DOI: https://doi.org/10.13161/kibim.2023.13.3.021

BIM 운용 전문가 시험을 통한 ChatGPT의 BIM 분야 전문 지식 수준 평가

Evaluating ChatGPT's Competency in BIM Related Knowledge via the Korean BIM Expertise Exam

최지원[1], 구본상[2], 유영수[3], 정유정[4], 함남혁[5]
Choi, Jiwon[1] · Koo, Bonsang[2] · Yu, Youngsu[3] · Jeong, Yujeong[4] · Ham, Namhyuk[5]

Received June 16, 2023; Received September 05, 2023 / Accepted September 10, 2023

ABSTRACT: ChatGPT, a chatbot based on GPT large language models, has gained immense popularity among the general public as well as domain professionals. To assess its proficiency in specialized fields, ChatGPT was tested on mainstream exams like the bar exam and medical licensing tests. This study evaluated ChatGPT's ability to answer questions related to Building Information Modeling (BIM) by testing it on Korea's BIM expertise exam, focusing primarily on multiple-choice problems. Both GPT-3.5 and GPT-4 were tested by prompting them to provide the correct answers to three years' worth of exams, totaling 150 questions. The results showed that both versions passed the test with average scores of 68 and 85, respectively. GPT-4 performed particularly well in categories related to 'BIM software' and 'Smart Construction technology'. However, it did not fare well in 'BIM applications'. Both versions were more proficient with short-answer choices than with sentence-length answers. Additionally, GPT-4 struggled with questions related to BIM policies and regulations specific to the Korean industry. Such limitations might be addressed by using tools like LangChain, which allow for feeding domain-specific documents to customize ChatGPT's responses. These advancements are anticipated to enhance ChatGPT's utility as a virtual assistant for BIM education and modeling automation.

KEYWORDS: ChatGPT, Large Language Model, Generative AI, BIM Expertise Exam, BIM
키워드: ChatGPT, 대규모 언어 모델, 제너레이티브 AI, BIM 운용 전문가 시험, 빌딩정보모델

1. 서론

1.1 연구의 배경 및 목적

최근 OpenAI사가 개발한 ChatGPT가 큰 각광을 받고 있다. ChatGPT는 거대 언어 모델(Large Language Model, LLM)인 GPT(Generative Pretrained Transformer) 모델을 대화형 에이전트로 학습 시킨 모델로서, 사용자의 질문(prompt)에 적합한 답변을 할 수 있게 되었다.

ChatGPT의 높은 성능으로 인해 검색 방법에 대한 패러다임이 바뀌고 있으며, GPT-3의 경우 175B개의 패러미터와 600GB에 이르는 대량의 웹 정보를 학습시켜 일상 대화 뿐 아니라 특정 전문 분야에 대한 질의에도 답변이 가능해지고 있다.

이에 따라 각 전문 분야에 대한 지식 수준을 평가하기 위해 주된 전문가 자격 시험에 ChatGPT의 성능을 테스트하는 연구들이 등장하였다. 대표적으로 의사(Kung et al., 2023; Kasai et al., 2023) 및 변호사 자격증 시험(Choi et al., 2023; Bommarito II and Katz, 2022; OpenAI, 2023)과 미국의 대학 및 대학원 능력 시험(OpenAI, 2023)에 적용한 사례를 들 수 있으며, 대부분 학생들의 평균에는 못 미치지만 최저 점수는 통과하는 결과를 보여주고 있다.

이들의 테스트를 통해 ChatGPT의 성능이 입증되고 있으며 이에 따라 본 기술을 각 전문 분야의 가상 보조 원(Virtual Assistant, VA) 및 개인 AI(Artificial Intelligence) 교습 도구로 사용할 수 있을 것으로 내다보고 있다.

[1]학생회원, 서울과학기술대학교 건설시스템공학과 석사과정 (wldnjs-823@seoultech.ac.kr)
[2]정회원, 서울과학기술대학교 건설시스템공학과 교수, 공학박사 (bonsang@seoultech.ac.kr) (교신저자)
[3]학생회원, 서울과학기술대학교 건설시스템공학과 박사과정 (youngsu@seoultech.ac.kr)
[4]학생회원, 서울과학기술대학교 건설시스템공학과 석사과정 (yujeong0819@seoultech.ac.kr)
[5]정회원, 한양사이버대학교 디지털건축도시공학과 교수 (nhham@hycu.ac.kr)

BIM(Building Information Modeling)의 적용이 국내에서 의무화되는 추세이지만 BIM의 이론 및 관련 전문 소프트웨어를 다룰 수 있는 전문 인력은 여전히 모자란 실정이다(KIBIM, 2023). 또한 BIM 소프트웨어를 다루기 위해서는 이론 뿐 아니라 상세한 조작 방법을 일일이 익혀야 하는데 이는 종종 BIM 활용의 실질적 진입장벽으로 작용한다(Seo and Ju, 2012). 이에 따라 기존 설계자나 기술자들이 새로운 도구를 쉽게 배울 수 있는 방법이 마땅치 않은 실정이다.

최근 BIM 분야에서도 ChatGPT 활용을 시도한 연구가 등장하였다. 대표적으로 BIM 모델 내 정보를 검색하는 기술(Zheng and Fischer, 2023), BIM과 LLM을 통합하여 공사 일정을 자동으로 관리하는 연구(Singh et al., 2023)가 존재한다. 그러나 이들 연구는 BIM 모델 내의 특정 정보를 탐색하는 기능으로 ChatGPT를 쓴 것으로 BIM 전반에 대한 지식수준을 평가하지는 못한다. 따라서 본 연구에서는 ChatGPT의 BIM과 관련된 지식수준을 평가하는 것을 목적으로 하고 있으며, 앞선 연구와 유사한 방식으로 공인된 BIM 전문가 시험을 통해 전문지식에 대한 성능을 평가하고자 하였다.

