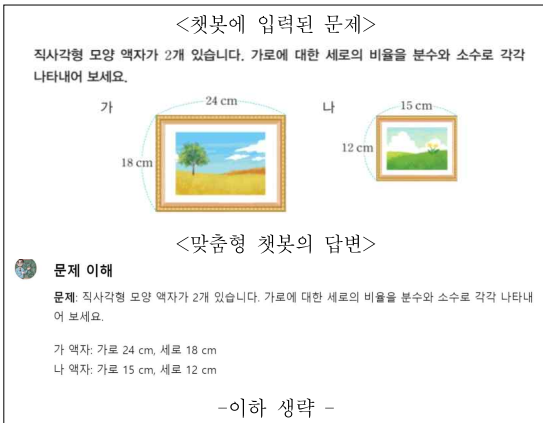
다음으로 문제해결 시 그림(표, 말주머니 등 제외)이 핵심적인 역할을 하는 경우와 그렇지 않은 경우를 구분하여 살펴보았다. 맞춤형 챗봇의 경우 표와 말주머니를 잘 인식하였으며 특히 표를 매우 잘 인식하여

표에 제시된 정보를 대부분 옳게 파악하였다. 따라서 표와 말주머니 등을 제외한 그림이 문제를 해결하는데 핵심적인 역할을 하는 문제들만 추출하여 정답률을 살펴보았다([표 9] 참조). 그 결과 그림이 문제를 해결하는 데 핵심적인 역할을 하지 않는 문제의 경우 91%의 높은 정답률을 보였다. 반면 그림이 문제를 해결하는 데 핵심적인 역할을 하는 문제의 경우 이보다 낮은 82.0%의 낮은 정답률을 보였다.

[표 9] 문제에서 그림의 역할에 따른 정답률

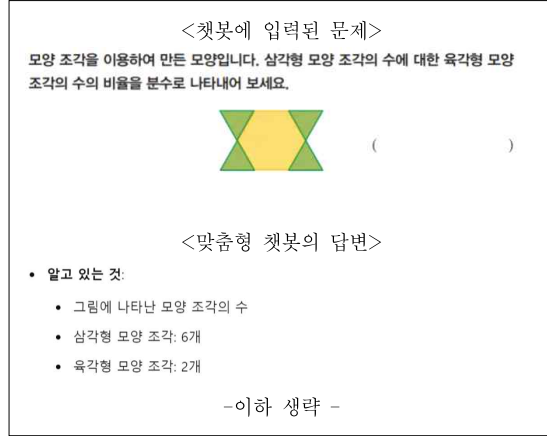
	문제 수		
	정답	오답	합계
그림이 핵심적이지 않은 문제 (N=50)	45.5 (91.0%)	4.5 (9.0%)	50 (100%)
그림이 핵심적인 문제 (N=15)	12.3 (82.0%)	2.7 (18.0%)	15 (100%)

구체적으로 맞춤형 챗봇은 문제에 제시된 그림에서 가로와 세로를 매우 잘 인식하는 것으로 나타났다(예, [그림 12] 참조). 문제에서 제시한 액자 그림에 가로 세로가 표시되어 있지 않음에도 완벽히 인식하였다.



[그림 12] 가로와 세로를 잘 이해한 예시 (안병곤 외, 2023, p. 99)

반면 일부 그림은 인식하는 데 지속적인 오류를 보였다(예, [그림 13] 참조). 특히 해당하는 물건 또는 도형이 몇 개인지 텍스트로 제공되어 있지 않은 경우, 물건 또는 도형의 수를 베타 버전 포함 총 4회 검증 가운데 2~3회 다르게 답하였다.



[그림 13] 그림을 잘못 인식한 예 (박성선 외, 2023, p. 96)

### 3. 맞춤형 챗봇의 오류 유형

맞춤형 챗봇이 초등학교 수학 문제를 해결하는 과정에서 나타난 오류 유형을 분석한 결과, 문제 인식 오류, 문제 이해 오류, 개념 혼동 오류, 단순 계산 오류, 출력 오류 등이 나타났다([표 10] 참조).

