컴퓨터교육학회 논문지 2024년 제27권 제4호
https://doi.org/10.32431/kace.2024.27.4.011

거대언어모델 기반 인공지능의 일반물리학 학습 도구로서의 가능성 탐색: 문제 풀이 능력을 중심으로

Investigating the potential of large language model-based AI as a learning tool in Introductory physics: A focus on problem-solving capabilities

강동열†
Dongyel Kang†

요 약

최신 거대언어모델 기반 인공지능(LLM-AI)인 GPT-4o, ChatGPT4, Gemini Advanced, 그리고 Claude-3 Opus를 대상으로, 실제 대학의 일반물리학 교과목 정규 시험에서 발췌한 45개 문제에 대한 풀이 능력을 측정하였다. LLM-AI를 학습 도구로 사용하는 학습자 관점에서, 단순 문제 입력 방식과 문제에 추가 정보를 포함한 방식으로 각 문제를 3번씩 시도하였다. 그 결과, 모든 LLM-AI 모델들은 학생들의 평균 정답률인 50%를 웃도는 80% 이상의 정답률을 기록하였으며 특히, GPT-4o는 90%를 넘는 성능을 보였다. 또한, LLM-AI의 정답률이 낮은 문제들과 이에 대한 LLM-AI의 오답 유형을 분석하여 LLM-AI의 확률적 추론 방식의 특성과 한계를 논의하였다. 본 연구 결과는 LLM-AI 가 대학 일반물리학 교육에서 지능형 맞춤 학습 도구 등으로 활용될 가능성을 시사한다.

주제어: 인공지능, 일반물리학, 맞춤형 교육, 자기주도학습, 인공지능 교육

ABSTRACT

This study evaluated the problem-solving capabilities of latest Large Language Model based Artificial Intelligence (LLM-AI) models, GPT-4o, GPT-4, Gemini Advanced, and Claude-3 Opus, on 45 problems extracted from actual university-level introductory physics course exams. From the perspective of a learner using LLM-AI as a learning tool, we attempted to solve each problem three times using two input methods: simple problem input and problem input with additional information. All LLM-AI models achieved accuracy rates exceeding 80% with the performance of GPT-4o more than 90%, which surpass the average student accuracy rate of approximately 50%. We also analyzed the problems with low accuracy rates of LLM-AI and their incorrect responses to discuss the characteristics and limitations of probabilistic inference reasoning of LLM-AI. The findings of this study suggest that the currently available LLM-AIs can be utilized as intelligent personalized learning tools in university-level introductory physics education.

Keywords: Artificial intelligence, Introductory physics, Personalized education, Self-directed learning, Artificial intelligence education

1. 서론

최근 거대언어모델(Large Language Model: LLM) 기반의 생성형(generative) 인공지능(Artificial Intelligence; AI)의 발전이 가속화되고 많은 분야에서 이 LLM 기반 AI의 성능을 측정하고 활용하려는 움직임이 커지고 있다[1-3]. Transformer 기법으로 가능해진 대규모 언어 자료에 대한 훈련(pre-training)과 인간에 의한 추가적인 강화학습으로 그 성능이 크게 향상된 LLM 기반 생성형 AI(LLM-AI)는 대화형 챗봇

†일반회원: 국립한밭대학교 노마드칼리지 기초과학부 부교수

논문투고: 2024년 06월 12일, 심사완료: 2024년 06월 28일, 게재확정: 2024년 07월 03일

(chatbot)과 결합하여 마치 대화하는 인격체처럼 사람과 상호작용하는 AI로 진화하였다[2,3]. 2022년 말에 OpenAI는 강력한 자연어 처리를 통해 인간 수준에 근접한 대화를 하는 ChatGPT를 공개하여 세상에 큰 반향을 일으켰으며, 이후 ChatGPT3.5, GhatGPT4, 그리고 최근의 GPT-4o까지 계속해서 성능이 향상된 후속 모델을 출시하고 있다. OpenAI 이외에 구글(Google)도 Bard를 거쳐 Gemini를 인간과 자연어를 통해 상호작용하는 LLM-AI로서 공개했으며 유료로 사용할 수 있는 더 강력한 Gemini Advanced를 서비스하고 있다. 최근에는 Anthropic에서 Claude를 필두로 Claude-3까지의 LLM-AI를 선보였는데 Claude-3은 IQ 테스트에서 AI 최초로 100을 넘는 점수를 기록하기도 하였다[4]. 특히, 최근에 서비스되고 있는 LLM-AI는 문자뿐만 아니라 그림이나 PDF 파일 내용도 인식할 수 있기에 여러 분야에서 효율적으로 활용될 수 있을 것으로 기대되고 있다.

교육 분야에서 LLM-AI의 활용은 학생에게는 개개인의 학업 능력과 요구에 맞춘 개인 맞춤형 학습을 제공하고 교수자에게는 다양한 교육 전략을 개발하고 적용할 수 있게 한다 [2,3]. 특히, 대면 및 온라인 강의에서 스마트폰, 태블릿 등의 IoT 기기로 LLM-AI를 활용하여 교수자와 학습자 간의 고차원적인 상호작용이 가능하므로[3], 교수자에서 학습자로의 일방적인 지식 전달 방식의 교육을 탈피하여 더욱 창의적인 교육으로의 전환을 실현할 수 있다. 또한 AI는, 기존의 전통적인 평가 방식을 넘어, 시험에 대한 실시간 피드백을 제공함은 물론, 시험 결과를 토대로 학습 진행 상황에 대한 진단과 학습 계획까지 제공하는 혁신적인 평가 방법을 가능하게 한다[2]. 이런 엄청난 잠재력으로 인해 다양한 교육 분야에서 LLM-AI를 활용한 연구들이 시도되었고, 관련 연구 논문은 ChatGPT가 공개된 2022년 말 이후로 급증하고 있다 [1-3,5]. 국내외를 막론하고 LLM-AI의 활용에 관한 대부분의 연구에서는 ChatGPT가 고려되었고 강력한 자연어 처리 능력에 의한 대화형 AI를 표방하는 만큼, 언어 영역에서의 교육적 활용에 대해 가장 많이 연구되었다[1,5].

그러나 교육 목적으로서의 AI의 활용에는 유의할 점들도 있는데, 기본적으로, 모든 학습자가 AI의 활용성에 동등하게 접근할 수 있도록 디지털 빈곤과 격차로 인한 교육적 불공평성이 해결되어야 한다[2]. 또한, AI를 활용한 표절이나 AI에 대한 맹신으로 인해 학업 성취도나 평가의 공정성이 저해되는 문제에 대비해야 한다[3]. 무엇보다도, LLM-AI는 입력된 내용에 대해 확률적인 추론 방식으로 반응 내용을 생성하기에, 편향되거나 부정확한 정보를 마치 사실처럼 출력하는 Hallucination의 특성을 보인다[6,7]. LLM-AI에 입력하는 정보의 종류, 순서 등의 알고리즘, 즉, 프롬프트(prompt)에 따라 이 Hallucination의 발생 확률과 정도를 완화할 수 있지만, 이것을 아직 완전히 회피할 수는 없다[6,7]. LLM-AI에 내재한 이 Hallucination 특성으로 인해, LLM-AI를 특정 분야의 교육에 활용하기 위해서는 그 활용 분야에서 검증된 내용으로 해당 AI의 수준을 객관적으로 파악하는 연구가 필수적으로 선행되어야만 한다.