구체적으로 한국BIM학회에서 주관하는 BIM 운용 전문가 시험 중 1급 필기시험 문제를 질의 형식으로 GPT 버전 3.5 및 4에 제공하고 이에 대한 올바른 답변 여부를 분석하고자 하였다.

본 실험을 통해 ChatGPT의 전반적인 성능을 파악하는 동시에 BIM과 관련된 세부 분야 중 취약한 부분에 대해서도 분석할 수 있으며, 이를 통해 GPT의 BIM 가상 보조원 내지 AI 교육 도구로써의 활용 가능 여부를 가늠할 수 있다. 또한 ChatGPT는 fine tuning[6] 및 LangChain[7] 등의 기술로 전문 분야 문서를 학습시킬 수 있는데, 본 연구 결과를 토대로 어느 정도의 추가 도메인 학습이 필요한지 파악하고자 하였다.

1.2 연구의 범위 및 방법

ChatGPT를 BIM 운용 전문가 1급 필기시험에 응시시키기 위해 시험지 3년치를 구하여 총 150 문항을 대상으로 실험을 진행하였다.

본 문항들을 GPT-3.5 및 GPT-4 버전의 ChatGPT에 각각 입력하고 해당 질문에 대한 정답 및 설명을 문의하는 식으로 진행한 후 두 버전의 성능 차이를 비교 분석하였다. 더불어 BIM 운용 전문가 시험은 5가지 세부 유형으로 구분되는데 이에 대한 유형별 정답률을 파악하였으며 추가로 문항을 단답형/서술형으로 구분하여 이에 대한 정답률 차이 여부를 파악하였다.

결과를 통계적으로 정리하여 1) 전체 결과에 대한 GPT 두 버전의 성능 비교, 2) 5개 세부 유형별 성능 비교, 및 3) 단답형/서술형에 대한 성능 비교를 진행하였다.

본 결과를 토대로 취약 부분에 대한 보완 방법 및 ChatGPT의 BIM 관련 정보 탐색 및 BIM 모델링 등과 같은 향후 활용 가능한 분야에 대해 제시하였다.

2. 연구 배경

2.1 BIM 운용 전문가 1급 시험

정부의 BIM 활성화 추진으로 인해 BIM 전문가 및 실무와 관련된 교육의 필요성이 증가함에 따라 관련 분야의 전문성과 실무 능력을 객관적으로 인증할 수 있는 자격증의 필요성이 대두되었다. 이에 한국 BIM 평가원은 한국 BIM 학회와 한솔아카데미 공동 주관 하에 BIM 운용 전문가 자격시험을 개설하였다. 해당 시험은 매년 3회 시행되며, 1차 필기시험과 2차 실기시험으로 구성되어 있다. 자격 등급은 1급, 2급, 3급으로 분류되어 있어 직무 내용에 따른 다양한 전문성을 평가하고 인증할 수 있다.

이 중 1급 BIM 전문가는 토목 및 건축 프로젝트의 BIM 업무 수행 계획을 수립하고 이에 따라 단계별, 분야별 업무 프로세스 설정, 정보의 구축 및 교환 정의, 참여자 간의 협업 및 커뮤니케이션을 통하여 프로젝트 협업 전반을 조율하며 관리하는 수행 능력을 갖춘 전문가를 의미한다. 해당 자격을 취득할 수 있는 BIM 운용 전문가 1급 시험은 필기 및 실기시험으로 구분되어 있으며, 1차 필기시험은 객관식과 주관식이 포함된 50문항으로 구성되어 있고 2차 실기시험은 면접 구술형으로 진행된다. 필기와 실기시험 모두 100점 만점 중 60점 이상 득점해야 합격이다.

본 연구에서는 1급 필기시험을 대상으로 진행되었으며, 문항 유형은 공종별로는 토목, 건축, 공통, 그리고 BIM 분야별로는 'BIM 일반사항', 'BIM 활용', 'BIM 환경 구축', 'BIM 소프트웨어' 및 '스마트 기술' 등의 5가지의 세부 분야로 나눠진다.

2.2 거대 언어 모델 GPT

GPT는 OpenAI에서 개발한 자연어 처리(Natural Language Processing, NLP)를 위한 거대 언어 모델(LLM)이다. 이 모델은 인터넷의 다양한 텍스트 데이터를 통해 학습되며, 이를 통해 문맥에 따른 적절한 답변을 생성하거나 문장을 완성하는 등의 작업을 수행할 수 있다. GPT는 GPT-1, GPT-2, GPT-3, GPT-3.5 그리고 GPT-4 등의 여러 버전이 존재하며 각 버전은 모델의 크기와 학습 데이터의 양에 따라 성능과 기능이 차이가 난다[8].

ChatGPT는 GPT 모델을 기반으로 하는 특정 애플리케이션이다. 이는 사용자와 대화를 나누는 것을 주요 목표로 하며, GPT의 모델 구조와 학습 방법을 이용하여 자연스러운 대화를 생성

6) https://platform.openai.com/docs/guides/fine-tuning

7) https://langchain.com/

한다. ChatGPT는 원래 GPT-2를 기반으로 개발되었지만, 이후 버전에서는 GPT-3 또는 GPT-4와 같은 업그레이드된 모델을 사용하기도 한다.

GPT가 일반적인 언어 생성 엔진이라면, ChatGPT는 그 엔진을 이용하여 개발된 특정한 대화형 애플리케이션으로 구분할 수 있다.

