개념 학습을 위한 LLM 기반 챗봇과의 질의문답 분석: 학업 성취도를 중심으로

Analysis of Question-and-Answer Interactions with LLM-based Chatbots in Concept Learning: an Academic Performance Perspective

양연선* 신아현* 송진영
Yeonsun Yang Ahyeon Shin Jean Y. Song
디지스트 디지스트 연세대학교
DGIST DGIST Yonsei University
diddustjs98@dgist.ac.kr ahyeon@dgist.ac.kr jeansong@yonsei.ac.kr

* Equal contribution

요약문

본 연구에서는 거대언어모델(LLM) 기반 챗봇을 개념 학습 맥락에서 사용할 때 나타나는 질문 패턴이 학습자의 학업 성취도와 어떤 관계를 보이는지 살펴보았다. 선행 연구를 통해 질의의 양과 다양성이 학업 능력과 상관 관계가 있을 것으로 가정하여 연구 가설을 설정하였다. 이를 검증하기 위해, 대학 기관의 '프로그래밍 입문' 수업에서 152명의 수강생을 대상으로 ChatGPT와의 질의응답 데이터 1391쌍과 시험 성적을 수집하였다. 질문의 양과 다양성 두 측면에서 학업 성취도와의 상관 관계를 분석한 결과, 학업 성취도와 질문의 개수 및 다양성 간에 통계적으로 유의미한 연관이 있음을 관찰했다. 이는 학업 성적이 높은 학습자들이 개념을 이해하기 위해 LLM 기반 챗봇과 더 많은 상호작용을 하며, 의미적으로 다양한 질문을 요청하는 과정을 통해 여러 관점에서 지식을 습득해 나감을 시사한다. 본 연구는 LLM을 활용한 학습에서 질의응답 방식이 학업 성과에 연관이 있음을 보여주며, 향후 LLM 기반 챗봇 인터페이스 개발 시 개인화된 학습을 위한 질의 가이드라인 도출의 중요성을 논의한다.

주제어

LLM 챗봇, 학습보조도구, 질의응답, 질의패턴, 개념학습, 학업 성취도

1 서론

거대언어모델(LLM; Large Language Model)의 발전으로 대화형 언어모델 기반 시스템이 다양한 분야에서 범용적으로 활용되고 있으며, 최근 교육 분야에서도 자료 검색, 내용 요약, 문제 생성, 에세이 작성, 프로그래밍 등의 작업에 적극적으로 도입되고 있다[1,2]. 이 중 많은 시스템들은 학습자들이 수업이나 과제 수행 중 마주하는 어려움과 의문 해소를 돕기 위해, LLM 과의 질의응답 방식 상호작용을 기반으로 개념을 이해하고 지식을 습득하는 것을 지원하고자 한다.

학습자 질의에 연관된 자료를 검색, 수집, 정리하는 과정은 LLM을 통해 자동적으로 수행될 수 있지만, 제공된 응답을 바탕으로 개념을 이해하고, 스스로 지식의 공백을 인식하며, 이를 정리된 문장으로 표현해 다음 질문을 형성하는 질의 생성(query formulation)은 여전히 학습자 주도적으로 이루어지는 과정이다. 이러한 과정은 교육학적으로 중요한 학습 역량으로 평가되며[3], 특히 LLM이 적극적으로 도입되고 있는 현대 교육 환경에서는 프롬프트 엔지니어링(prompt engineering)과 연계되어 더욱 필수적인 능력으로 강조되고 있다.

이에 기존 연구에서는 학습자와 LLM 간의 대화에서 나타나는 일반적인 상호작용 패턴과 질문 유형을 탐구하였지만[4,5,6], 이를 학습자의 학업 능력 관점에서 실증적으로 분석한 연구는 수행되지 않았다. 본 연구는 학습자의 학업 성취도가 질의 생성과 연관된 주요한 요소로 작용할 수 있다는 점에 주목하여, 대학

기관의 '프로그래밍 입문' 수업에서 수강생들을 대상으로 수집한 GPT-3.5 기반 ChatGPT[7]와의 대화 데이터 및 학업 성적을 분석하였다. 분석 결과, 학업 성취도와 질문 패턴 간에 유의미한 상관관계가 있음을 발견하였다.