일반물리학은 다양한 공학 전공 분야를 학습할 수 있는 기초적인 개념과 원리를 제공하기에 이에 대한 단단한 학습은 개별 공학에서의 학습 성취도는 물론, 문제 해결 능력을 향상하는 데도 필수적인 토대가 된다. 일반물리학에서 문제를 풀기 위한 개념들을 선별하고 이들을 적절하게 적용하여 정답을 도출해 내는 과정은 여러 공학에서 문제를 해결하는 과정과 유사하다[8]. 이런 측면에서, LLM-AI를 다양한 공학 분야의 교육에 활용하는 연구에 앞서, 일반물리학에서의 다양한 문제들을 적용하여 LLM-AI의 일반물리학에 대한 수준을 파악하는 연구는 그 가치가 상당하다. 본 연구에서는 국내 H 대학의 공대 신입생들이 수강하는 일반물리학 교과목에서 실제로 출제된 중간 및 기말고사의 문제들을 현재까지 공개된 최신의 LLM-AI 모델들에 적용하여 이들의 문제 풀이 성능을 비교 및 분석하였다.

본 연구는 교수자의 교육 보조 도구로서의 LLM-AI의 물리 문제 풀이 성능을 진단하는 측면도 있지만, 학습자의 자기주도학습 도구로서 LLM-AI의 활용 가능성을 알아보는 측면이 크다. 그러므로 물리 문제를 해결하는 데 필요한 물리 개념과 그 응용 방법을 모르는 학습자의 입장을 가정하여, LLM-AI의 문제 풀이 능력을 측정하고 분석하였다. 만약 LLM-AI가 같은 문제에 대해 정답과 오답을 번갈아 출력한다면, 그 LLM-AI는 해당 문제와 개념을 배우기 위한 학습 도구로서 적절하지 않다. 문제를 입력하는 프롬프트의 내용이나 형식을 변환하여 LLM-AI의 출력 내용을 향상할 수 있지만[9], LLM-AI의 출력 정보를 수동적으로 받아들이는 학습자로서는 능동적인 프롬프트 작업은 쉽지 않다. 예를 들어, 해당 문제를 입력하기 전에 유사한 예제와 풀이로 LLM-AI의 해결 방향을 미리 정해주는 few-shot, 해당 문제만 입력하는 zero-shot, 해당 문제 바로 뒤나 해당 문제에 대한 LLM-AI의 출력에

내용을 추가로 덧붙이는 CoT(Chain-of-Thought) 등으로 프롬프트를 분류할 수 있는데[9], 학습자 대부분은 zero-shot이나 매우 간단한 CoT 정도만 가능하다고 볼 수 있다. 따라서, 본 연구에서는 zero-shot이나 간단한 CoT 프롬프트 방식으로 입력한 물리 문제에 대해, LLM-AI의 물리 문제 풀이 성능과 확률적 추론에 의한 출력 정보의 변화 특성을 조사하였다. 이 연구 과정과 결과는 일반물리학을 배우는 학습자들의 수준별 개인 교사로서의 LLM-AI의 가능성뿐만 아니라, LLM-AI를 타 공학 전공 교육에 적용할 수 있을지에 대한 객관적인 기준을 세우는 데도 도움이 될 것이다.

2. 연구 방법

2.1 연구 배경

ChatGPT를 필두로 한 본격적인 LLM-AI의 도래 이후에 물리 분야에서 LLM-AI의 문제 풀이 능력을 진단하는 연구가 크게 확장되었다. B. Gregorcic과 A. Pendrill은 ChatGPT의 초기 버전으로 자유 낙하 개념에 대한 여러 질문에 대한 반응을 분석하였다 [10]. 추후, R. Santos는 동일한 물리 개념에 대해 ChatGPT3.5, ChatGPT4, Bing Chat, 그리고 Bard를 테스트하였고 ChatGPT4가 가장 좋은 성능을 보임을 보고하였다[11]. G. Kortemeyer는 뉴턴 역학의 기본 개념 문제들로 구성된 Force Concept Inventory(FCI)와 일반물리학 교과목의 숙제와 시험의 몇몇 문제들로 ChatGPT3의 문제 풀이 능력을 분석하였다 [12]. 2023년 초였던 이 당시, ChatGPT의 일반물리학에 대한 최종 성적은 학생들의 학점 기준에 맞춰 4.0 만점에 1.5가 부여되었다. C. West는 FCI 문제들을 2023년 상반기에 ChatGPT3.5와 ChatCPT4에 시도하였고 각각 50~60%와 28/30의 정답률을 보고하였다 [13]. 이 연구는 ChatGPT4의 FCI에 대한 정답률이 ChatGPT3.5보다 월등하게 높을 뿐만 아니라 출력 내용의 일관성도 훨씬 뛰어남을 보여주었다. D. Tong 등은, FCI 문제들을 벗어나, 전자기학 개념이 포함된 중학교와 고등학교 수준의 물리 문제들로 ChatGPT3.5와 ChatGPT4의 수준을 알아보았는데, ChatGPT4는 중학생과 고등학생의 평균 점수보다 훨씬 높은 100점에 가까운 성능을 보여주었다 [14]. W. Yeadon과 D. Halliday는 영국 Durham 대학의 물리학과 학부 및 석사 학생들이 수강하는 물리학 관련 교과목들의 593개 문제를 자체 코딩한 소프트웨어로 ChatGPT3.5와 ChatGPT4의 API를 통하여 zero-shot 프롬프트 형식으로 질문하였고, 각각 49.4%와 38.6%의 정답률을 얻었다 [15]. 그들은 이 결과를 바탕으로 ChatGPT 등의 LLM-AI가 아직 인간에게 큰 위협을 가할 만한 수준이 아니라고 결론지었다. G. Polverini와 B. Gregorcic는 LLM-AI의 작동 방식에 대한 이해와 다양한 프롬프트 기법을 통해 개념적인 물리 문제에 대한 ChatGPT4의 풀이 능력을 탐구하고 분석하여 기존 연구를 보완하였고, LLM-AI를 물리 교육에 적용할 방향을 논의하였다 [7]. K. Wang 등은 일반물리학에서 문제를 해결하기 위한 숫자나 조건들이 문제에 모두 제시되는 '특정된 문제(specified problem)'와 문제에서 조건과 숫자들이 일부만 제시되고 문제를 푸는 주체가 나머지 필요한 것들은 현실적으로 가정하여 문제를 풀어야 하는 '비특정 문제(unspecified problem)'로 ChatGPT4의 물리 문제 풀이 능력을 측정하였다 [8]. 특정 문제와 비특정 문제에 대한 ChatGPT4의 정답 비율은 각각 10/16과 2/24로서 비특정 문제에서 ChatGPT4의 문제 해결 능력이 매우 낮았다. K. Wang 등은 문제 뒤에 단순한 추가 문구를 넣은 CoT 프롬프트 방식으로 ChatGPT4가 틀린 28개의 문제를 다시 시도하였는데, ChatGPT4가 이 중 3개의 문제를 풀었지만 단순 계산 오류는 고쳐지지 않는다고 보고하였다. 국내에서는 ChatGPT3.5 plus를 활용하여 고등학교 물리 및 일반물리학의 수업 계획서, 수업 지도안, 수업 자료 준비 보조, 그리고 몇몇 문제 풀이와 문제 채점 등, 전반적인 교육 활동을 보조하는 수준으로서의 AI의 역할에 관한 연구가 수행되었다[16].

