컴퓨터교육학회 논문지 2026년 제29권 제6호
https://doi.org/10.32431/kace.2026.29.6.007

거대언어모델 활용 초등 수학 문항 난이도 자동 예측: 초등학교 5학년 수학 문항을 중심으로

Automatic Prediction of Elementary Mathematics Item Difficulty Using Large Language Models: Focusing on Grade 5 Mathematics Items

추성엽†
Seongyeub Chu†

요약

본 연구는 거대언어모델(LLM)을 활용하여 초등 5학년 수학 문항 난이도를 자동으로 예측하는 방법을 제안하고 그 가능성을 분석한다. 이를 위해 인코더 기반, 인코더-디코더 기반, 그리고 디코더 기반 LLM을 공개된 초등 수학 문항 데이터로 파인튜닝하고, 기존에 많이 사용되는 기계학습 기반 모델 및 파인튜닝을 하지 않은 LLM과 예측 성능을 비교하였다. 그 결과, 학습 데이터와 동일한 학기·내용영역뿐만 아니라 학습하지 않은 타 학기·타 내용영역 문항에서도 파인튜닝된 디코더 기반 LLM이 가장 높은 성능을 보였다. 또한 주의집중(attention) 분석을 통해 LLM의 난이도 예측 성능이 높을수록 교육과정 개념과 관련한 문항 내 단어를 더욱 참조함을 확인하였다.

주제어 초등 수학, 문항 난이도 예측, 거대언어모델, 기계학습, 주의집중

ABSTRACT

This study proposes a method for automatically predicting the difficulty of grade 5 elementary mathematics items using large language models (LLMs). To this end, encoder-based, encoder-decoder-based, and decoder-based LLMs were fine-tuned on a public dataset and compared with machine learning–based models and non–fine-tuned LLMs. The results show that fine-tuned decoder-based LLMs achieved the best performance on both seen and unseen items across semesters and domains. Furthermore, attention-based analysis indicates that higher-performing models attend more to item words that are closely related to curriculum-based concepts.

Keywords Elementary Mathematics, Item Difficulty Prediction, Large Language Models, Machine Learning, Attention

†정회원 한국과학기술원 데이터사이언스대학원 박사과정

논문투고 2026년 02월 10일
심사완료 2026년 04월 27일
게재확정 2026년 05월 18일
발행일자 2026년 06월 30일

1. 서론

수학 문항의 난이도는 학습자의 성취 수준을 정확하게 측정하고, 적절한 학습 경험을 설계하는 데 있어 핵심적인 요소이며, 제작한 문항의 질을 평가하는 데에도 유용한 평가 기준으로 활용된다. 또한, 문항 난이도는 학습자의 수준에 맞는 문제를 제공하기 위한 적응형 학습 시스템(adaptive learning system)의 핵심 기준으로도 활용된다. 따라서 수학 문항의 난이도를 사전에 정확하게 예측하는 것은 교육 평가와 학습 지원 전반에서 중요한 과제로 자리 잡아 왔다[1].

전통적으로 문항 난이도는 대규모 사전 검사(pre-test) 기반의 경험적 난이도(empirical difficulty)나 전문가 판단 기반의 인지적 난이도(cognitive difficulty)를 통해 추정되어 왔다[2, 3]. 그러나 이러한 방식은 상당한 시간과 비용을 요구하며, 다양한 학습자 집단과 교육 환경에서 반복적으로 검증되어야 한다는 점에서 현실적인 한계를 지닌다[4]. 특히 대규모 표준화 평가(large-scale assessment)나 국가 수준 학업성취도 평가에서는 문항의 신뢰성을 확보하기 위해 여러 차례의 현장 검증이 필요하며, 이로 인한 비용 부담은 지속적으로 증가하고 있다. 이러한 맥락에서 학습자 응답 데이터에 의존하지 않고 AI 모델을 활용하여 자동으로 문항 난이도를 예측하는 것은 비용 효율적이고 확장 가능한 대안으로 주목받고 있다[5].