2.2.1 ChatGPT의 학습 및 커스터마이징 방법

ChatGPT는 GPT-3.5 버전을 대화형 에이전트로 만들기 위해 Figure 1에서와 같이 지도형 미세조정(Supervised Fine Tuning), 보상 모델(Reward Model), 및 강화학습(Proximal Policy Optimization, PPO) 훈련을 진행하였으며, 이러한 일련의 과정을 Reinforcement Learning by Human Feedback (RLHF)라고 한다. 최근에는 본 과정을 GPT-4에도 적용하여 가장 최신 버전으로 제공하고 있다(OpenAI, 2023).

ChatGPT는 현재 2021년 9월 이전의 웹상 데이터 및 책 코퍼스 데이터로만 학습이 된 상태이며 특정 산업이나 도메인에 특화돼 있지 않다. 따라서 답변의 정확성을 향상시키기 위해 특화된 데이터를 Fine-Tuning, LangChain 등의 기술을 이용해 추가학습이 가능하다. Fine tuning의 경우 OpenAI사에서 제공하는 API end-point를 이용해 질의-답변 쌍 (Prompt-completion)으로 구성된 데이터 세트를 학습시켜 모델을 특화시킬 수 있다. LangChain의 경우 ChatGPT가 학습하지 못한 2021년 9월 이후의 데이터나 사용자가 추가로 학습시키고자 하는 정보가 담긴 문서를 추가학습시켜 특정 작업에 특화된 모델로 사용자화가 가능하다.

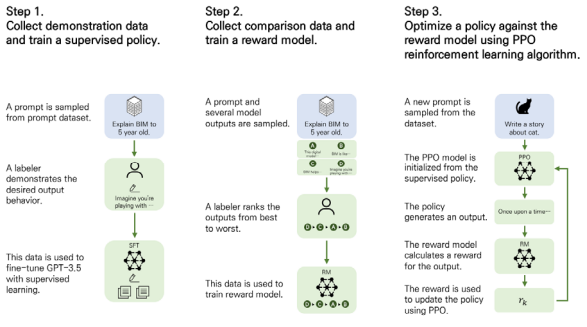

Figure 1. Three steps of method for reinforcement learning (Introducing ChatGPT(2022, November 30) OpenAI: https://openai.com/blog/chatgpt)

또한 ChatGPT는 대화형 에이전트이기 때문에 질문, 즉 Prompt를 어떻게 작성하느냐에 따라 답변이 달라질 수 있다. 그러므로 사전에 일관된 Prompt를 구성하는 게 중요하다. 동일한 질문에도 ChatGPT는 유사하지만 다른 문구의 답변을 제시할 수 있기 때문에 이를 감안해야 하는데 이때 자체 하이퍼파라미터(Hyperparameter)인 temperature, max_tokens 등을 이용해 답변에 대한 무작위성 및 답변 길이를 조정할 수 있다.

2.3 ChatGPT 시험 적용 연구 사례

ChatGPT가 LLM의 실용화를 가져오고 지능적 답변이 가능해지면서 본 모델을 각종 자격증 시험에 응시시켜 성능을 파악해 보는 연구가 최근 등장하였다. 특히 대표적인 전문가 자격증 시험인 변호사 및 의사 자격 시험에 적용한 사례 및 OpenAI가 직접 성능 검사를 실시한 연구가 존재한다.

2.3.1 변호사 자격증 시험

미국의 변호사 자격 시험(이른바 "Bar Exam")에 GPT의 성능을 분석한 연구들이 여럿 존재한다(Choi et al., 2023; Bommarito II and Katz, 2022; OpenAI, 2023)

이 중 Bommarito II and Katz(2022)는 미국의 변호사 자격 시험 중 객관식 시험(Multistate Bar Examination (MBE)) 부분에 GPT-3.5 모델을 적용하여 모델 성능을 평가하였다.

미국의 변호사 자격 시험은 National Conference of Bar Examiners (NCBE)에서 주관하며, 주마다 시험 내용 및 구성이 다르지만, 최근에는 대부분의 주가 일관된 시험인 Uniform Bar Exam(UBE)를 도입하고 있다. 본 시험 형식에서는 1) 객관식, 2) 에세이 및 3) 시나리오 기반 시험으로 구분되며, 본 연구에서 집중한 객관식 시험은 전체 성적의 50%를 차지하고 있다. 객관식 시험은 200개(8개 분야 25문제씩)의 사지선다 문제로 구성된다. 응시자들은 평균적으로 본 시험에서 68%밖에 못 맞춘다고 한다.

본 시험을 GPT가 응시하도록 하되, Prompt에는 여러 가지 응답 조건을 제시하였는데, 대표적으로 1) 정답만 제시 및 2) 3개의 답을 제시하되 이에 대한 ranking을 제시하도록 하였다. 첫 번째 응답 조건의 경우 GPT-3.5는 50.3%의 정답률을 보였으며, 이는 무작위로 답을 선택하는 25%보다는 훨씬 높은 정답률을 보였으나 응시 학생들의 68%에는 미치지 못하였다. 두 번째 조건, 즉 정답의 근접한 정도를 평가할 경우, 본 평균이 88%까지 상승하였으며 이는 GPT가 난해한 문제를 어려워한다는 것을 시사하였다. 또한 8개의 분야 중 전문 법률 용어가 유난히 많은 형사법(Criminal law)에 가장 낮은 점수를 획득해 특정 전문 분야에는 취약한 것으로 드러났다.