이처럼 LLM-AI의 물리 문제 해결 능력에 관한 심층적인 연구는 주로 해외에서 활발하게 수행되었고 가장 최근의 연구 결과는 ChatGPT4에 대한 것들임을 확인할 수 있었다. 최근 공개된 LLM-AI는 자체 번역 기능을 탑재하고 있지만, LLM-AI를 훈련(pre-training)하는 자료가 영문이 국문보다 월등하게 많을 수밖에 없으므로 영문과 국문 작동 환경에서 성능 차이가 있을 수 있다. 또한, LLM-AI의 성능에 영향을 미치는 각종 프롬프트 기법도 영어와 한국어에 따라 미묘한 차이가 있을 수 있다. 무엇보다도, 지금까지는 완전히 같은 물리 문제에 대해 동일 LLM-AI 모델의 출력 내용이 확률적 추론 특성으로 인해 정답과 오답으로 달라지는 현상에 관한 체계적인 연구가 없었다.

2.2 연구 대상

국내 H 대학교에서는 대부분의 공대 1학년 학생들이 일반물리학을 필수적으로 수강하는데, 수준별 교육과정의 일환으로 공대 신입생 중, 입학 전의 물리 관련 학업 성적이 뛰어난 학생들을 대상으로 하는 고급물리학이 개설된다. 본 연구에서는 2023학년도에 개설되었던 고급물리학1의 중간과 기말고사에서 45개의 문제를 선별하여 최신 LLM-AI의 일반물리학 문제 풀이 능력을 측정하였다. 이 문제들은 필요한 물리 개념을 적용하고 문제에서의 주어진 값들을 적절하게 대입하여 정답을 산출하는 유형, 즉, 특정된 문제(specified problem)이다. 24명의 학생이 응시하였던 중간과 기말고사 시험에서는 이들 문제가 6지 선다형과 단답형 방식으로 출제되었다. 6지 선다형은 학생이 도출한 답안이 제시된 보기에 없으면 해당 문제를 다시 시도하므로 단순 실수에 의한 오답을 완화할 수 있다. 그러나 단답형은 실수가 그대로 오답으로 연결되므로, 단답형의 정답률이 6지 선다형에 비해 좀 더 낮아지는 경향이 있다. 본 연구에서는 LLM-AI에 적용한 문제들에 대한 학생들의 정답률, 즉 학생 입장으로서의 각 문제의 난이도를 활용하여 LLM-AI의 문제 풀이 성능을 비교하고 분석하였다.

선별된 45개의 문제를 현재까지 공개된 최신 LLM-AI인 OpenAI의 GPT-4o와 ChatGPT4, 구글의 Gemini Advanced, 그리고 Anthropic의 Claude-3 Opus에 적용하여 이들 4개의 LLM-AI 모델들이 제시한 풀이 내용을 분석하고 정답과 오답을 결정하였다. 일부 LLM-AI 모델들은 인터넷에 접속할 수 있지만 본 연구에서는 인터넷 접속은 허용하지 않았다. 그림이 포함된 문제를 입력할 때는 [그림: 그림 내용]으로 그림에 관해서 설명하는 문구를 해당 문제 앞에 넣었다. 이들 LLM-AI는 비정기적으로 버전이 업데이트되는데, 본 연구가 행해진 시기는 2024년 4월부터 6월이었다.

2.3 연구 절차

본 연구 절차를 수립하는 과정에서 본 논문의 저자가 파악한 LLM-AI의 문제 풀이 성능과 관계된 몇몇 중요한 특성은 다음과 같다. 현재 공개된 최신의 LLM-AI는 동일 대화창에서 이전 대화 내용을 상당한 부분까지 기억하고 있다. 만약 특정 대화창에서 LLM-AI가 단순한 계산 오류 등을 일으켰다면, 그 대화창에서 다른 물리 문제로 질문하였을 때, 앞에서의 단순한 계산 오류를 반복할 확률이 높다. 또한, LLM-AI의 관점에서 어려운 물리 문제의 경우에는, 확률적 추론 방식으로 인해 한 대화창, D_A에서는 틀린 풀이를, 다른 대화창, D_B에서는 맞는 풀이를 출력할 수 있다. 만약, 두 대화창, D_A와 D_B에서 '지금까지의 모든 대화 내용을 잊고 초기화해라' 는 초기화 프롬프트를 입력하면 LLM-AI는 초기화했다는 대답을 출력한다. 그러나 이 초기화 프롬프트 입력 이전의 물리 문제 내용을 물어보면, 어떤 LLM-AI는 그 내용을 기술하며, 또 다른 LLM-AI는 초기화해서 모른다고 답한다. 더 나아가, 초기화 수행 직후에 각 D_A와 D_B 대화창에서 초기화 프롬프트 이전의 물리 문제를 한 번 더 입력해서 물어보면, D_A 대화창에서는 계속해서 틀린 내용을 출력하고 D_B 대화창에서는 계속해서 맞는 내용을 출력할 확률이 높다. 심지어는 매우 높은 확률로, D_A 대화창에서 초기화 프롬프트를 입력한 전후의 오답 출력 내용을 비교하면 틀린 부분의 세부 내용과 오답까지도 같다.

이런 특이한 LLM-AI의 특성과 물리를 배우는 학습자의 입장을 고려하여 다음과 같이 단계별 연구 절차를 수립하였다.

[연구 절차]

1) 독립된 두 대화창에 22개와 23개, 총 45개의 물리 문제를 별도의 추가 설명이나 지시 없이 오직 물리 문제만을 입력하는 zero-shot 프롬프트로 실험.
2) 1)에서 최종적으로 틀린 문제들에 대해, 문제 바로 뒤에 간단한 CoT를 추가한 zero-shot-CoT 프롬프트로 실험.
3) 2)에서 매우 낮은 정답률을 보이는 문제들에 대해, 문제 내용의 단어와 구문을 변형하여 zero-shot-CoT 프롬프트로 실험.

[공통 사항]

1) 모든 문제에 대해, 각 LLM-AI의 독립된 대화창에서 3번씩 시도. 3번 모두에서 올바른 풀이와 정답을 출력하는 경우에만 최종 정답으로 인정.
2) 같은 대화창에서 오답이 2번 연속으로 나오거나 단순 계산 오류가 몇 문제 안에 다시 나올 때는, 2번째 오답은 무효로 하고 새 대화창을 열어서 그 2번째 물리 문제를 다시 질문.

LLM-AI가 풀이를 출력하다가 풀이 중간에 멈추거

나, '*… 와 같이 풀면 구할 수 있습니다*' 와 같이 정답을 출력하지 않고 끝맺는 경우가 가끔 발생하였는데, 이런 상황에서도 새 대화창을 열어서 다시 해당 물리 문제를 질문하였다. 입력한 문제에 대해 LLM-AI가 출력한 내용이 오답이면, 개념 적용 자체에 오류가 있는 경우와 개념과 풀이 과정은 맞지만 단순한 계산 실수로 인한 오답인지를 구별하였다. 앞의 [연구 절차] 2)에서의 CoT 문구는 LLM-AI의 물리 문제 풀이 능력에 관한 연구를 수행했던 이전 문헌[6-9]과 본 논문 저자의 연구 경험으로 결정하였는데, 문제 바로 뒤에 덧붙인 CoT의 내용은 다음과 같았다.

이 문제를 물리 교수처럼 차근차근 정확하게 풀고 다음 사항들을 지켜라.
1. 수학 계산 부분은 방정식의 이항, 교환 및 결합법칙, 사칙연산 계산 등, 모든 계산을 빠짐없이 단계별로 보여라.