구체적으로 오류를 유형별로 분석한 결과, 문제 인식 오류는 5.9%를 차지하였으며, 문제 이해 오류는 약 15%, 개념 혼동 오류는 3.3%, 출력 오류는 0.5%로 나타났다. 베타 버전에서 나타났던 단순 계산 오류는 3회 동안 나타나지 않았다. 이러한 오류 유형은 맞춤형 챗봇이 문제를 인식하는 데 가장 큰 어려움을 겪고 있음을 시사하며, 텍스트가 아닌 시각적인 표현에서는 아직 능력의 한계가 있음을 뜻한다. 또한 이는 문제를 이미지가 아닌 텍스트로 입력한다면 현재 평균 정답률인 89.0%보다 더 높은 정답률을 보일 수 있다는 것을 의미하기도 한다.

실제 이러한 문제 인식 오류는 문제를 그림이 아닌 텍스트로 제공하거나 문제를 인식한 부분에 대해 대화로 맞춤형 챗봇에게 틀렸으므로 수정해 달라고 요청했을 경우 대부분 해결되었다. 하지만 본 논문에서는 교과서 문제 자체를 입력했을 때의 수학적 성능을 살펴보고 있어, 텍스트를 추가 입력하거나 문제 인식을 잘못했을 경우에 맞춤형 챗봇에게 요청한 내용에 대한 분석 결과는 구체적으로 제시하지 않았다.

예를 들어, [그림 14]의 그림의 경우 3회 동안 지속

[표 10] 맞춤형 챗봇의 답과 풀이 과정의 유형 분석 결과

(N=65)

분석 기준		문항 수				
		1차	2차	3차	평균(%)	
정답	풀이 과정이 옳은 경우	학년 수준에 적합한 풀이	49.5	49	56	51.5 (79.2%)
		학년 수준에 적합하지 않은 풀이	8	8	4	6.7 (10.3%)
	풀이 과정이 옳지 않은 경우	문제 이해 오류	0	1	2	1 (1.5%)
		개념 혼동 오류	0	0	0	0 (0%)
오답	풀이 과정이 옳은 경우	학년 수준에 적합한 풀이	8.5	8	5	7.2 (11%)
		학년 수준에 적합하지 않은 풀이	0	0	0	0 (0%)
	풀이 과정이 옳지 않은 경우	문제 인식 오류	6	3.5	2	3.8 (5.9%)
		문제 이해 오류	0	2	1	1 (1.5%)
		개념 혼동 오류	2	2.5	2	2.2 (3.3%)
		단순 계산 오류	0	0	0	0 (0%)
		출력 오류	1	0	0	0.3 (0.5%)

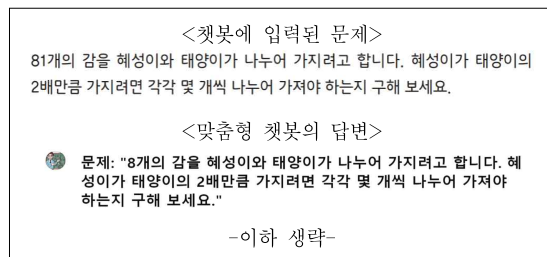
해서 승용차는 5대, 버스는 2대로 인식하여 오답을 제시하였다. 이에 맞춤형 챗봇에 승용차의 수가 틀렸으니 다시 확인해 보도록 요청하거나 그림의 내용을 텍스트로 제공했을 때 옳은 답을 제시하였다. 이러한 이미지 인식 오류에 대해서는 생성형 AI 자체의 성능이 좋아져야 수정이 가능할 것으로 보이나 본 연구에서는 학생들과 자연어로 대화가 가능한 맞춤형 챗봇을 개발했으므로 맞춤형 챗봇이 문제를 잘못 인식했을 경우 학생들이 직접 챗봇과의 대화를 통해 해당 부분을 수정하여 무리 없이 사용할 수 있으리라 판단된다.



[그림 14] 문제 인식 오류 중 그림 인식 오류의 예 (박성선 외, 2023, p. 96)

또한 일부 문제의 경우 [그림 15]와 같이 수와 문자 사이의 간격이 좁은 경우 지속적으로 수를 인식하지 못하는 오류를 보였으며, 이를 텍스트로 형태로 제공

했을 때는 옳게 해결하였다. 이러한 것을 통해 문제 안에 수와 문자의 간격 또는 글씨체에 따라 인공지능이 인식하지 못할 수 있으며, 앞으로 인공지능을 활용하기 위해서는 이러한 부분까지 점검이 필요해 보인다.