더 나아가 최근 지능형 튜터링 시스템(intelligent tutoring systems)과 적응형 학습 환경의 확산은 문항 난이도 예측의 중요성을 더욱 부각시키고 있다. 효과적인 적응형 학습을 위해서는 학습자의 결손 개념 습득을 지원하기에 적합한 문항을 순차적으로 제시하는 문항 배열(item sequencing)이 필수적이며, 이를 위해서는 각 문항의 난이도를 사전에 정확히 파악하는 과정이 요구된다. 수학과 같이 개념 간 위계 구조가 뚜렷한 교과에서는 부적절한 난이도의 문항 제시는 학습자의 과도한 인지 부담을 초래하거나[6, 7], 목표하는 개념 형성을 저해할 수 있기 때문에 정확한 난이도 예측 기술의 필요성은 더욱 커지고 있다[3]. 본 연구에서의 문항 난이도는 단순히 학생의 정답률 기반의 경험적 난이도보다는 각 문항의 내재 개념에 대한 인지적 난이도에 주된 초점을 맞춘다. 이것은 학생의 성취도 향상 목적의 문항 선별을 위해서는 내용영역의 요소 및 성취기준과 연결되는 문항 배열이 중요하며 이 과정에서 교과 전문가의 관점이 필요하기 때문이다[2, 3].

이러한 맥락에서 기존 연구들은 문항의 메타데이터(metadata)에서 나타나는 특성을 기반으로 기계학습(machine learning) 모델을 활용하여 문항 난이도를 예측하는 방법을 제안해 왔다[8, 9]. Random Forest[8]나 Gradient Boosting[9] 계열 모델은 비교적 안정적인 성능을 보였으나, 문항의 메타데이터와 표면적 특성에 대한 의존도가 높아 문항 텍스트에 내재된 개념이나 교육과정의 내용을 충분히 반영하는 데에는 한계가 있었다. 최근에는 거대언어모델(large language models, LLM)의 발전에 따라 문항의 의미적 맥락과 개념 정보를 보다 풍부하게 활용할 수 있는 새로운 접근이 제시되고 있으나[5, 10], LLM 기반 문항 난이도 예측 연구는 아직 초기 단계에 머물러 있으며, 특히 국내 교육과정과 연결한 분석은 충분히 이루어지지 않았다.

이에 본 연구는 초등학교 5학년 수학 문항을 대상으로 LLM 기반 모델의 문항 독해 난이도 예측 성능을 체계적으로 비교·분석한다. 구체적으로, 데이터 기반 파인튜닝(fine-tuning)을 적용한 다양한 LLM이 기존 기계학습 모델에 비해 어떠한 성능적 이점을 가지는지 탐구하고, 학기 및 내용영역이 변화하는 상황에서도 일반화 성능을 유지할 수 있는지 분석한다. 나아가 LLM의 주의집중(attention) 양상을 분석하여, 문항 텍스트 독해를 통한 난이도 예측 과정에서 2022 개정 교육과정 내용체계의 핵심어가 어떻게 참조되는지 정량적·정성적으로 살펴봄으로써, LLM 기반 난이도 예측의 교육적 해석 가능성을 탐색한다. 이를 위해 본 연구는 다음과 같은 연구 문제를 설정하였다.

연구 문제 1) LLM을 파인튜닝하였을 때 초등 5학년 수학 문항 난이도 예측에서 어떠한 양상을 보이는가?

연구 문제 2) LLM을 파인튜닝하였을 때 학습 데이터에 포함되지 않은 학기와 내용영역의 초등 5학년 수학 문항 난이도 예측에서 어떠한 양상을 보이는가?

연구 문제 3) 초등 5학년 수학 문항 난이도 예측 과정에서 LLM이 주의집중하는 문항 내 단어는 교육과정과 비교하였을 때 어떤 특성을 보이는가?