이 CoT 내용은 LLM-AI가 잘못된 물리 개념을 적용하거나 단순 계산 오류를 수행하는 것을 최대한 방지하기 위함인데, 정확하게 위의 내용과 같을 필요는 없다. 만약, LLA-AI가 출력한 오답의 내용을 분석하여 추가적인 CoT 내용이 더 필요하다고 판단될 때는 해당 내용을 '2. …'로 추가한 zero-shot-CoT 프롬프트를 적용하였다. 추가 내용은 해당 물리 문제에 대한 힌트나 중요한 개념이 아닌, 학습자 관점에서 문제의 의도를 다시 한번 명확하게 설명하는 것이다. [연구 절차] 3)의 목적은 인간 관점에서는 같거나 유사한 물리 문제지만 그 문제 내용을 표현하는 단어나 구문이 달라질 때, LLM-AI의 Hallucination 특성을 더 체계적으로 관찰함에 있다. 이들 변형된 문제들을 활용하여 LLM-AI의 확률적 추론 방식에 의한 문제 풀이 성능 변화를 분석하고 이를 통해 LLM-AI의 한계를 가늠하려고 하였다.

3. 연구 결과

3.1 Zero-shot 프롬프트

Table 1.과 2.에 중간과 기말고사 문제들을 zero-shot 프롬프트로 각 LLM-AI에 질문한 결과를 나타내었다. Table 1.과 2.의 A1, A2, A3, 그리고 A4는 각각 GPT-4o, ChatGPT4, Gemini Advanced, 그리고 Claude-3 Opus를 나타낸다. Table 1.과 2.에 각 문제에 해당하는 일반물리학의 소주제와 학생들의 시험에서의 단답형 출제 여부(s), 그리고 학생들의 정답률(%)이 표시되어 있다. LLM-AI의 정답과 오답은 각각 O와 ×로 표기하였는데, ×뒤의 숫자는 독립된 대화창에서 시도한 3번 중, 오답을 출력한 횟수를 나타낸다. Table 1.의 8번과 Table 2.의 13번 문제는 그림을 글로 설명하기 불가능하여 그림 파일 업로드 기능이 없는 ChatGPT4에는 시도하지 않았다(pic로 표기함). Table 1.과 2.의 마지막 줄에는 학생들이 각 시험에서 맞힌 문제들의 평균 개수와 각 LLM-AI 모델의 (3번 시도에서 모두 맞은 문제 개수)/(전체 문제 개수)를 표시하였다.

Table 1. Correct answer rate(CAR) of students and performance of LLM-AI models for midterm exam problems

No.	Area in Physics	CAR	A1	A2	A3	A4
1	Motion in 1 dimension(s)	75.0	O	O	O	O
2	Motion in 1 dimension(s)	45.8	O	O	O	O
3	Vectors	58.3	O	O	O	O
4	Motion in 2 dimensions	83.3	O	O	O	O
5	Motion in 2 dimensions(s)	41.7	O	O	O	O
6	The law of motion	79.2	O	O	O	×3
7	The law of motion(s)	87.5	O	O	O	×1
8	The law of motion	45.8	O	pic	×3	×2
9	The law of motion(s)	33.3	O	×2	O	O
10	The law of motion	62.5	O	×2	×1	O
11	Energy of a system(s)	75.0	O	O	O	O
12	Energy of a system(s)	41.7	O	O	O	O
13	Energy of a system	58.3	O	O	O	O
14	Conservation of energy(s)	54.2	O	O	×1	O
15	Conservation of energy	62.5	O	O	×1	O
16	Linear momentum & collisions	66.7	O	×2	×3	O
17	Linear momentum & collisions(s)	33.3	O	O	O	×1
18	Physics measurement	16.7	×3	×3	×3	×3
19	Circular motion of Newton law	37.5	O	×1	×3	×1
20	Linear momentum & collisions(s)	20.8	×1	×1	×2	O
21	Motion in 1 dimension(s)	25.0	O	×1	×2	×1

22	The law of motion(s)	79.2	O	O	×1	×1
	Student average score, O/Total	11.8	20 /22	14 /21	12 /22	14 /22

Table 1.과 2.에서 살펴보면, 모든 LLM-AI의 성적이 학생들의 평균보다 높음을 알 수 있다. 맞힌 문제 비율은 GPT-4o이 가장 높고, ChatGPT4와 Claude-3 Opus는 비슷하며, Gemini Advanced가 이들보다 소폭 낮다. LLM-AI 전체로 볼 때, 정오답 비율의 일관성이 없는, 즉, 정답과 오답이 반반인 문제는 Table 1.과 2.에서 각각 3개이다. 나머지 문제들은 정답이든 오답이든 3:1 이상으로 LLM-AI의 정오답 비율에 쏠림이 있다. 또한, LLM-AI의 오답 비율이 높은 문제들은 학생들의 정답률도 낮은 경향이 관찰된다. 세부적으로 보면, Table 1.과 2.에서 모든 문제에 대한 학생들의 정답률 평균값은 각각 53.8%와 44.0%인데, 4개의 LLM-AI가 틀린 문제들만의 학생 정답률 평균값은 Table 1.과 2.에서 각각 43.5%와 33.2%였다. 이 차이는 LLM-AI가 틀린 문제들의 난이도가 전체 문제의 평균 난이도와 비교해서 높다는 것을 의미한다. 즉, LLM-AI가 틀리는 문제들은 무작위가 아니고 전체적으로 특정한 분포를 이루고 있음을 알 수 있다. 그러나 Table 1.의 6번과 9번 문제, 그리고 Table 2.의 15번 문제와 같이, 오답이 특정한 하나의 LLM-AI 모델에만 집중되는 상황도 관찰된다.

Table 2. Correct answer rate(CAR) of students and performance of LLM-AI models for final exam problems

No.	Area in Physics	CAR	A1	A2	A3	A4
1	Linear momentum & collisions(s)	58.3	O	O	O	O
2	Linear momentum & collisions	66.7	O	O	×1	O
3	Rotation of rigid objects(s)	20.8	O	O	O	O
4	Rotation of rigid objects(s)	29.2	O	O	O	O
5	Rotation of rigid objects	62.5	O	O	×1	×2
6	Rotation of rigid objects	41.7	O	O	O	O
7	Rotation of rigid objects	20.8	×3	×3	×3	×3
8	Angular momentum(s)	33.3	O	O	O	O
9	Angular momentum	75.0	O	O	O	O
10	Angular momentum	29.2	O	×1	O	O
11	Fluid mechanics(s)	66.7	O	O	O	O
12	Fluid mechanics	70.8	O	O	×1	O
13	Fluid mechanics(s)	75.0	O	pic	O	O
14	Fluid mechanics	54.2	O	O	×1	×1
15	Oscillatory motion	29.2	O	O	×3	O
16	Oscillatory motion(s)	25.0	O	O	O	×1
17	Oscillatory motion	20.8	×2	×1	×3	×3
18	Wave motion(s)	16.7	O	O	O	O
19	Wave motion(s)	54.2	O	O	O	O
20	Wave motion	70.8	O	O	O	O
21	Superposition & standing wave	50.0	×2	×1	O	×2
22	Superposition & standing wave(s)	4.2	×3	×2	×3	×3
23	Superposition & standing wave	37.5	O	×2	×2	O
	Student average score, O/Total	10.1	19 /23	16 /22	14 /23	16 /23