2 관련 연구

2.1 LLM 챗봇의 학습효과와 질의응답 상호작용 분석

학습자의 지식 습득을 돕기 위한 LLM 기반 대화형 학습 도구가 활발히 개발됨에 따라, 학습자의 LLM 사용 방식을 심도 있게 이해하기 위해 질의응답 패턴을 분석하는 여러 연구들이 수행되고 있다. 예를 들어, Kazemitabaar(2023)은 프로그래밍 학습 맥락에서 학습자가 LLM에 요청하는 질문 유형과 프롬프트 작성 방식을 분석하고, 이러한 상호작용 패턴이 학습 성과에 미치는 영향을 탐구하였다[4]. 또한 Jin(2024)과 Han(2024)은 학생들이 LLM에 요청하는 질문을 분석해 분류체계를 제시하고 이를 정량적으로 평가하였다[5,6].

한편, 최근 연구들은 LLM 기반 학습 효과가 학습자의 선행지식, 문해력 등 개인의 인지 능력 및 학업 능력과 밀접한 상관관계가 있음을 보고하고 있다. Kazemitabaar(2023)은 프로그래밍에 대한 사전지식이 LLM 챗봇 사용에 미치는 영향을 탐구하기 위해, 상위 그룹과 하위 그룹으로 나누어 학습결과를 비교분석 하였다[8]. Otis(2023)은 개인의 능력 수준에 따라 LLM 챗봇 활용 능력에 유의미한 차이가 있음을 발견하였다[9].

이러한 연구들은 LLM 챗봇이 효과적인 학습성과와 개인화된 학습 경험을 제공하기 위한 실험적 기초를 마련한다. 그러나 기존 연구들은 상호작용 패턴 또는 학업 능력 중 한 가지 관점에만 초점을 맞추어 진행되었으며, 두 관점을 종합하여 질의 패턴을 학업 능력의 관점에서 심층적으로 분석한 연구는 부족하다. 이에 본 연구는 개념 학습을 위한 학습자의 LLM 챗봇 질의 데이터를 분석하고, 학업 성취도와의 상관 관계를 탐구하고자 한다.

2.2 교육 분야에서의 질의 생성의 중요성

질문은 교육에서 학습 내용을 깊이 이해하고, 더 나아가 사고를 확장하는 데 핵심적인 역할을 한다. 학습자들은 질문을 스스로 생성해내는 과정에서 자신이 이해한 것과 그러지 못한 것을 인지하고, 이해가 부족하다고 느끼는 내용을 보완하며 성장해나간다. 이는 학업 능력으로도 이어져 학업 성취도가 높은 학습자일수록 더 구체적이고 체계화된 질문을 생성하는 경향을 보인다[3]. 이에 따라 학습자의 질의능력을 체계적으로 평가하기 위한 연구가 활발히 이루어져 왔다.

Zhang(2011)은 질의의 양과 길이를 평가하여 이를 학습자의 숙련도를 예측하는 데 활용할 수 있음을 보였다[10]. 또한 White(2009)는 학업능력이 높은 학습자일수록 지식 습득을 위해 더 다양하고 복잡한 질문을 던지는 경향이 있음을 발견하고[11], 이에 Ide(2021)는 질의 다양성 평가를 위해 코사인 유사도(cosine similarity)를 측정하여 문장 간 유사도를 분석하였다[12]. 그러나 이러한 연구들은 웹 기반 학습 맥락에서 수행된 연구로, LLM 챗봇과의 상호작용에서 일반화되기 위해서는 추가적인 분석 및 논의가 필요하다.

따라서 본 연구에서는 LLM을 통한 학습 상황에서 질의 능력과 학업 능력의 상관 관계를 확인하기 위해 아래와 같은 가설을 설정하였다.

가설1(H_1). 학업 성취도가 높은 학습자일수록 질문의 양적인 측면에서 LLM 챗봇에 요청한 질의의 개수와 질의의 총 길이의 값이 클 것이다.

가설2(H_2). 학업 성취도가 높은 학습자일수록 질문의 다양성 측면에서 LLM 챗봇에 요청한 질의 간 의미적 유사도가 적을 것이다.

3 연구방법

3.1 연구 데이터

본 연구는 '프로그래밍 입문 (BE101a)' 교과목에서 수행된 과제의 제출물을 연구 데이터로 사용하였으며, 기관 내 IRB 승인을 받은 후 진행되었다 (DGISTIRB-202402-001). 과제는 수업에서 다룬 합병정렬(Merge Sort)의 개념을 GPT-3.5 기반 ChatGPT를 사용하여 이해하도록 설계되었으며, 과제 수행 과정에서 생성된 모든 ChatGPT 대화 데이터와 참가자들의 중간고사 및 기말고사 성적이 수집되었다. 수집된 대화 데이터 중 참가자들이 직접 작성한 질의가 아닌 코드나 과제의 지시사항을 그대로 복사해 붙여넣은 데이터는 질문 길이 계산 시 편차를 유발할 수 있다. 따라서 분석의 정확성을 위해 복사-붙여넣기된 문자는 전처리 과정에서 제거하였다.