2.3.2 국가 의사 자격증 시험

미국 및 일본 의사 자격증 시험에 GPT의 성능을 평가한 연구

8) 각 버전의 매개변수 개수 및 학습데이터 크기: GPT-1 (117M, 5GB); GPT-2(1.5B, 40GB); GPT-3(175B, 600GB); GPT-3.5(175B, 공개 안됨); GPT-4(1T 추정, 공개 안됨).

가 존재한다(Kung et al., 2023; Kasai et al., 2023).

이 중 Kasai et al.(2023)은 일본의 의사국가시험에 ChatGPT, GPT-3 및 GPT-4를 응시시켜 각각의 성능을 비교 평가하였다.

일본의 의사국가시험 중 객관식 시험은 오지선다로써 6개 분야(총 28개 세부 분야)로 나눠지며 분야별로 50-75문제, 총 400문제로 구성된다. 또한 문제들은 유형별로 필수, 일반, 및 금기 문제로 나눠지며, 최저 통과 점수는 필수 문제 80%, 일반 70% 이상이다. 더불어, 금기 문제가 존재하며 이는 3개 이하의 오답만 허락된다. 의료 및 공중 보건과 관련된 문제가 출제되며 이미지가 주어지는 문제도 존재한다. 응시자는 의과대학 졸업생으로서 시험 합격 후 레지던트 지원이 가능해지는데 평균 합격률이 91.7%로서 높은 편이다.

세 개의 GPT 버전을 테스트하기 위해 각각의 LLM API를 활용하였으며, GPT-3는 text-davinci-003(Brown et al., 2020), ChatGPT 는 gpt-3.5-turbo-0301, 그리고 GPT-4는 OpenAI (2023) 전용 API를 활용하였다.

5년치의 시험 문제를 풀도록 하였으며 모두 일어 원문을 그대로 사용한 점이 가장 주목할 만하다. 결과적으로 성능은 GPT-4, ChatGPT, GPT-3 순이었으며, 실제로 5년치 시험을 모두 통과한 모델은 GPT-4뿐이었다. 또한 금기 문제 3개 이하 기준 또한 GPT-4만이 통과하였으나, GPT-4도 실제 수험생의 평균에는 미치지 못했다.

2.3.3 미국 대학 및 대학원 진학 자격 시험

이외에 OpenAI는 자체적으로 미국의 대학 입학시험인 SAT(Scholastic Assessment Test), 대학원 입학시험인 GRE (Graduate Record Examination) 및 고등학교 선행 시험인 AP(Advanced Placement) 교과목 시험에 GPT-3.5와 GPT-4의 성능을 분석하였다. 제반 시험에서 GPT-4가 GPT-3.5보다 우월한 성능을 보였으며, 이미지 판독이 가능한 GPT-4버전이 가장 우수한 성능을 보였다. Figure 2에서와 같이 AP 과목 중 심리학, 미국 역사 등은 90% 이상의 성적을 받은 반면, 수학 시험에 속하는 AP 미적분학 및 AMC(American Mathematics Competitions)에는 저조한 성적을 보였다. 이는 언어 위주의 문제에 GPT가 더 강점을 보이는 것으로 해석된다.

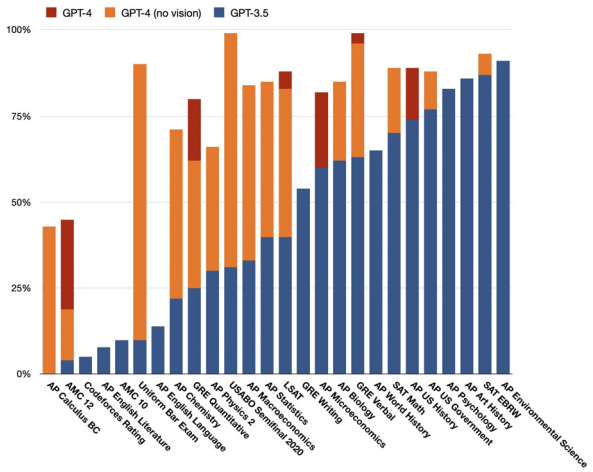

Figure 2. GPT performance on academic and professional exams(OpenAI, 2023)

3. 연구 방법

BIM 운용 전문가 1급 필기시험에 GPT의 성능을 평가하기 위해 다음과 같은 절차 및 범위를 설정하였다.

3.1 Step 1: 시험지 문항 수집 및 문제 유형 정리

BIM 운용 전문가 1급 시험 중 필기시험 3년치 (2020, 2021, 2022년도)에 해당하는 150 문항을 대상으로 하였다. 각 문항은 2장에서 소개된 바와 같이 5가지 유형인 'BIM 일반사항', 'BIM 활용', 'BIM 환경 구축', 'BIM 소프트웨어' 및 '스마트 기술'로 구분이 되며 각 유형에 대한 세부 내용은 Table 1에 소개하였다.

Table 1. Problem categories for the BIM expertise exam

Type	Content
BIM generalities	BIM concepts and purpose, BIM system and policy
BIM applications	3D visualization, Simulation, Engineering analytics, Reviewing volume, etc.
BIM environment establishment	Environment establishment requirements for BIM application in design and construction
BIM software	Software type, UI(User interface), Family, Object oriented and parametric modeling
Smart construction technology	Laser scanning, Drone, AR(Augmented Reality)/VR(Virtual Reality), IoT(Internet of Things), AI, etc.

3년 치 문항 구성을 보면 5개 유형별로는 Figure 3와 같이 'BIM 일반사항'이 45%로 가장 많았고 그다음으로 'BIM 환경 구축'이 27%로 많았다.

본 연구에서는 이와 함께 각 문항을 단답형 내지 서술형으로 구분하여 성능분석에 활용하였다. 단답형은 답이 숫자나 용어 등 단순하고 짧은 형태이며 서술형은 주어진 문제에 대한 설명이나 용어의 의미 등을 자세하게 서술하여 긴 문장으로 구성된 경우이다. Figure 4에 일례로 두 가지 유형을 제시하였다.