[그림 15] 문제 인식 오류 중 수 인식 오류의 예 (한대희 외, 2023, p. 99)

다음으로 많이 발생한 오류는 개념 혼동 오류이다. 맞춤형 챗봇의 경우 비와 비율을 같은 개념으로 혼동하거나 대부분 비를 비율로 제시하는 경향을 보였다. 이는 각 용어의 의미를 정확히 파악하지 못하는 것으로 판단된다. 또한 A에 대한 B의 비를 해석할 때 지속적으로 오류를 보였다. [그림 11]과 같이 여름옷의 수에 대한 겨울옷의 수를 비로 써 보라는 문제에서 '여름옷 : 겨울옷 = 8 : 9'로 제시하거나 [그림 10]과 같이 '상추밭 넓이에 대한 전체 밭의 백분율'의 경우 '상추밭 넓이'가 기준량이므로  $(25 \div 15) \times 100$ 을 제시해야 하나

반대로  $(15 \div 25) \times 100$ 을 제시하였다. 이처럼 특히 비교하는 양이 기준량보다 큰 경우에는 비교하는 양과 기준량을 바꾸어 제시하는 경향을 보였다. 이는 일반적으로 1 이상의 비율보다는 1 이하의 비율이 확률적으로 더 많이 다루어지기 때문에 LLM 모델에서 확률을 기반으로 바꾸어 제시한 것으로 유추해 볼 수 있다.

세 번째로 문제 이해 오류는 문제에서 구하고자 하는 것을 적절하게 파악하지 못하고 오답을 보인 경우이다. [그림 16]과 같이 문제에서 물건의 정가와 판매가를 제공한 후, 할인하는 비율이 높은 물건부터 차례로 쓰도록 요구하였으나 맞춤형 챗봇에서 구하고자 하는 것을 ‘정가에 대한 판매가의 비율을 구하여 물건의 비율이 높은 순서대로 나열하기’로 인식하여 오답을 제시하였다. 반면 다른 회차에서 구하고자 하는 것을 ‘주스, 과자, 아이스크림의 할인율을 계산하여 높은 순서대로 나열’하는 것으로 인식했을 경우에는 옳은 답을 제시하였다. 이처럼 맞춤형 챗봇의 경우에도 문제에서 구하고자 하는 것을 제대로 파악하지 못하여 오류를 범하는 경우가 있으며, 이는 단순 계산 오류보다 더 많이 발견되었다.

마지막으로 출력 오류가 1회 나타났으며 [그림 17]과 같이 ‘유산균 음료를 만드는 데 우유 1000mL와 요구르트 150mL가 필요합니다. 유산균 음료를 만드는 데 필요한 우유 양에 대한 요구르트 양의 비율을 구하고, 백분율로 나타내어 보세요.’라는 문제를 해결할 때 비율을 분수와 소수로 구했으며 백분율로 옳게 제시하였다. 하지만 답으로 출력할 때에는 백분율은 15%로 옳게 제시하고 비율의 경우 3:20으로 비를 제시하였다. 이처럼 문제는 적절히 해결하였으나 답을 제시할 때에도 매우 적은 비율이나 오답을 제공하기도 하였다.

오류 유형 이외에 풀이 과정의 경우 학년 수준에 적합하도록 많이 수정되었다. [표 5]와 [표 10]을 비교해보면, 맞춤형 챗봇 베타 버전의 경우 정답과 오답을 포함하여 풀이 과정이 학년 수준에 적합한 경우는 65개 문제 중 51문제(78.5%)인 반면, 맞춤형 챗봇 최종 버전에서는 65개 문제 중 58.7문제(90.2%)로 나타났다. 구체적으로 맞춤형 챗봇 최종 버전에서는 대부분 미지수  $x$ 를 대신하여 □, ○등을 사용하였으며, 일부 문제에서만 미지수를 문자로 제시하였다. 이처럼 완벽하지는 않지만 풀이 과정 또한 학년 수준에 적합하도록 어느 정도 조정된 것으로 보인다.