3.2 연구 대상

연구 대상자는 '프로그래밍 입문' 교과목을 수강하는 원내 대학생 152명 (여:51명, 남:101명, 나이: M=19, SD=0.96)이며, 연구 내용을 이해하고 자발적으로 동의한 대상자만 연구에 포함되었다.

3.3 분석 방법

본 연구에서 수집한 데이터는 대학생 152명의 학업성취도 (중간고사 및 기말고사 점수)와 ChatGPT와의 질의응답 총 1391 쌍으로 구성된다. 학업성취도는 중간고사와 기말고사 점수의 총점을 기준으로 계산한 백분위 (percentile, 상위 0~100%)로 나타내었다.

연구 가설에 따라 참가자들의 질의 패턴을 분석하기 위해 세 가지 기준을 적용하였다. 질문의 양과 관련된 분석을 위해서 1) 질문의 총 개수, 2) 질문의 총 글자 수를 계산하였다. 질문의 다양성에 초점을 두어 분석하기 위해서 3) 모든 질문 쌍에 대해 코사인 유사도를 계산한 뒤 평균값을 산출하였다. 코사인 유사도[16]는 두 벡터 간의 방향적 유사성을 측정하는 방법으로, 이를 통해 질문 간 내용적 유사성을 수치화할 수 있다. 값이 1에 가까울수록 유사도가 높음을 의미하며, 0에 가까울수록 상이함을 의미한다. 구체적인 통계는 표 1에 나타내었다. 최종적으로 질의 패턴과 학업 성취도 간의 상관관계를 평가하기 위해 성적 백분위와 위 세 항목 간의 Pearson 상관분석을 수행하였다.

	평균	표준 편차	최소값	최대값
백분위	48.20	28.67	0.5	99
총 개수	9.15	5.83	0	31
총 글자수	801.66	716.74	32	3873
평균 코사인 유사도	0.15	0.06	0.04	0.34

표 1. 실험자들의 백분위, 질문의 총 개수, 질문의 총 글자 수, 질문 간 평균 코사인 유사도

4 결과

	Pearson's r	p
백분위-총 개수	−0.165	0.043*
백분위-총 글자수	−0.071	0.421
백분위-평균 코사인 유사도	0.172	0.049*

표 2. 학업성취도(백분위)와 질의패턴(개수, 글자수, 코사인 유사도) 간의 Pearson 상관분석 (*$p<0.05$)

4.1 질문의 양과 학업 성취도 간의 상관 관계 분석

질문의 양과 학업 성취도 간의 상관 관계에 대한 가설 H_1에 대한 분석 결과, 표 2에 제시된 바와 같이 성적 백분위와 질문의 총 개수 간 약한 상관 관계가 관찰되었으며 이는 통계적으로 유의미했다 ($r=-0.165$, $p<0.05$). 그러나 백분위와 질문의 총 글자수 간 상관분석을 수행한 결과, 통계적으로 유의미한 연관성이 보이지 않았다 ($p>0.05$). 즉, 개념 학습을 위해 LLM 챗봇과의 문답 시 작성하는 질문의 개수는 학습자의 학업능력과 연관이 있지만 질문의 길이와의 연관은 밝히지 못했으므로 이는 초기 연구가설 H_1에 일부 부합하였다. 그림 1은 백분위와 질문의 개수 간 상관 관계를 시각화한 산점도와 회귀선이다.

상관분석 시 총 글자수는 학업 성취도와 연관성이 관찰되지 않았지만, 표 1에서 학습자 간 편차가 상대적으로 크게 나타난 요소였다. 이는 일부 학습자의 경우 포괄적인 질문을 작성하는 반면, 구체적인 예시를 포함하거나 스스로 공부한 내용을 정리하여 올바르게 이해했는지 반복적으로 확인을 요청하는 등 학습자 간 질문 작성 전략이 다르기 때문으로 보인다. 이러한 차이가 총 글자수의 개인 간 편차에 큰 영향을 미친 것으로 판단된다.