Table 1.의 8번은 줄 2개로 연결된 평형 상태에 있는 물체의 그림에서 줄에 작용하는 장력을 구하는 문제였고, Table. 2의 13번은 자동차 유압 장치 그림에서 파스칼의 원리를 적용하여 힘을 구하는 문제였다. 학생들의 정답률을 보면 알 수 있듯이, 8번 문제는 까다로운 문제인 반면, 13번은 매우 잘 알려진 쉬운 문제이다. 두 문제 모두 LLM-AI가 그림을 인식해야 하는 문제인데, 13번은 모두 정답이지만, 8번 문제에 대해서는 Gemini Advanced와 Claude-3 Opus의 정답률이 낮다. 이 두 LLM-AI가 출력한 내용에서는 힘의 평형에 대한 개념 적용이 그림과 전혀 맞지 않아, 마치 이들 LLM-AI가 그림을 제대로 인식하지 못하는 것처럼 보였다. 또한, LLM-AI가 출력한 내용에서 개념 적용과 풀이는 맞는데도 계산기로도 쉽게 정답을 도출할 수 있는 단순한 계산에 대한 오류로 오답 처리되는 경우가 많았다. GPT-4o는 단순 계산 오류의 빈도가 낮았지만, 나머지 LLM-AI 모델들에서는 좀 더 빈번한 계산 오류가 관찰되었다. 특이한 점은 오류가 발생한 단순 계산 부분만 발췌하여 해당 LLM-AI의 새로운 대화창에서 물어보면 그 LLM-AI가 정확한 답을 제시하는 확률이 매우 높았다는 것이다. 이는 LLM-AI의 단순 계산 오류는, 물리나 공학 문제의 풀이처럼, 개념을 적용하고 설명하는 부분과 숫자들을 대입해서 계산하는 부분이 혼재된 작업을 수행할 때, 훨씬 더 빈번하게 발생함을 의미한다.

각 문제당 3번의 독립적인 풀이 시도에서 모든

LLM-AI가 2번 이상 정확한 풀이와 정답을 제시한 문제들이 있다. 즉, 이들 문제에 대해서는 LLM-AI가 매우 낮은 확률로 오답을 제시한다고 볼 수 있는데, Table 1.에서 7번, 14번, 15번, 17번, 그리고 22번이며, Table 2.에서 2번, 10번, 12번, 14번, 그리고 16번이 해당 문제들이다. 만약 본 연구에서처럼 같은 문제를 독립적으로 3번 시도하지 않았다면 이 10개의 문제는 4종의 LLM-AI가 항상 정답을 맞힌다고 잘못 인식되었을 가능성이 높다. 이들 문제에 대한 LLM-AI의 문제 풀이 결과는 인간과 다르게 작동하는 LLM-AI의 확률적 추론 방식의 특성을 잘 보여준다.

3.2 Zero-shot-CoT 프롬프트

Table 1.과 2.에서 각 LLM-AI가 최종적으로 틀린 문제들에 대해 앞서 '2.3 연구 절차' 에서 명시하였던 CoT 문구를 문제 바로 뒤에 추가해서 다시 실험을 수행하였다. Table 3.에 이 새로운 프롬프트를 적용한 결과와 4개 LLM-AI의 최종 정답 비율 값들이 표시되어 있다. Zero-shot에서와 같이, 모든 문제를 독립적인 대화창에서 3번씩 시도하였다. [문제번호:O]는 zero-shot에서는 틀린 문제가 zero-shot-CoT의 3번의 시도에서 모두 정답으로 변경된 경우이고, [문제번호:×(숫자)]는 zero-shot-CoT에도 불구하고 3번의 시도 중 한 번이라도 오답이 나타난 경우이다. 여기서 (숫자)는 3번의 시도 중에 오답을 출력한 횟수를 나타낸다. 또한 각 LLM-AI에 표시된 N(×) 값들은 모든 틀린 문제들에 대해, zero-shot에서 zero-shot-CoT로 프롬프트를 변경함으로써 해당 LLM-AI가 제시했던 총 오답 횟수의 변화를 나타낸다.

Table 3.은 zero-shot-CoT 프롬프트의 문구들이 LLM-AI의 작동 방식에 영향을 미쳐, 모든 LLM-AI의 물리 문제 풀이 성능이 확률적으로 향상되었음을 보여준다. 참고로, 새로운 프롬프트의 수학 계산과 관련한 문구로 인해 단순 계산 오류가 발생하는 비율이 줄었지만, 여전히 단순 계산 오류는 관찰되었다. 세부적으로 보면, A2(ChatGPT4)와 A4(Claude-3 Opus)는 zero-shot-CoT 프롬프트에 의한 성능 향상 폭이 비슷했고 A3(Gemini Advanced)의 성능 향상 폭이 가장 컸다. 반면, A1(GPT-4o)은 새로운 프롬프트로 성능이 향상된 것은 확실하지만, zero-shot에서도 정답률이 이미 높았던 영향인지, 최종 정답률로만 보면 중간과 기말시험에서 각 1문제씩만 정답으로 상향되었다. 또한, Table 3.의 결과는 현재 최신 LLM-AI의 관점에서 난이도가 일정 수준 이상인 일반물리학 문제는 본 논문에서 제시한 CoT 문구가 포함된 프롬프트라도 여전히 공략하기 어렵다는 것을 보여준다.

Table 3. Performance of LLM-AI models with zero-shot-CoT prompts for Introductory physics problems

	Midterm problems		Final problems	
A1	[18:×2], [20:O]	21 /22	[7:×3], [17:×1], [21:O], [22:×2]	20 /23
	N(×): 4 → 2		N(×): 10 → 6	
A2	[10:×1], [16:O], [18:×3], [19:×1], [20:O], [21:×1]	17 /21	[7:×2], [10:O], [17:×1], [21:×2], [22:×2]	18 /22
	N(×): 11 → 6		N(×): 10 → 7	
A3	[8:×3], [10:×1], [14:O], [15:O], [16:O], [18:×3], [19:×3], [20:O], [21:O], [22:O]	18 /22	[2:O], [5:O], [7:×3], [12:O], [14:O], [15:×2], [17:O], [22:×2], [23:×1]	19 /23
	N(×): 20 → 13		N(×): 18 → 8	
A4	[6:×1], [8:×3], [17:O], [18:×3], [19:×2], [21:O], [22:O]	18 /22	[5:×1], [7:×3], [14:O], [16:O], [17:×2], [21:×1], [22:×3]	18 /23
	N(×): 12 → 9		N(×): 15 → 10	

3.3 LLM-AI의 확률적 추론 특성과 한계

이번 절에서는 4개의 LLM-AI 모델 모두가 공략하지 못했던 3개의 물리 문제를 통해 이들 LLM-AI의 물리 문제 풀이 출력의 특성과 한계를 유추해 본다. Table 3.을 살펴보면 모든 LLM-AI가 중간고사의 18번, 그리고 기말고사의 7번과 22번 문제에 대해서 정답을 제시하지 못하였다. 18번 문제는 유효숫자 법칙에 따른 측정값의 계산에 관한 문제로서, 문제 내용은 다음과 같다.

유효숫자로 이루어진 측정값의 연산, sin(76°) ×4.739 - 2.74를 올바르게 나타낸 것은?