**<챗봇에 입력된 문제>**

어떤 가게에서 판매하는 각 물건의 정가와 판매가입니다. 물음에 답해 보세요.

물건	주스	과자	아이스크림
정가(원)	2000	2500	1000
판매가(원)	1800	2300	700

☐ 안에 알맞은 수를 써넣으세요.

- 정가가 2000원인 주스를  % 할인하여 1800원에 판매합니다.
- 정가가 2500원인 과자를  % 할인하여 2300원에 판매합니다.
- 정가가 1000원인 아이스크림을  % 할인하여 700원에 판매합니다.

• 주스, 과자, 아이스크림 중 할인하는 비율이 높은 물건부터 차례로 써 보세요.

( )

**<맞춤형 챗봇의 틀린 답변(3회)>**

**🔍 구하고자 하는 것:**

- 정가에 대한 판매가의 비율을 구하여 물건의 비율이 높은 순서대로 나열하기

**📌 알고 있는 것:**

- 주스: 정가 2000원, 판매가 1800원
- 과자: 정가 2500원, 판매가 2300원
- 아이스크림: 정가 1000원, 판매가 700원

-중략-

비율이 높은 순서대로 나열하면:

- 과자: 0.92
- 주스: 0.9
- 아이스크림: 0.7

**📌 답:**

- 주스, 과자, 아이스크림을 할인하는 비율이 높은 물건부터 차례로 쓰면 과자, 주스, 아이스크림입니다.

**<맞춤형 챗봇의 옳은 답변(2회)>**

- 구하고자 하는 것: 주스, 과자, 아이스크림의 할인율을 계산하여 높은 순서대로 나열.
- 문제에서 알 수 있는 것:
  - 주스: 정가 2000원, 판매가 1800원
  - 과자: 정가 2500원, 판매가 2300원
  - 아이스크림: 정가 1000원, 판매가 700원

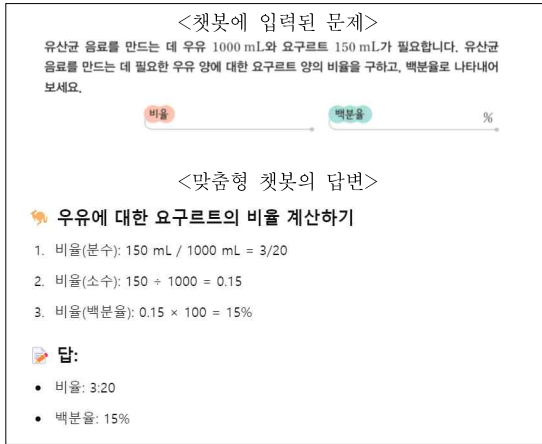
-중략-

**📌 답:**

- 아이스크림 (30%)
- 주스 (10%)
- 과자 (8%)

[그림 16] 문제 이해 오류의 예 (류희찬 외, 2023, p. 109)





[그림 17] 출력 오류의 예 (안병곤 외, 2023, p. 99)

## V. 결론 및 제언

본 논문은 문제해결 역량 신장을 위해 폴리아의 문제해결 단계에 따라 풀이를 제공하는 GPTs 기반 문제해결 맞춤형 챗봇을 개발하여 그 수학적 성능을 분석하는 데 초점을 두었다. 연구 결과를 토대로 도출한 결론 및 제언은 다음과 같다.

첫째, GPTs 기반 문제해결 맞춤형 챗봇의 전체 문제에 대한 정답률은 약 89.0%로 범용적인 챗봇인 ChatGPT보다 높게 나타나 학생들의 학습을 돕는데 더 효과적인 것으로 기대된다. 구체적으로 권오남 외(2023)에서 일반 ChatGPT(3.5 turbo 버전)를 이용하여 국가수준 학업성취도 평가 수학 문제를 풀어보게 한 결과 정답률은 37.1%, 대학수학능력시험 수학 문제를 풀어보게 한 결과 정답률은 15.97%로 나타났다. 이는 중·고등학생용 문제에 대한 수학적 성능을 평가했기 때문에 난도가 높아 맞춤형 챗봇보다 정답률이 더 낮게 나타날 수 있다. 하지만 본 논문에서 개발한 맞춤형 챗봇과는 정답률에서 최대 73%p로 많은 차이가 나는 것을 알 수 있다. 또한 강윤지(2024)에서 본 논문과 동일한 교과서 단원의 평가 문제를 활용하여 ChatGPT의 문제해결 능력을 평가한 결과 정답률은 81.08%로 나타났으며, 본 논문에서 제작한 맞춤형 챗봇의 경우 정답률은 89.6%([표 8] 참조)로 8.52%p 더 높은 정답률을 보였다. 또한 본 논문에서는 선행 연구에서 텍스트 형태로 문제를 직접 ChatGPT에 입력한