2. 이론적 배경

2.1 수학 문항 난이도 자동화 예측

수학을 비롯하여 다양한 교과의 문항 난이도 예측은 대규모 평가에서의 균형있는 문항 배열, 학습 과정에서 학생의 개념 습득을 위한 적절한 문항 제시 등을 위해 반드시 필요한 문항 분석 절차이다[5]. 문항 난이도 분석은 인지적, 경험적, 체감 난이도로 분류된다[2, 3]. 먼저, 인지적 난이도는 교과 전문가가 개념별 인지적 요구 수준을 고려하여 판단하는 난이도이고, 경험적 난이도는 임의의 학생 집단을 대상으로 사전 검사를 통한 정답률을 기반으로 측정하며, 체감 난이도는 문항에 대한 학생들의 설문 혹은 인터뷰를 통해 측정된다. 수식이나 도표로 문항이 구성되는 수학교과는 문항 텍스트에 존재하는 형태적 특성(예: 계산하기, 설명하기, 그리기), 교육과정 내용의 포함 비율, 문항의 유형(예: 선다형, 주관식) 등이 특성으로 활용되어 난이도가 측정된다[11, 12]. 이처럼 문항 난이도 분석 과정은 필연적으로 많은 시간과 비용이 발생한다.

수학 문항의 난이도를 자동으로 예측하려는 연구는 문항 개발 비용을 절감하고, 대규모 평가 및 학습 시스템에서 효율적인 사전 문항 검증을 가능하게 한다는 점에서 교육측정 및 학습분석 분야에서 지속적으로 주목받아 왔다[5,

13]. 이와 관련하여 초기 연구들은 문항의 난이도가 문항에서 나타나는 인지적 요구[14]와 메타정보에 의해 설명될 수 있다는 가정 하에, 이들을 활용하여 통계적 방식으로 접근하였다[15]. 구체적으로, 수학 역량(예: 추론, 모델링 등)에 대한 요구 여부를 통해 통계적으로 PISA 제공 문항 난이도를 예측하거나[16], 문항 관련 성취기준(achievement-level descriptor)에서 명시된 인지적 요구 수준과 독해 부담(reading load) 등이 실제 난이도와 밀접한 관련이 있다는 점에 주목하여 이를 문항 난이도 예측에 활용하려는 시도가 지속적으로 이루어져 왔다[17]. 이러한 방법은 비록 현대적 기계학습 기법을 사용하지는 않았으나, 이후 문항 난이도 예측 연구의 개념적 토대를 마련하였다.

이후 기계학습 기술의 발전에 따라, 문항 난이도 예측 자동화 연구는 문항 특성을 변수로 활용한 모델링을 적용하는 연구로 이어졌으며, 주로 복수의 의사결정 트리(decision tree) 알고리즘을 결합한 앙상블(ensemble) 모델이 많이 활용되었다. 대표적으로, 다수의 결정 트리를 독립적으로 학습한 뒤 예측 결과를 종합하는 배깅(bagging) 기반의 Random Forest 모델[8, 18]과 여러 트리 기반 모델을 순차적으로 학습하여 이전 단계의 오류를 보완하는 방식인 그래디언트 부스팅(gradient boosting) 기반의 XGBoost 모델[9, 19]이 활용되었다. 이러한 기계학습 기반 모델은 선행연구에서 문항 난이도와 연관성이 높다고 보고된 문항 텍스트 내 단어 수(word count)와 문장 수(sentence count)[20], 평균 문장 길이(average sentence length)[21], 그리고 문항 자체의 다양한 메타 데이터(예: 단원, 주제, 연산자 수 등)[9]에 대한 특성 공학(feature engineering)으로 산출된 변수를 활용하여 문항 난이도를 예측하였다. 그러나 이러한 방법들은 특성 공학으로 산출된 표면적인 변수에 대한 높은 의존성과 함께 데이터 과적합으로 인해 학습된 범위 이외의 문항에 대한 제한된 일반화 성능이 그 한계점으로 꾸준히 지적되어 왔다[13].

이러한 기존 기계학습 기반 방법들의 한계를 극복하기 위해 문항 텍스트가 갖는 의미적 특성에 관심이 집중되었고, 주의집중 기법을 활용한 트랜스포머(transformer)[22] 구조 기반의 LLM 활용이 시도되고 있다. 최근 문항 난이도 예측 분야에서는 인코더 기반 LLM인 BERT[23]를 데이터로 파인튜닝함으로써 기존의 기계학습 기반 방법을 능가하는 정확도를 보여 파인튜닝의 효과를 확인하였다. 그러나, 이러한 인코더 기반 LLM은 입력되는 텍스트 정보에 대한 의존도가 높기 때문에 파인튜닝에 활용되는 데이터 특성에 매우 민감하다[24, 25].