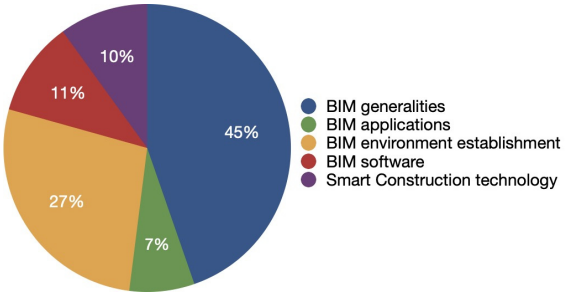

Figure 3. Distribution of questions by type (average)

Example of short-answer question

복수의 3차원 스캐닝을 통해 얻은 데이터를 합칠 때 기준을 삼을 수 있는 target은 하나의 스캔에서 최소한 몇 개가 필요한가?

① 1개 ② 2개 ③ 3개 ④ 5개

Example of essay question

BIM에 있어 상호연동성에 대한 설명으로 옳지 않은 것은?

1) 상호연동성이 필요한 이유는 프로젝트 참여자 간 신속한 정보교환을 통한 의사결정과 프로세스 간 정보 재활용을 위해서이다.
2) 상호연동성은 정확하고 신속한 의사결정을 통해 고객이 원하는 가치를 주기 위한 정보교환 행위라 할 수 있다.
3) 데이터를 운용하는 시스템 모델이 서로 상이해도 상호 연동성의 문제는 발생하지 않는다.
4) BIM에서 상호 연동성을 위해 주로 사용되는 모델링 언어는 EXPRESS와 UML언어이다.

Figure 4. Example of short-answer/essay question

본 기준으로 나눌 경우 Figure 5에서와 같이 150문항은 62%가 단답형, 38%가 서술형으로 구분되었다.

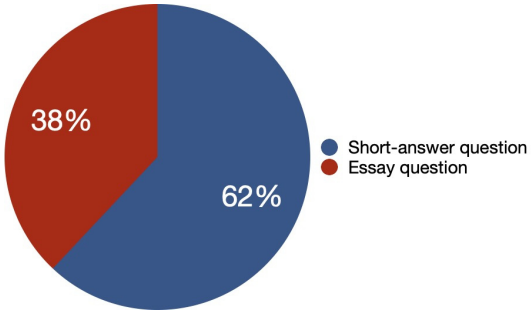

Figure 5. Percentage of short-answer/essay question (average)

3.2 Step 2: 문항 입력 및 답변 출력

150 문항을 가장 최근 버전인 GPT-3.5 및 GPT-4에 입력하고 정답을 제시하도록 하였다. 2장의 선행연구에서 보듯이 GPT-3은 두 최신 버전보다 성능이 낮게 나온 것으로 파악되어 본 연구에서는 제외시켰다.

Prompt에는 해당 질문을 각 GPT 버전에 입력한 후, 이와 함께 "본 질문에 가장 올바른 답변을 골라주고 그 이유를 간단하게 설명해줘"라고 지시하였다. 2장에서 언급한 대로 성능의 향상을 위해 Hyperparameter 조정 및 Prompt-completion 미세조정을 실시할 수 있었으나 본 연구에서는 이러한 특화 작업 이전 ChatGPT 고유의 성능을 평가하기 위해 원기능 그대로 시험을 진행하였다.

3.3 Step 3: 결과 분석

결과를 통계적으로 정리하여 1) 전체 결과에 대한 GPT 두 버전의 성능 비교, 2) 5개 세부 유형별 성능 비교, 및 3) 단답형/서술형에 대한 성능 비교를 진행하였다.

4. 연구 결과

4.1 전체 결과

Table 2 및 Figure 6에 3년 치 점수와 정답률을 제시하였다. GPT-3.5 및 GPT-4 는 3년 평균으로 각각 68 및 85점을 획득하여 최저 점수인 60점을 통과하였다. GPT-4가 17점을 더 획득하였으며 각 해 점수에서도 GPT-3.5보다 우세한 성적을 거두었다.

Table 2. ChatGPT results by year

Model	2020		2021		2022		3-year average	
	# of correct answers	Score	# of correct answers	Score	# of correct answers	Score	# of correct answers	Score
GPT-3.5	37	74	32	64	33	66	34	68
GPT-4	44	88	42	84	42	84	42.7	85
Total	50	100	50	100	50	100	50	100
Passing Score	30	60	30	60	30	60	30	60

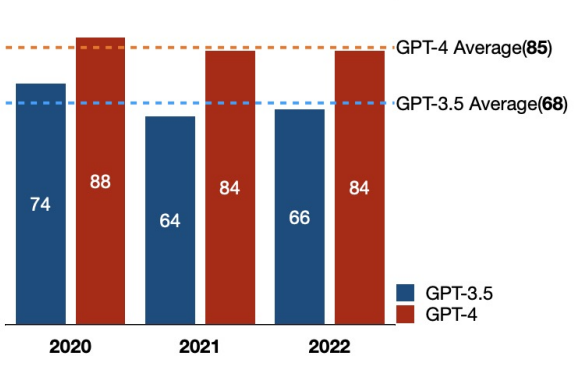

Figure 6. ChatGPT results by year

4.2 5가지 세부 유형별 결과

5가지 유형별 분석 결과 Figure 7에서와 같이 GPT-3.5 대비 GPT-4의 성능이 'BIM 활용' 분야에서는 동등하고 나머지 분야에서는 모두 월등한 것으로 나타났다. 특히 'BIM 소프트웨어' 부분에서 GPT-3.5 대비 36%로 가장 높은 정답률 차이를 보였다.