방식과 달리, 이미지 형태로 그림, 표, 말주머니 등을 포함하여 제공하였기 때문에 맞춤형 챗봇의 활용 범위와 수학적 성능 또한 높아졌음을 알 수 있다. 이러한 챗봇의 수학적 성능의 향상은 GPT-4o 버전에서 검색증강 생성(RAG)과 적절한 프롬프트 엔지니어링을 통해 챗봇을 개발하였기 때문으로 보인다.

둘째, 범용적인 챗봇인 ChatGPT와 맞춤형 챗봇이 제시한 오답 유형에도 차이가 있는 것으로 나타났다. ChatGPT를 이용하여 본 논문과 동일한 교과서 단원의 문제로 실험을 실시한 강윤지(2024)의 경우, 문제 이해 오류가 6.84%, 식 세우기 오류가 9.5%, 계산 오류가 4.74%인 반면, 본 논문에서는 문제 인식 오류가 5.9%, 문제 이해 오류가 1.5%, 개념 혼동 오류가 3.3%, 출력 오류가 0.5%로 나타났다. 이를 통해 범용적인 챗봇인 ChatGPT의 경우 맞춤형 챗봇보다 식 세우기 오류와 계산 오류를 더욱 많이 범하는 것을 알 수 있다. 반면 맞춤형 챗봇 최종 버전에서는 식 세우기 오류와 계산 오류가 나타나지 않았다. 이는 맞춤형 챗봇을 처음 제작할 때 프롬프트 엔지니어링 기법 중 생각의 사슬(CoT)을 사용하여 생성형 AI가 식 세우기 오류 없이 문제를 더욱 잘 해결한 것으로 보이며, 베타 버전에서 계산 오류에 대한 지침을 제공하여 문제를 해결할 때 계산 오류를 범하지 않은 것으로 판단된다. 이처럼 맞춤형 챗봇의 경우 제공된 지식과 프롬프트 엔지니어링으로 범용적인 챗봇인 ChatGPT보다 좀 더 미세하게 오류를 조정할 수 있기 때문에 수학교육에 더 효과적으로 사용이 가능할 것이다. 한편, 맞춤형 챗봇의 경우 범용적인 챗봇에서 나타나지 않은 문제 인식 오류가 가장 많이 발생했다. 맞춤형 챗봇의 경우 모든 문제를 이미지 형태로 제공하였기 때문으로 보이며, 학생들이 맞춤형 챗봇을 이용할 때 이미지 형태를 가장 많이 활용할 수 있으므로 그림을 단순화하거나 수와 단위 사이에 적절한 간격을 주는 등 이미지를 잘 인식할 수 있도록 교과서를 제작할 때 주의를 기울인다면 좀 더 효과적으로 학생들이 맞춤형 챗봇을 활용할 수 있을 것으로 생각한다.

셋째, 맞춤형 챗봇의 경우 비와 비율을 혼동하여 사용하는 경향을 보였으며 이러한 현상은 범용적인 챗봇인 ChatGPT에서도 동일하게 발생하였다. 구체적으로 맞춤형 챗봇에서 처음 문제를 인식할 때, 비를 비율로 인식하여 비율로 답을 하거나 비율로 인식하였으나 답

은 비로 하는 경우가 있었다. 또한 처음에 인식은 비로 하였으나 답은 비율로 제시하는 등 대부분의 오류에서 비를 비율로 제시하는 경향을 보였다. 또한 A에 대한 B의 비를 해석할 때도 오류를 보였으며 특히 비교하는 양이 기준량보다 큰 경우에는 비교하는 양과 기준량을 바꾸어 제시하는 오류 양상을 보였다. 이처럼 LLM을 기반으로 하는 생성형 AI는 일반적인 상황은 잘 처리하나, 각 용어의 의미를 정확히 알고 판단하기보다는 확실적인 답변을 하기 때문에 미세한 부분에서 오류를 보였다. 이에 맞춤형 챗봇을 개발할 때에는 이러한 부분에 주의하여 전형적이지 않은 상황에 대해서는 미리 프롬프트 엔지니어링을 통해 조정하여야 할 것이다.