Zero-shot으로 위의 문제만 입력했을 때는 모든 LLM-AI가, 문제에 '유효숫자로 이루어진 측정값' 이라는 문구가 있음에도 불구하고, sin(76°)를 수학적 계산으로 임의로 가정하여 계산값에 무한대의 유효숫자를 채택하였다. 이에, zero-shot-CoT에서는 연구 계획에서 명시한 CoT에 '2. 문제의 모든 숫자는 측정값이다.' 라는 문구를 추가하여 실험을 수행하였다. 그러나 이 경우에는 대부분 LLM-AI가 sin(76°)의 계산값에서 유효숫자를 정확하게 산출하였음에도 불구하고, 단 한 번 GPT-4o의 정답을 제외한 나머지 모든 경

우에서 최종적으로는 오답을 제시하였다. 이는 LLM-AI가 유효숫자 법칙에 기반한 연속적인 수학 계산에 취약하다는 것을 의미한다.

기말고사의 7번 문제는 ChatGPT4가 3번의 시도 중 한 번만 정확한 풀이와 답을 제시하였는데, 문제는 다음과 같다.

한쪽 끝이 회전축인 막대가 수평 상태에서 정지해 있다. 이 막대를 놓으면 무게로 인해 한쪽 끝을 회전축으로 하여 회전한다. 이 막대의 질량은 2.3 kg이고 길이는 4.0 m이다. 막대가 수평선과 막대 사이의 각도가 35° 까지 회전했을 때, 이 막대 질량 중심의 회전 접선 속력[m/s]은?

이 문제는 수평 막대가 회전하면서 질량 중심의 높이가 낮아지므로 막대의 중력 퍼텐셜에너지(위치 에너지) 감소량이 회전운동에너지로 전환되는 개념을 적용해야 한다. 특이한 점은 LLM-AI가 출력한 내용의 상당수에서 오답의 유형이 같았는데, 다음과 같이 막대의 중력 퍼텐셜에너지 변화를 잘못 산정한 경우였다.

(중략)
위치 에너지 계산
$\triangle U = mg\left(\frac{L}{2} - \frac{L}{2}\cos(35^\circ)\right)$
$\triangle U = 2.3 \times 9.8 \times (2.0 - 2.0\cos(35^\circ))$
(중략)

중력 퍼텐셜에너지의 감소량은 막대 질량 중심의 높이 차이에 의존하는데, 코사인 함수가 포함된 괄호 부분이 잘못되어 있다.

마지막 22번 문제는 도플러 효과에 의한 음파의 주파수 변이에 관한 것이었는데, 문제 내용은 다음과 같다.

747 비행기가 1770 Hz 주파수로 소리를 내면서 음속의 0.5 배의 속도로 경비행기로 접근한다. 이 경비행기는 음속의 0.05 배로 747 비행기로부터 멀어지고 있다. 이 경비행기에 탄 사람이 듣게 되는 747 비행기 소리의 주파수는 몇 Hz인가? (소리의 속도는 343 m/s)

음원인 747 비행기와 청취자인 경비행기가 모두 움직이는데, 접근하거나 멀어지는 경우를 파악하여 각 비행기의 속도에 대한 부호를 결정해야 하는 것이 관건이다. 대부분 LLM-AI가 정확한 도플러 주파수 변이 식을 제시하였으나, 몇몇 경우를 제외하고 이 식에 문제의 값들을 정확하게 대입하지 못하였다. Table 3.에서 보면, 각 4종의 LLM-AI에 3번씩 시도한 총 12번 중, 3번을 맞혔는데, 이는 마치 LLM-AI가 도플러 주파수 변이에 대한 정확한 이해 없이 문제에서 제시된 각 비행기의 속도와 부호를 무작위로 시도해서 우연히 맞힌다는 느낌이 들었다. 연구 계획에서 명시한 CoT에 '2. 747 비행기는 접근하고 경비행기는 멀어진다.' 라는 강조 문구를 추가하여 실험을 수행해도 결과는 Table 3.과 비슷했다. 또한, LLM-AI의 한국어 번역 능력에 의한 문제 해석 오류에 의한 오답인지 확인하기 위해, 원본 문제에서 비행기 부분만을

747 비행기가 1770 Hz 주파수로 소리를 내면서 음속의 0.5 배의 속도로 경비행기 방향으로 움직인다. 이 경비행기는 음속의 0.05 배로 747 비행기의 반대 방향으로 움직인다.

와 같이 바꿔서 의도를 명확하게 표현하였는데도 4종의 LLM-AI의 오답 비율은 Table 3.과 크게 달라지지 않았다.

본 논문 저자의 생각으로는, 국내외의 대학에서 사용하는 일반물리학 교과서들에 위의 문제들에 적용되는 개념에 관한 설명과 예시 문제들은 있지만, 정확하게 이 논문에서와 같은 형태의 문제는 거의 없다고 본다. 예를 들어, 두 유효숫자의 곱셈이나 뺄셈에 대한 예시는 있지만 삼각함수가 포함되면서 이 둘을 결합한 문제는 거의 찾아보기 어려울 것이다. 마찬가지로, 기말고사 7번은 수평 막대가 회전하여 가장 낮은 위치(즉, 수직)인 순간에 대한 문제는 있지만, 회전하여 내려오는 중간 상태에서 질량 중심의 접선 속력을 구하는 문제는 찾아보기 어려울 것이다. 도플러 주파수 변이 문제는 음원과 청취자가 모두 움직이면서 해당 문제와 같이 한쪽은 다가가고 다른 한쪽은 멀어지는 경우는 드물 것이다.

LLM-AI 관점에서 드문 형태의 물리 문제의 의미와 이에 대한 출력 반응을 이해하기 위해 기말시험의 7번과 22번 문제를 변형한 아래의 문제들로 추가 실험을 시행하였다.

[7-1]
(중략) 막대가 수평선으로부터 가장 낮은 위치까지 회전했을 때, 이 막대 질량 중심의 회전 접선 속력[m/s]은?

[7-2]
(중략) 막대가 수평선으로부터 각도, θ = 90° 까지 회전

했을 때, 이 막대 질량 중심의 회전 접선 속력[m/s]은?

[22-1]
747 비행기가 1770 Hz 주파수로 소리를 내면서 음속의 0.5 배의 속도로 정지해 있는 사람에게서 멀어지는 방향으로 움직인다. 이 사람이 듣게 되는 747 비행기 소리의 주파수는 몇 Hz인가?

[22-2]
747 비행기가 1770 Hz 주파수로 소리를 내면서 음속의 0.5 배의 속도로, 경비행기가 음속의 0.05 배로, 두 비행기가 서로 멀어지는 방향으로 움직인다. 경비행기에 탄 사람이 듣게 되는 747 비행기 소리의 주파수는 몇 Hz인가?

[22-3]
747 비행기가 1770 Hz 주파수로 소리를 내면서 음속의 0.5 배의 속도로 경비행기로부터 멀어지는 방향으로 움직인다. 이 경비행기는 음속의 0.05 배로 747 비행기로부터 멀어지는 방향으로 움직인다. 이 경비행기에 탄 사람이 듣게 되는 747 비행기 소리의 주파수는 몇 Hz인가?

문제, 7-1과 7-2은 사실상 같은 문제인데, 막대가 제일 낮은 부분에 위치한다는 부분을 서술한 방식만 다르다. 22-1은 LLM-AI가 도플러 주파수 변이 물리 문제를 풀 수 있는지 알아보기 위해 원래의 22번 문제보다 쉽게 만든 문제이고, 22번과 유사한 난이도의 22-2와 22-3도 사실상 같은 문제인데, 두 비행기가 서로 멀어지는 방향으로 움직인다는 부분의 표현만 다르다. 이 추가 실험도 앞의 zero-shot-CoT 방식으로 각 3번씩 독립된 대화창에서 시행하였는데, 그 결과가 Table. 4에 표시되어 있다. 비교를 위해, Table. 3의 7번과 22번 문제의 결과를 다시 표시하였다. 참고로, *의 표시는 해당 LLM-AI가 출력한 전체 풀이는 맞는데, 문자와 숫자의 단순한 계산 오류 때문에 오답이 된 경우이다.