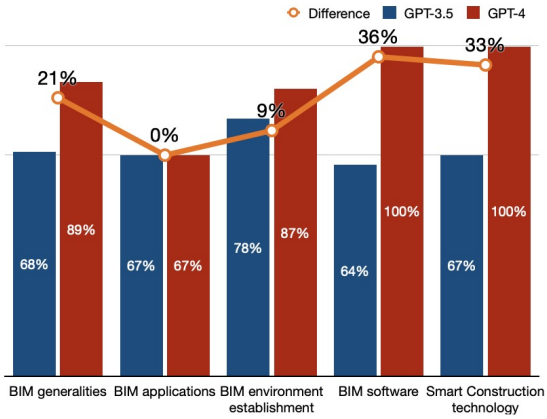

Figure 7. Performance of ChatGPT by question type

GPT-4의 결과만 볼 경우, 'BIM 소프트웨어' 및 '스마트 기술'의 경우 모두 정답을 맞춰 가장 높은 성능을 보였다. 'BIM 소프트웨어' 유형은 개별 BIM 소프트웨어의 기능 및 사용 방법을 묻는 문항들로서 실무에서 특히 필요한 지식들이므로 BIM 모델링 작업 보조에 GPT-4가 적정한 것으로 판단된다. 그 다음으로는 'BIM 일반사항' (89%) 및 'BIM 환경 구축'(87%)으로, 문제 중 국내 기준 및 표준에 관한 문항에서 오답들이 존재하였으며, 국내 정책 및 규정에 관련된 문항에 대해서는 명확한 설명을 하지 못하였다.

GPT-4가 가장 취약한 부분은 'BIM 활용'(67%) 유형이었다. 해당 유형의 문제에 대해 오답의 비율이 가장 높을 뿐 아니라 복수의 답을 도출하는 등 해당 질문에 어려움을 겪는 것이 확인되었다. 'BIM 활용'은 BIM 기반으로 가능한 다양한 사업 관리 작업을 포함하는 유형으로서 광범위하면서도 절차를 규정하기 힘든 부분이 있어 이런 문제가 발생한 것으로 판단된다.

Figure 8에 GPT-4가 강점을 보인 'BIM 소프트웨어' 및 '스마트 기술' 문항, 그리고 Figure 9에서는 'BIM 활용' 문제에서 오답의 예시를 제시하였다.

Example of 'BIM software'

일반적으로 BIM 사업의 CDE 시스템에서 준공 시 저장되는 데이터는 어떤 상태인가?

① WIP ② Approved ③ Published ④ Authorized

Example of 'Smart Construction technology'

다음 중 드론 촬영이나 3D스캐너를 이용한 데이터 활용 방법으로 가장 적절하지 않은 것은?

1) Cloud Point Data를 이용한 역설계
2) 광범위한 지역의 정사사진 작성
3) 대지 현황 측량 및 토공량 산출
4) 작업자들의 위치 실시간 파악
5) 시공오차 검토

Correct answer Answer of ChatGPT

Figure 8. Example of ChatGPT correct answers

Examples of 'BIM applications'

Example 1

공정관리를 위한 4D 시뮬레이션 절차를 올바른 순서로 나열한 것을 고르시오

1) 3차원 모델링 시 시공순서 및 분할 → 자동 연결 → 3차원 모델 적용 → 표준 코드 정의 → 스케줄 작성 및 적용 → 공정협의 → 4D 시뮬레이션 광범위한 지역의 정사사진 작성
2) 3차원 모델링 시 시공순서 및 분할 → 표준 코드 정의 → 3차원 모델 적용 → 공정협의 → 자동 연결 → 스케줄 작성 및 적용 → 4D 시뮬레이션
3) 3차원 모델링 시 시공순서 및 분할 → 3차원 모델 적용 → 스케줄 작성 및 적용 → 공정협의 → 표준 코드 정의 → 자동 연결 → 4D 시뮬레이션
4) 3차원 모델링 시 시공순서 및 분할 → 공정협의 → 표준 코드 정의 → 3차원 모델 적용 → 스케줄 작성 및 적용 → 자동 연결 → 4D 시뮬레이션

Example 2

스마트설계지침에서 BIM 지형데이터의 상세 설계 단계에서 필요한 정밀도 수준은?

1) 1:5000
2) 1:2000
3) 1:1000
4) 1:500

Figure 9. Examples of ChatGPT incorrect answers

4.3 단답형/서술형 유형별 결과

Figure 10에서와 같이 GPT-3.5는 단답형, 서술형 정답률에 차이를 보이지 않았으나, GPT-4는 단답형에서 상대적으로 11%

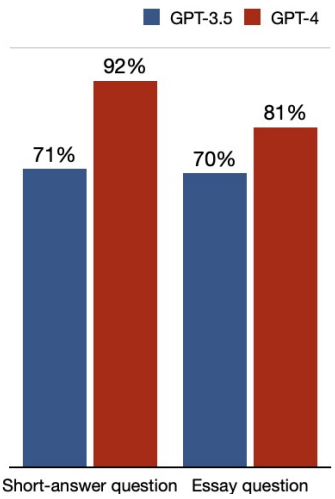

Figure 10. ChatGPT performance with respect to short-answer vs. essay questions

더 나은 성능을 보여 단답형에서 월등한 성능을 보였다. 일례로, Figure 11에서와 같이 전환설계 BIM 수행 시 담당자별 역할을 묻는 질문에서 GPT-3.5는 1)번을 오답으로 답변한 반면, GPT-4는 정답인 4)번을 선택하였다.