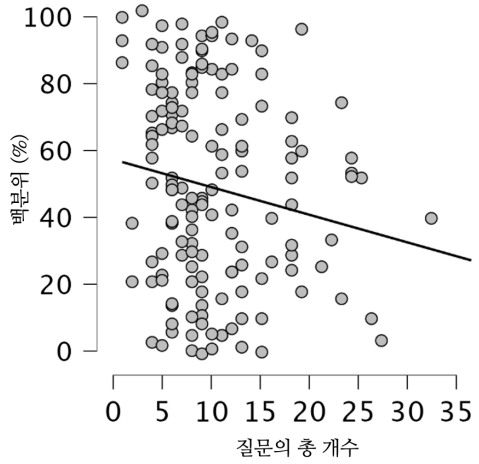

그림1. 학업성취도(백분위)와 실험자의 질문 총 개수 간 상관관계를 나타낸 산점도

4.2 질문의 다양성과 학업 성취도 간의 상관 관계 분석

질문의 다양성과 학업 성취도 간의 상관 관계에 대한 가설 H_2에 대한 분석 결과, 표 2에 제시된 바와 같이 성적 백분위와 질문 간 평균 코사인 유사도 간에 약한 상관 관계가 관찰되었으며 이는 통계적으로 유의미했다 ($r=0.172$, $p<0.05$). 즉, 학업 능력이 높은 학습자일수록 LLM 챗봇에 요청하는 질문의 의미적 다양성이 더 큰 것으로 보아, 연구가설 H_2에 부합하는 결과이다. 그림 2는 백분위와 질의의 평균 코사인 유사도 간 상관 관계를 시각화한 산점도와 회귀선이다.

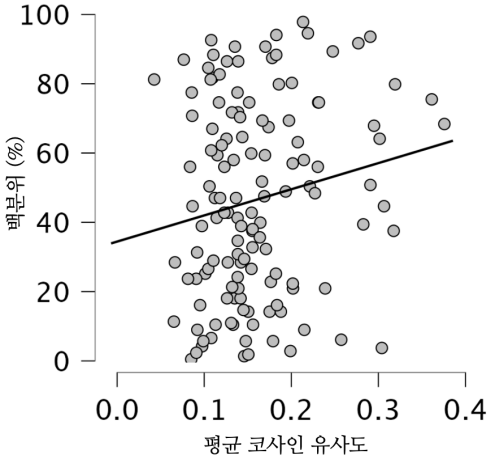

그림2. 학업성취도(백분위)와 실험자의 질문 평균 코사인 유사도 간의 상관관계를 나타낸 산점도

5 결론 및 논의

본 연구의 목적은 개념 학습 과정에서 LLM 기반 챗봇과의 질의응답 방식 및 패턴이 학습자의 학업 성취도와 어떤 연관성을 가지는지 탐구하는 것이었다. 이를 위해 '프로그래밍 입문' 수강생들을 대상으로 ChatGPT와의 문답을 통해 합병 정렬 알고리즘을 학습하도록 하고, 수집된 대화 데이터와 학업성적 간의 상관관계를 분석하였다.

연구 결과에 따르면, 학업성취도를 나타내는 백분위와 질문의 개수 간에는 유의미한 음의 상관관계가 나타났지만, 질문의 총 글자 수와는 상관관계가 관찰되지 않았다. 이는 학업 성적이 높을수록 더 깊은 이해와 지식습득을 위해 LLM 챗봇과의 상호작용 횟수가 많아지는 경향이 있으나, 질문의 작성방식 (예시, 내용 정리 및 확인 등)과는 뚜렷한 연관이 관찰되지 않음을 의미한다. 또한, 백분위와 질문 간 평균 코사인 유사도 사이에서 유의미한 양의 상관관계가 관찰되었다. 이는 학업 성적이 높은 학습자가 개념을 이해하기 위해 더 다양한 주제와 유형의 질문을 던지는 경향이 있음을 시사한다.

본 연구의 한계점은 다음과 같다. 첫째로, 연구에서 수행한 상관분석은 통계적으로 유의미했으나, 상관관계가 약하게 관찰되어 결과의 강도가 두드러지지 않는다는 점이 연구의 한계로 지적될 수 있다. 둘째로, 학업성취도 분석 시 프로그래밍 교과목에 대한 점수만 포함하였으므로, 전체 교과 과정의 학업 성취도를 대표하기는 어렵다. 마지막으로, 실험에서 사용한 LLM 기반 챗봇은 GPT-3.5 기반 무료버전 ChatGPT이므로, GPT-4 및 4o과 같은 유료버전을 사용할 경우 응답의 질과 상호작용 패턴이 달라질 가능성이 있다. 따라서 이에 대한 추가적인 후속 연구가 필요하다. 그럼에도 불구하고, 본 연구는 합병 정렬이라는 제한된 개념 학습 맥락에서 수집된 질의 데이터를 바탕으로 통계적으로 유의미한 결과를 도출했다는 점에서 의의가 있다. 이는 학업 성취도와 질의 패턴 간에 상관관계가 존재한다는 초기 가설을 부분적으로 지지하는 결과로 해석된다.