7-1은 일부 일반물리학 교과서 등에서 잘 알려진 형태의 문제인데, 예상대로 A1(GPT-4o)과 A2(ChatGPT4)가 3번의 시도에서 모두 정확한 풀이를 출력하였다. 그러나 A3(Gemini Advanced)과 A4(Claude-3 Opus)는 저조한 정답률을 보여주었다. A3는 단순 계산 오류가 있지만, A4는 전혀 풀지 못한 결과가 매우 의외였는데, A4는 이 7번 물리 문제가 속한 소분야(회전하는 강체의 운동)에 대해서 다른 LLM-AI보다 학습(pre-training)이 충분하지 않은 상태라고 짐작된다. 7-2는 A3와 A4가 7-1보다 한 문제씩 더 정답을 출력하였지만, 모든 LLM-AI가 3번의 시도를 통과하지 못했다. 또한, 앞에서 언급했던 막대 중심의 높이 차이를 잘못 산정한 코사인 함수가 포함된 오류 식에 90도를 대입하면 우연히 7-2문제에서 요구하는 정확한 높이 차이가 도출된다. 이런 이유로 Table. 4에 표시된 7-2에 대한 LLM-AI의 정답률이 실제보다 좀 더 높게 나왔을 것이라 본다. 결과적으로, 7-1과 7-2에서 4종의 LLM-AI가 출력한 전체 오답 횟수는 같았지만, 잘 알려진 7-1의 형태보다는 7-2문제에 대한 최종 정답률이 낮았다고 결론지을 수 있다.

Table 4. Performance of LLM-AI models with modified problems of 7 and 22 in the final exam.

	7	7-1	7-2	22	22-1	22-2	22-3
A1	×3	○	×1	×2	○	×1	×2
A2	×2	○	×1	×2	○	○	×3
A3	×3	×2*	×1	×2	○	×1*	×2
A4	×3	×3	×2	×3	○	○	×2

Table. 4에서 22-1번의 결과는 모든 LLM-AI가 도플러 주파수 변이 식을 적절하게 사용하고 정답을 도출할 수 있다는 것을 보여준다. 그러나 두 비행기 모두 움직이는 상황을 묘사한 22-2와 22-3문제에서 '두 비행기가 서로 멀어지는' 이라는 문구가 포함된 22-2문제의 정답률이 22나 22-3보다는 훨씬 높다. 22-3문제는 22-2문제와 그 내용은 완전히 같은데도 LLM-AI의 정답률이 매우 낮았는데, 이는 LLM-AI가 훈련(pre-trainig)한 물리 분야의 언어 자료에서 두 물체가 서로 멀어지는 방향으로 움직이는 상황을 22-3처럼 표현한 문구들의 빈도수가 상대적으로 매우 작았다고 유추할 수 있다. 결론적으로, LLM-AI가 친숙하지 않은, 그래서 정답률이 낮은 물리 문제는 그 문제에서 사용된 단어들과 구문들이 훈련(pre-training)에 사용되었던 대규모 언어 자료에서 매우 낮은 빈도수로 나타났었다는 의미이다.

인간 학습자나 전문가 관점에서는 7번과 7번의 변형 문제들, 혹은 22번과 22번의 변형 문제들은 각각 같은 물리 개념을 활용하는 문제들이다. 결국, LLM-AI가 이런 유사한 물리 문제 모두에 대해 정답을 출력하는 비율을 높이기 위해서는 훈련으로 습득한 '물리 개념의 이해와 적용' 부분과 '언어적 해석' 부분, 이 두 부분을 동시에 유기적으로 혼합하고 응용하는 수준이 되어야만 한다. 이는, 앞서 언급했던 LLM-AI의 단순한 수학 계산 오류가 물리나 공학 문제 풀이에 포함되어 있을 때만 상대적으로 높은 확률로

나타나는 LLM-AI의 고질적인 오류와 그 맥락이 같다. LLM-AI가 여러 영역들을 유기적으로 혼합해서 응용하는 능력은 쉬운 물리 문제들에서는 높은 확률로 정답을 출력할 만큼 구현된 것처럼 보인다. 하지만, 본 연구의 결과는 현재 공개된 최신 LLM-AI 모델들이라 할지라도, 난이도가 상향된 물리 문제들에 대해서는 아직 이런 수준까지는 도달하지 못했다는 것을 보여준다.

3. 결론 및 제언

본 연구에서는 국내 H 대학에서 실시하였던 일반물리학 교과목의 중간과 기말고사에서 발췌한 45개의 문제로 현재 공개된 최신 4종의 LLM-AI 모델들의 문제 풀이 성능을 측정하였다. 각 LLM-AI의 독립된 대화창에서 문제마다 3번씩 시도하여 모두 정답인 경우만 최종 정답으로 인정하였다. 오직 물리 문제만 입력하는 zero-shot 프롬프트 방식으로 조사한 결과, GPT-4o, ChatGPT4, Gemini Advanced 그리고 Claude-3 Opus가 각각 86.7%, 69.8%, 57.8%, 그리고 66.7%의 정답률을 보였다. 본 연구에서 제시한 CoT가 포함된 zero-shot-CoT를 적용하면 각 LLM-AI의 정답률이 91.1%, 81.4%, 82.2%, 그리고 80.0%로 전반적으로 향상되었다. 이 새로운 프롬프트에서는 LLM-AI에서 고질적으로 빈번하게 나타나는 단순 계산 오류가 어느 정도까지는 줄어드는 것을 관찰하였다. 이 45개의 시험문제에 응시한 학생들의 정답률은 48.7%였기에 LLM-AI의 물리 문제 풀이 능력이 학생들보다는 훨씬 높다고 결론지을 수 있었다.

이전의 유사한 연구와는 달리, 본 연구에서는 모든 문제에 대해 LLM-AI의 확률 추론에 의한 출력 내용의 Hallucination을 체계적으로 연구하였다. 통상적인 인간의 관념으로는 LLM-AI가 특정 물리 문제를 풀 수 있는지 없는지의 두 가지 기준으로만 LLM-AI의 수준을 평가하려는 경향이 있다. 그러나 본 연구의 결과는 LLM-AI의 문제 풀이 성능 평가는 문제의 정답을 출력하는 확률로서 접근해야만 한다는 것을 보여준다. 예를 들어, LLM-AI 기준에서 쉬운 문제는 정확한 풀이와 정답을 출력할 확률이 매우 높은 문제이다. 입력하는 문제가 점점 어려워지게 되면, LLM-AI가 해당 문제의 정답을 출력할 확률이 점점 낮아지게 된다. 그러나 여전히 매우 높은 확률로 정확한 풀이와 정답을 보여줄 수도 있다. LLM-AI 기준에서 어려운 문제는 대부분 오답을 출력하지만 아주 가끔 정답을 출력할 때도 있다. 또한, 같은 개념의 유사한 물리 문제들을 활용하여 LLM-AI의 확률적 추론 방식은 입력 내용의 단어와 구문에 큰 영향을 받는다는 특성과 이로 인한 한계를 알아보았다. 본 연구의 결과가 보여준 LLM-AI의 확률 추론적 특성을 고려하면 앞으로 AI의 반응을 분석하는 연구에서는 입력 내용의 변화와 시도 횟수에 따른 출력 내용의 변화, 특정 내용을 출력할 확률 등에 대한 논의가 병행되어야 한다고 본다.