서술형 문제 중 GPT-4의 오답을 보면, Figure 11에서와 같이 IPD(Integrated Project Delivery) 수행 효과 질문에서 3)번 대신 1)번을 오답으로 처리하였다. IPD의 효과가 사업참여자간 투명성을 제공해 불확실성을 줄여주지만, 이는 꼭 해당 사업에서 분쟁이나 리스크를 줄여주는 보장은 없다. 그러므로 이는 GPT-4가 자의적인 추론을 한 것으로 볼 수 있어, 본 사항은 유의해야 하는 부분으로 판단된다.

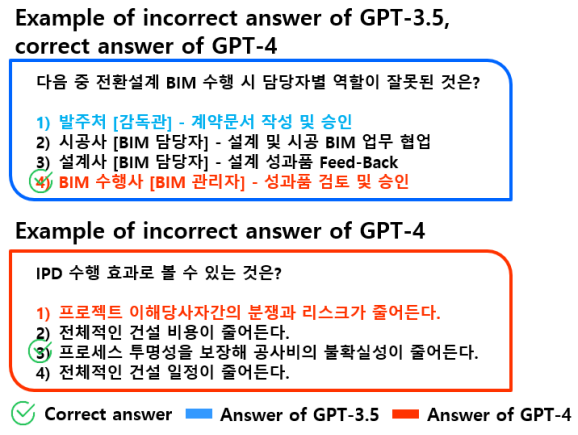

Figure 11. Example of ChatGPT correct/incorrect answers

4.4 추가 오류 사례

상기 예시들과는 별도로 GPT-4가 취약한 문제들은 정책과 제도에 관한 부분이었다. 특히 국제적으로 통용되는 표준 정보에 대해서는 정확하지만, 국내 한정 자료에는 오답을 제시하였다. 일례로 MasterFormat, Uniclass와 같이 국제적으로 활용되는 정보에 대한 답변은 정확하지만, 국내 국토교통부, 도로공사의 BIM 관련 정보에는 부정확한 경우가 다수 있었다. 대부분의 국내 자료를 문서로 제공하고 있어 인터넷 텍스트 데이터를 기반으로 학습하는 GPT의 경우 검색으로 취득할 수 있는 정보를 제외한 문서 내 존재하는 정보 취득의 한계가 존재하여 이러한 문제가 발생한 것으로 판단된다.

또한 국내 설계지침, 가이드라인과 같이 재개정된 최근 정보에 대한 질의에도 취약한 것으로 드러났다. 이는 GPT-4가 2021년 9월 기준 이전의 정보에 학습한 점과 국내에 한정된 정보에는 접근이 어려워서 발생한 것으로 판단된다.

실제 상기 문제들을 2021년 12월 까지의 정보로 학습한 구글의 LLM인 Bard에 제시했을 경우 정답을 맞추는 것으로 드러나, 정보의 접근 제한과 관련된 문제인 것으로 파악되었다.

5. 시사점

5.1 GPT-4가 전체 및 세부 유형별로도 성능 우세

상기 연구 결과를 요약하면, 우선 두 버전 모두 BIM 운용 전문가 1급 필기시험은 최저 점수인 60점을 통과하였다. 3년치 평균 기준 GPT-3.5 및 GPT-4 각각 68점, 85점을 기록하여 GPT-4가 상대적으로 우수한 성능을 보였다. 이는 GPT-4가 이전 버전에 비해 거의 여섯 배 (175B 대 ITB)에 달하는 매개변수가 존재하는 것으로 추정되며 공개는 되지 않았으나 더 방대한 데이터 기반으로 학습된 데서 기인한 것으로 보인다.

5.2 BIM 분야 중 취약 부분에 대한 추가학습 필요

GPT-4는 BIM과 관련된 국내 정책, 제도 및 지침 등에 대한 질문에 취약한 것으로 드러났다. 이들 정보 및 관련 문서는 접근성이 제한적이라 어찌 보면 당연한 결과이다.

이에 대한 보완은 fine tuning 및 LangChain 기술과 같이 ChatGPT를 커스터마이징할 수 있는 기법을 통해 로 해결이 가능할 것으로 예상된다.

- Fine tuning: ChatGPT는 2021년 9월까지의 데이터로만 학습되어 있으며 기본적으로 'few shot' learning에 의해 답변을 하게 된다. 이에 대한 보완으로 Prompt-completion 형식의 학습 데이터를 추가하여 답변을 특성화할 수 있으며 이때 최소 100-200개의 데이터셋이 필요하다. Fine tuning은 특히 분류 작업 및 조건부 텍스트 생성에 유효한 것으로 알려졌다. 단, 현재 GPT-3 모델인 text-davinci-003(Brown et al., 2020)에만 적용 가능한 제약 있다.
- LangChain: ChatGPT API와 함께 활용하여 외부 데이터 또는 특정 파일을 ChatGPT가 직접 활용하여 답변 가능케 하는 오픈소스 프레임워크이다. 주어진 문서를 Vector-Store 형태로 변환하고 Embedding 기술을 통해 내용을 판독하여 이를 기반으로 답변하게 된다. 따라서, 국토교통부의 BIM 기본 및 시행 지침 및 발주처별 적용 지침 또는 BIM 소프트웨어 사용 설명서를 제공하여 답변을 특성화할 수 있을 것으로 기대된다.

5.3 활용 방안

ChatGPT가 기본적인 BIM 전문지식은 갖춘 것으로 평가되었으며, 취약 부분에 따라 추가학습이 가능하다는 전제하에 다음과 같은 분야에서 활용이 가능할 것으로 보인다.