넷째, 맞춤형 챗봇은 이미지 인식 성능이 매우 좋은 것으로 나타났다. 표의 경우 100% 완벽하게 인식하였으며 그림의 경우에도 텍스트로 된 표현 없이 가로 길이, 세로 길이, 키, 그림자의 길이 등을 정확히 인식하였다. 또한 문제해결과 관련이 없는 그림이 포함된 문제에서는 해당 그림은 인식하지 않았다. 이처럼 맞춤형 챗봇은 그림이 포함된 문제를 제공했을 때 그림에서 문제를 해결하기 위한 정보만을 추출할 수 있는 정도의 수준 높은 성능을 보였다. 초등학교 수학교과서에서는 텍스트로만 제시된 문제보다는 그림 또는 표가 함께 있는 문제들이 많이 제시된다. 따라서 이러한 맞춤형 챗봇의 우수한 이미지 인식 성능을 고려해볼 때, 초등학생들을 위한 맞춤형 챗봇을 활용할 수 있는 가능성을 엿볼 수 있다. 다만 대부분의 그림은 잘 인식하였으나 일부 그림의 경우 지속해서 인식하지 못하는 경향을 보였다. 물론 챗봇이기 때문에 문제 인식의 오류가 있을 경우 이를 대화를 통해 바로 수정할 수 있지만 학생들에게는 불편한 작업일 수 있다. 특히 맞춤형 챗봇이 인식하지 못한 그림 중 일부는 모호하게 그려져 있어 직접 학생들이 문제를 해결할 때도 혼란을 줄 수 있는 경우도 있었다. 이에 교과서 발행 시 인식하기 모호한 그림이 있는지 맞춤형 챗봇을 통해 확인하여 수정도 가능하리라 생각한다.

다섯째, 맞춤형 챗봇의 수학적 성능은 학습 주제에 따라 다르게 나타났다. 연구 결과, 비와 비율에 관한 문제의 정답률은 84.6%, 비례식과 비례배분에 관한 문제의 정답률은 92.6%로 나타났다. 비와 비율에 관한 문제에서는 대부분 비와 비율에 대한 개념을 혼동하는

오류가 많이 나타났으나 비례식과 비례배분에서는 이러한 개념 혼동 오류가 적어 정답률이 더 높게 제시되었다. 이처럼 맞춤형 챗봇의 정답률은 문제가 다루고 있는 주제에 따라 다르게 나타날 수 있으며, 그러한 이유를 파악하여 프롬프트를 조정한다면 학생들에게 좀 더 효과적인 맞춤형 챗봇을 개발할 수 있을 것이다.

마지막으로 맞춤형 챗봇의 경우에도 문제 이해 단계가 매우 중요한 과정임을 알 수 있다. 본 연구에서는 맞춤형 챗봇에게 답변의 처음 부분에 ‘문제에서 구하고자 하는 것’을 제시하도록 개발하였다. 그 결과 동일한 문제일 경우에도 문제에서 구하고자 하는 것을 정확히 파악했을 때는 정확한 풀이 과정과 정답을 제시한 반면 문제에서 구하고자 하는 것을 파악하지 못했을 때는 전혀 다른 내용의 풀이 과정과 오답을 제시하였다([그림 16] 참조). 실제 학생들 역시 문장제를 해결할 때 식을 세우거나 계산을 하지 못해 어려움을 겪는 것보다 문제를 이해하지 못해 어려움을 겪는 경우가 더 많은 것으로 나타났다(Gallagher et al., 2021). 이처럼 문제를 해결하기 위해서는 문제 이해가 매우 중요한 단계임을 다시 한번 생각해 볼 수 있다.