이를 통해 실 교육 분야에 도입되고 있는 LLM 기반 학습보조 에이전트 개발 시, 학습자의 질문패턴을 기반으로 효과적인 지식 습득을 지원하는 디자인 가이드라인을 논의할 수 있다. 예를 들어, LMS(Learning Management System)[13], Khan

Academy[14], Quillbots[15] 등의 시스템 상에서 학습자 질문의 상호작용 횟수에 따라 추가적인 질문을 유도하는 인터랙션이 제공될 수 있다. 더불어 학습자 질의 다양성을 분석하여 더 넓은 스펙트럼의 주제와 개념에 대해 포괄적으로 학습할 수 있도록 키워드를 제안할 수 있다.

추후 프로그래밍 과목에서의 한정된 개념 학습을 넘어, 다양한 과목과 주제에서 나타나는 질문 패턴을 분석하고, 이를 학업 성취도와 연관 지어 탐구하는 후속 연구가 필요하다. 이를 통해 향후 LLM 챗봇 인터페이스 개발 시 개인화된 학습을 지원할 수 있는 질의 디자인 가이드라인을 도출하는 데 기여할 것으로 기대된다.

사사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2024-00353125).

참고 문헌

1. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences, 103, 102274.
2. MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., & Huang, Z. (2022, August). Generating diverse code explanations using the gpt-3 large language model. In Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2 (pp. 37-39).
3. Davoudi, M., & Sadeghi, N. A. (2015). A Systematic Review of Research on Questioning as a High-Level Cognitive Strategy. English Language Teaching, 8(10), 76-90.
4. Kazemitabaar, M., Hou, X., Henley, A., Ericson, B. J., Weintrop, D., & Grossman, T. (2023, November). How novices use LLM-based code generators to solve CS1 coding tasks in a self-paced learning environment. In Proceedings of the 23rd Koli Calling International Conference on Computing Education Research (pp. 1-12).
5. Jin, H., Lee, S., Shin, H., & Kim, J. (2024, May). Teach AI How to Code: Using Large Language Models as Teachable Agents for Programming Education. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-28).
6. Han, J., Yoo, H., Myung, J., Kim, M., Lee, T. Y., Ahn, S. Y., & Oh, A. (2024). RECIPE4U: Student-ChatGPT Interaction Dataset in EFL Writing Education. arXiv preprint arXiv:2403.08272.
7. OpenAI. (n.d.). ChatGPT (GPT-3.5) [Free tier]. OpenAI. Retrieved January 4, 2025, from https://www.openai.com
8. Kazemitabaar, M., Chow, J., Ma, C. K. T., Ericson, B. J., Weintrop, D., & Grossman, T. (2023, April). Studying the effect of AI code generators on supporting novice learners in introductory programming. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-23).
9. Otis, N., Clarke, R. P., Delecourt, S., Holtz, D., & Koning, R. (2023). The uneven impact of generative AI on entrepreneurial performance. Available at SSRN 4671369.
10. Zhang, X., Cole, M., & Belkin, N. (2011, July). Predicting users' domain knowledge from search behaviors. In Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval (pp. 1225-1226).
11. White, R. W., Dumais, S. T., & Teevan, J. (2009, February). Characterizing the influence of domain expertise on web search behavior. In Proceedings of the second ACM international conference on web search and data mining (pp. 132-141).
12. Ide, E., & Olivares-Rodriguez, C. (2021, November). Semantic expansion to improve diversity in query formulation. In 2021 IEEE

Latin American Conference on Computational Intelligence (LA-CCI) (pp. 1-6). IEEE.
13. Blackboard Inc. Learning Management System (LMS). https://www.blackboard.com/. December 1, 2024.
14. Khan Academy. Khan Academy. https://www.khanacademy.org/. December 1, 2024.
15. QuillBot. QuillBot. https://quillbot.com/. December 1, 2024.
16. G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE transactions on knowledge and data engineering, vol. 17, no. 6, pp. 734 – 749, 2005.