- BIM 가상 튜터: 최근 ChatGPT를 가상 튜터로 개발한 사례들이 등장하고 있으며 대표적으로 Khan Academy의

Khanmigo[9]를 들 수 있다. 본 기업은 미국 초중고 온라인 교육 기업으로서 각광을 받고 있는데 당사 플랫폼에 ChatGPT를 탑재하여 학생과 대화형으로 문제를 풀어나가는 방안을 제시하고 있다. 본 모듈을 통해 즉각 답을 제시하지 않고 학생의 문제 풀이를 진행하면서 정답을 유도하는 방식이다.

이처럼 BIM 관련 학습 또는 실무에서 BIM 모델링 진행 시 가상 보조원 역할을 할 수 있을 것으로 내다본다.

- BIM 모델 정보 탐색: Zheng and Fischer(2023)는 클라우드 DB 형태로 저장된 BIM 모델에 질의를 순차적으로 설계하여 BIM 모델의 정보를 탐색(Information Retrieval, IR)하는 가상 보조(VA) 기술 제안하였고 이를 BIM-GPT라 명명하였다. 기존의 BIM IR은 BIMQL(Mazairac and Beetz, 2013) 또는 IfcOWL(Beetz et al., 2009)과 같은 특정 DB를 별도로 선 구축한 후 가능해졌는데 본 방식은 이러한 전처리 단계를 최소화할 수 있어 활용도가 높은 것으로 기대된다.

텍스트 기반 BIM 자동 모델링: Hypar사는 텍스트 입력 기반으로 BIM 모델을 자동 작성하는 이른바 'text-to-BIM[10]' 기술을 시연한 바 있다. 모델링 작업 시 요구되는 수작업을 줄일 수 있는데 ChatGPT에 개별 소프트웨어 설명서를 학습시키면 가능성이 있을 것으로 사료된다.

6. 결론

본 연구에서는 BIM 분야에서 업무 보조 도구로서 GPT의 활용 가능성을 평가하기 위해 BIM 운용전문가 1급 필기시험을 토대로 성능 평가를 실시하였다. 그 결과 GPT-3.5 및 GPT-4 모두 최저 합격 점수인 60점을 통과하였다. GPT-4는 3년치 시험에 대해 평균 85점을 획득하여 상대적으로 우수한 모델로 평가되었다.

5개 세부 유형에서 GPT-4는 'BIM 소프트웨어' 및 '스마트 기술'에서는 만점을 얻은 반면, 'BIM 활용' 분야에서는 가장 낮은 67점을 획득하는 데 그쳤다. GPT-4가 단편적 지식에는 강점이 있으나, 복합적 실무 지식에는 아직 취약한 것으로 드러났다. 이 밖에도 접근성이 어려운 국내 BIM 정책 및 제도와 관련된 질의에도 오답률이 높은 것으로 파악되었다.

본 연구에서는 ChatGPT의 BIM 전문지식에 대한 성능 평가를 위해 대표적인 국내 자격증 시험을 활용하고 이를 정량적으로 파악하였다는 데 의의를 둔다. 기존 유사 연구에서는 응시생들의 평균 정답률과 비교를 하였으나 본 연구에서는 자료의 제약으로 이를 수행하지 못하였다. 또한 두 버전의 성능 차이의 통계적 유의미까지 분석하지는 못하였다.

향후에는 이를 위한 추가 연구를 진행할 계획이며, 더불어 LangChain 등의 기술을 적용하여 BIM 분야에 특화된 GPT 모델을 구축할 계획이다.

감사의 글

본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1A2C1100741).

References

Beetz, J., Van Leeuwen, J., De Vries, B. (2009). IfcOWL : A Case of Transforming EXPRESS Schemas into Ontologies. Ai Edam, 23(1), pp. 89-101.

Bommarito II, M., Katz, D. M. (2022). GPT Takes the Bar Exam. arXivpreprint arXiv:2212.14402.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P. Sastry, G. Askell, A., Agarwal, S., Herbert-Voss, A. Krueger, G, Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020).Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, pp. 1877-1901.

Choi, J. H., Hickman, K. E., Monahan, A., Schwarcz, D. (2023). ChatGPT Goes to Law School. Available at SSRN.

Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y., Radev, D. (2023). Evaluating GPT-4 and ChatGPT on Japanese-Medical Licensing Examinations. JMIR Medical Education, 9.

KIBIM (2023). A study on countermeasures for construction engineering in accordance with the digitalization of the construction industry, Final Report.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon,

9) https://openai.com/customer-stories/khan-academy

10) https://aecmag.com/ai/hypar-text-to-bim-and-beyond

L., Elepano, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted Medical Education Using Large Language Models. PLOS Digital Health, 2(2).

Mazairac, W., Beetz, J. (2013). BIMQL-An Open Query language for Building Information Models. Advanced Engineering Informatics, 27(4), pp. 444-456.

OpenAI (2023). GPT-4 Technical Report. ArXiv, abs/2303.08774.

Seo, M. B., Ju, K. B. (2012). Strategies to Revitalize BIM (Building Information Modeling) by the Survey Question-Naires from Design Experts in Field of Civil Engineering. Journal of the Korea Contents Association, 12(11), pp. 446-457.

Singh, A. K., Pal, A., Kumar, P., Lin, J. J., Hsieh, S. H. (2023). Prospects of Integrating BIM and NLP for Automatic Construction Schedule Management. ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction. IAARC Publications, 40, pp. 238-245.

Zheng, J., Fischer, M. (2023). BIM-GPT: A Prompt-based Virtual Assistant Framework for BIM Information Retri-eval. arXiv preprint, DOI: 10.48550/arXiv.2304.09333