본 연구는 문제해결 역량의 신장을 위한 문제해결 맞춤형 챗봇을 개발하고 수학 문제를 단계별로 얼마나 잘 해결할 수 있는지 그 수학적 성능을 평가하고자 하였다. 연구 결과, GPTs 기반 문제해결 맞춤형 챗봇은 전체적으로 높은 정답률을 나타내어 초등 수학 문제해결에 유용한 도구로 활용될 수 있는 가능성을 엿볼 수 있었다. 다만 생성형 AI를 이용한 맞춤형 챗봇이기 때문에 100%의 정답을 제공한다고 보장할 수는 없다. 하지만 베타 버전의 제작 후 1회 수정을 통해 약 89.0%의 정답률을 보였으며, 대부분의 오류는 이미지를 사용하였기 때문에 생긴 것들이었다. 이미지 인식 기술 역시 계속 발전하고 있기 때문에 정답률은 더 높아질 것으로 예상된다. 또한 맞춤형 챗봇의 프롬프트를 지속적으로 수정한다면 수학적 오류가 좀 더 적은 정확도 높은 챗봇을 개발할 수 있을 것으로 생각한다. 이 연구가 다양한 수학교육용 맞춤형 챗봇을 개발하여 사용하는 시발점이 되어 학생들의 학습뿐만 아니라 교사 교육 등 다양한 분야에서 활용되길 기대한다.



## 참 고 문 헌

- 강윤지(2024). 수학 문장제 해결과 관련한 ChatGPT의 교수학적 활용 방안 모색. 수학교육 논문집, 38(1), 49-67.
- 교육부(2021). 수학 6-1. 비상교육.
- 교육부(2022). 수학과 교육과정. 교육부 고시 제 2022-33호 [별책 8]. 교육부.
- 권오남, 오세준, 윤정은, 이경원, 신병철(2023). ChatGPT의 수학적 성능 분석: 국가수준 학업성취도 평가 및 대학수학능력시험 수학 문제 풀이를 중심으로. 수학교육 논문집, 37(2), 233-256.
- 김귀훈, 고우영, 윤혜진, 이제현, 임준호, 류운하, 송석리, 이규호, 이도영, 한준구(2023). 생성형 AI 교육 자료: ChatGPT 사례 중심으로. 서울특별시교육청.
- 김덕진(2023). AI 2024 트렌드 & 활용백과. 스마트북스.
- 김덕진, 서승완(2024). 챗봇 2025 트렌드 & 활용백과. 스마트북스.
- 김성여, 강언진, 강요한, 고창수, 김보현, 김아롱, 노시현, 박소연, 박용준, 박희정, 송정아, 안효은, 이대현, 정선혜, 정유화, 황윤정(2023). 수학 6-1. 아이스크립 미디어.
- 김진중(2024). 최고의 프롬프트 엔지니어링 강의. 리코멘드.
- 류희찬, 유현주, 이종영, 조영미, 탁병주, 최인숙, 정미진, 이환규, 전종호, 김광식, 이경선, 박순덕, 김경희, 원선희, 선종희, 김백균, 이영배, 김종욱, 김해동, 장연지, 나미연, 김은미, 박순희, 서우림, 윤현철, 김유리(2023). 수학 6-1. 금성출판사.
- 박성선, 류성림, 김상미, 권성룡, 김남균, 강호진, 김경탁, 김보경, 김영진, 김용성, 김용관, 성장근, 오해진, 이명희, 정인수, 최병훈, 최주영(2023). 수학 6-1. YBM.
- 선주오(2024). 대화형 수학 튜터링 시스템을 위한 데이터셋 설계 및 구축. 건국대학교 대학원 석사학위논문.
- 손태권(2024). 수학교육에서의 인공지능: 지능형 교수시스템에 대한 체계적 문헌 검토. 수학교육학연구, 34(2), 187-208.
- 서지영(2024). 랭체인으로 LLM기반의 AI 서비스 개발하기. 길벗.
- 신항균, 김태환, 김리나, 정나영, 최혜령, 황혜진, 서형주, 김숙현, 김현미, 남미선, 박재한, 조웅래, 최하철, 김승희, 문행운, 홍혜원, 임연주(2023). 수학 6-1. 비상교육.
- 안병곤, 나귀수, 김민경, 이광호, 류현아, 최지선, 조진우, 김남준, 정연숙, 박은정, 허지연, 황창훈, 문은혜, 손태권, 추성엽, 장송이, 정소영, 김주창, 윤재원, 정은희, 차은경(2023). 수학 6-1. 두산동아.
- 유호석(2024). 제가 만든 GPT는 당신이 만든 GPT와 전혀 다릅니다. 리코멘드.
- 이규남, 조우진, 김동민(2023). 프롬프트 엔지니어링으로 인공지능 제대로 일 시키기. 제이펍.
- 이지혜, 허난(2020). 인공지능을 활용한 맞춤형 수학 학습 프로그램 개발. East Asian Mathematical Journal, 36(2), 273-289.
- 한대희, 고은성, 조형미, 한상희, 이희석, 신희영, 이자미, 문석현, 서현진, 김나라, 김가영, 김태범, 임훈택, 박이서로, 황은지, 오민영(2023). 수학 6-2. 천재교과서.
- 허민(2023). 제가 쓰는 챗GPT는 당신이 쓰는 챗GPT와 전혀 다릅니다. 리코멘드.
- Afzaal, M., Nouri, J., Zia, A., Papapetrou, P., Fors, U., Wu, Y., Li, X., & Weegar, R. (2021). Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation. *Frontiers in Artificial Intelligence*, 4, 1-20.
- Almeida, D. R. (2024, April 10). CLIP embeddings to improve multimodal RAG with GPT-4 Vision. *OpenAI Cookbook*. [https://cookbook.openai.com/examples/custom\\_image\\_embedding\\_search](https://cookbook.openai.com/examples/custom_image_embedding_search)
- Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J., Clowez, G., Boileau, P., & Ruetsch-Chelli, C., (2024). Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26, e53164. Retrieved from <https://www.jmir.org/2024/1/e53164>.
- Doc, N. V., Nam, N. T. H., Thanh, N. T., & Giam,