비록 본 논문의 일반물리학 문제를 해결하는 성능 연구에서 LLM-AI 모델 4종이 격차를 보였지만, 이 연구 결과가 이들 LLM-AI의 전반적인 성능 차이를 의미하는 것은 아니다. 본 연구는 통상적으로 한 학기 동안 행해지는 일반물리학의 일부 주제들에 대한 45개의 문제만을 다루었다. 또한 각 문제를 3번씩만 시도하였기에, 적은 시도 횟수로 인한 통계적인 한계도 있다. 일반물리학 이외의 수많은 분야, 그리고 무엇보다도 전체적인 글의 맥락을 이해하고 처리하는 능력 등, LLM-AI의 전반적인 성능을 평가하는 지표들은 본 연구에서 행해진 일반물리학 문제 풀이보다 더 확장된 개념이다.

이런 점에도 불구하고, 본 연구에서 실제 물리 시험문제들에 대해 LLM-AI 모델들이 보여준 정답률은 인상적이었다. 특히 GPT-4o는 중위권 이하의 대학에서 다루는 일반물리학 교과의 거의 모든 문제에 대해 높은 확률로 올바른 풀이 과정과 정답을 제시할 수 있는 수준에 도달했다는 것을 보여준다. 또한, 본 저자는 이들 LLM-AI가 올바른 풀이와 정답을 출력한 뒤, 풀이의 특정 부분에 대해 더 자세한 설명을 요청하면, 학습자가 충분히 이해할 만큼 더 자세한 설명을 제공하는 것을 확인하였다. 이는 LLM-AI가 물리 문제 풀이뿐만 아니라 학습 내용에 대한 설명까지 학습자를 위한 효율적인 학습 보조 도구로서의 가치가 높다는 의미이다.

대학 신입생들은 여러 가지 요소들로 인해 같은 대학의 학생들이라도 일반물리학 등의 기초 교과목들에 대한 학업 역량에 차이가 있다. 단일화된 강의 중심의 교육은 이런 다양화된 학생 집단에 대한 보편적인 교육으로서는 적합하지 않다. 대신, 강의뿐만 아니라 다양한 학습 자원을 활용하여 학생들의 개별적인 학습 요구를 충족하는 동시에, 개별 역량 차이를 고려한 개인별 맞춤형 교육이 필요하다. 본 연구 결과는 현재 공개된 LLM-AI가 학생 개개인의 수준과 역량에 맞춘 개인 수준별 학습을 보조할 수 있는 도구로서의 충분

한 가능성을 가지고 있음을 보여준다. 더 나아가, AI의 급격한 발전 속도를 고려하면 전통적인 교육 방식에 큰 변화가 있어야만 한다. 예를 들어, 이제는 전통적으로 행해졌던 과제나 take-home 시험과 같은 방식은 더 이상 평가로서의 의미가 없다고 판단된다. 또한, 교육을 위한 새로운 문제들의 제작과 검증도 LLM-AI를 활용하면 훨씬 다양하고 효율적으로 수행할 수 있을 것이다. 본 연구에서 보여준 LLM-AI의 수준과 나날이 발전하는 AI 기술까지 고려하면, AI를 활용한 교육적 변화와 요구는 필수적인 흐름이며, 이에 대처하는 방편으로서 AI의 수준과 특성을 파악하여 AI를 교육에 융합하는 연구가 더욱 활성화되기를 기대한다.

참고문헌

[1] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology, 1*(1), 100017. DOI : /10.1016/j.metrad.2023.100017

[2] Gill, S. S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., Fuller, S., Singh, M., Arora, P., Parlikad, A. K., Stankovski, V., Abraham, A., Ghosh, S. K., Lutfiyya, H., Kanhere, S. S., Bahsoon, R., Rana, O., Dustdar, S., Sakellariou, R., Uhlig, S., & Buyya, R. (2024). Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots. *Internet of Things and Cyber-Physical Systems, 4*, 19-23. DOI : 10.1016/j.iotcps.2023.06.002

[3] Onesi-Ozigagun, O., Ololade, Y. J., Eyo-Udo, N. L., & Ogundipe, D. O. (2024). Revolutionizing education through AI: A comprehensive review of enhancing learning experiences. *International Journal of Applied Research in Social Sciences, 6*(4), 589-607. DOI: 10.51594/ijarss.v6i4.1011

[4] Jones, M. (2024, March 5). A*Is ranked by IQ: AI passes 100 IQ for first time, with release of Claude-3*. Maximum Truth. URL: https://www.maximumtruth.org/p/ais-ranked-by-iq-ai-passes-100-iq

[5] Lee, S., & Song, K. (2023). Exploration of Domestic Research Trends on Educational Utilization of Generative Artificial Intelligence. *The Journal of Korean Association of Computer Education, 26*(6), 15-27. DOI : 10.32431/kace.2023.26.6.002

[5] Lee, S., & Song, K. (2023). Exploration of Domestic Research Trends on Educational Utilization of Generative Artificial Intelligence. *The Journal of Korean Association of Computer Education, 26*(6), 15-27. DOI : 10.32431/kace.2023.26.6.002

[6] Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023, December). Mathematical Capabilities of ChatGPT. *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.* (pp. 1-46). New Orleans, USA

[7] Polverini, G., & Gregorcic, B. (2024). How understanding large language models can inform the use of ChatGPT in physics education. *European Journal of Physics, 45*(2), 1-35. DOT : 10.1088/1361-6404/ad1420

[8] Wang, K. D., Burkholder, E., Wieman, C., Salehi, S., & Haber, N. (2024). Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. *Frontiers in Education, 8*, 1330486. DOI : 10.3389/feduc.2023.1330486

[9] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).* (pp. 1-15)

[10] Gregorcic, B., & Pendrill, A.-M. (2023). ChatGPT and the frustrated Socrates. *Physics Education, 58*(3), 035021. DOI : 10.1088/1361-6552/acb2e1

[11] dos Santos, R. P. (2023). Enhancing Physics Learning with ChatGPT, Bing Chat, and Bard as Agentsto-Think-With: A Comparative Case Study. *arXiv preprint*, arXiv : 2306.00724v1

[12] Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course?. *Physical Review Physics Education Research, 19*, 010132. DOI : 10.1103/PhysRevPhysEducRes.19.010132

[13] West, C. G. (2023). Advances in apparent conceptual physics reasoning in GPT-4. *arXiv preprint* arXiv : 2303.17012.

[14] Tong, D., Tao, Y., Zhang, K., Dong, X., Hu, Y., Pan,

S., & Liu, Q. (2023). Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education. *Asia Pacific Education Review.* DOI: 10.1007/s12564-023-09913-6

[15] Yeadon, W., & Halliday, D. P. (2023). Exploring Durham University Physics Exams with Large Language Models. *arXiv preprint* arXiv : 2306.15609.

[16] Kim, J., & Yoo, H. (2023). Exploring the use of ChatGPT in physics education: Focusing on high school and general physics classes. *The Journal of Korean Association for Science Education, 17*(3), 216-233.

강 동 열

1999년 부산대학교 물리학과(이학사)
2001년 한국과학기술원 물리학과(이학석사)
2008년 College of Optical Sciences, Univ. of Arizona(이학박사)
2013년 ~ 현재 국립한밭대학교 기초과학부 부교수
관심분야: 교과교육, 이러닝, 광 시스템 분석, 광 데이터 처리
E-Mail: dykang@hanbat.ac.kr