- N. M. (2023). Teaching mathematics with the assistance of an AI chatbot to enhance mathematical thinking skills for high school students. *International Journal of Current Science Research and Review*, 6(12), 8574-8580.
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching & learning*. 정제영, 이선복 역(2020). 인공지능 시대의 미래교육: 가르침과 배움의 함의. 박영story.
- Gallagher, M. A., Ellis, L., & Weiland, T. (2021). Making word problems meaningful. *Mathematics Teaching: Learning & Teaching PK-12*, 114(8), 580-590.
- Yeh, C. Y. C., Cheng, H. N. H., Chen, Z. H., Liao, C. C. Y., & Chan, T. W. (2019). Enhancing achievement and interest in mathematics learning through Math-Island. *Research and Practice in Technology Enhanced Learning*, 14(5), 1-19.

## Development and mathematical performance analysis of custom GPTs-Based chatbots

**Kwon, Misun**

Shinpoog Elementary School

E-mail : annitj@naver.com

This study presents the development and performance evaluation of a custom GPT-based chatbot tailored to provide solutions following Polya's problem-solving stages. A beta version of the chatbot was initially deployed to assess its mathematical capabilities, followed by iterative error identification and correction, leading to the final version. The completed chatbot demonstrated an accuracy rate of approximately 89.0%, correctly solving an average of 57.8 out of 65 image-based problems from a 6th-grade elementary mathematics textbook, reflecting a 4 percentage point improvement over the beta version. For a subset of 50 problems, where images were not critical for problem resolution, the chatbot achieved an accuracy rate of approximately 91.0%, solving an average of 45.5 problems correctly. Predominant errors included problem recognition issues, particularly with complex or poorly recognizable images, along with concept confusion and comprehension errors. The custom chatbot exhibited superior mathematical performance compared to the general-purpose ChatGPT. Additionally, its solution process can be adapted to various grade levels, facilitating personalized student instruction. The ease of chatbot creation and customization underscores its potential for diverse applications in mathematics education, such as individualized teacher support and personalized student guidance.

---

\* 2020 Mathematics Subject Classification : 97P80

\* Key Words : GPTs, GPT, customized chatbot, problem-solving, artificial intelligence, conversational tutoring system.