이에 따라, 텍스트 생성에 집중하여 텍스트에 표면적으로 드러나지 않는 정보를 파악하는 높은 추론 능력이 발현되어 최근 관심도가 높아지고 있는 디코더 기반 LLM을 문항 난이도 예측에 활용하고자 하는 시도가 점진적으로 증가하고 있지만, 여전히 GPT와 같이 개발사에서 블랙박스 형태로 제공하여 보안 유지가 어려운 Proprietary 모델에 의존하고 있고, 단순히 프롬프팅(prompting) 기반으로 LLM을 활용하여 문항 텍스트로부터 정보를 추출하고 이것을 기존의 기계학습 모델과 결합하는 방식에 국한되어 있다[5, 10]. 이러한 방법은 지속적인 유료 API 방식으로 인해 사용량에 비례하여 비용이 발생함과 동시에 교육 데이터의 외부 노출 위험으로부터 자유롭지 못하다. 또한, LLM을 수학 문항 난이도 예측에 맞게 파인튜닝하지 않아 LLM의 언어 이해 및 추론 능력을 충분히 활용하지 못하게 된다.

이러한 기존 연구의 한계를 해결하기 위해 본 연구는 최근 선행연구에서 확인된 인코더 기반의 LLM에 대한 문항 난이도 데이터 활용 파인튜닝 효과에 착안하여, 인코더-디코더 기반 모델과 디코더 기반 모델을 포함한 다양한 LLM을 국내 초등 5학년 수학 문항 데이터로 파인튜닝한 후 난이도 예측 성능을 비교·분석한다. 나아가 LLM의 주의집중 양상을 분석하여, 파인튜닝된 LLM이 주의집중하는 문항 텍스트의 부분과 성능의 연관성을 분석함으로써 국내 교육 환경에서의 파인튜닝 기반 LLM의 활용 가능성을 탐구하고자 한다.

2.2 주의집중 기반 거대언어모델

트랜스포머는 입력된 데이터의 모든 부분을 참고하던 순환 신경망(recurrent neural networks, RNN) 혹은 합성곱 신경망(convolutional neural networks, CNN)과 같은 기존 신경망 구조와 달리 목표하는 과업의 성격에 따라 입력된 데이터에서 연관된 부분만 선택적으로 참고하는 주의집중 메커니즘을 기반으로 맥락 정보를 모델링한다는 점에서 텍스트 혹은 이미지 처리 딥러닝 분야에 중요한 전환점을 마련하였다[22].

트랜스포머의 핵심 구성 요소인 자기 주의집중(self-attention)은 입력 문장을 문장 구성의 최소 의미 단위인 토큰(token) 단위로 분할한 뒤, 토큰별로 문장 내의 다른 토큰과의 관계를 참조하여 관련도에 따라 주의집중 정도를 차등적으로 부여하도록 설계된 기법이다. 여기서 토큰은 전체 단어가 될 수도 있고, 분할된 단어의 일부(subword) 단위가 될 수도 있다. 자기 주의집중을 통해 트랜스포머는 텍스트 내에서 위치상 멀리 떨어져 있지만 의미적으로 관련성이 높은 단어들 간의 관계를 효과적으로 모델링할 수 있다. 이 과정에서 각 토큰은 서로 다른 주의집중 가중치(attention weight)를 부여받으며, 이를 통해 언어모델은 텍스트 내에서 핵심 정보와 상대적으로 중요도가 낮은 정보를 구분한다.

주의집중 기법은 단순히 언어모델 성능을 향상시키는 것을 넘어, 언어모델이 텍스트 해석과 과업 수행 과정에서 어떤 언어적 단서에 집중하는지에 대한 근거를 제공한다. 특히 문항 난이도 예측과 같이 복합적인 의미 해석 및 교육과정 기반 판단이 요구되는 과업에서는, 주의집중 양상이 모델의 난이도 예측 전략을 간접적으로 파악할 수 있는 단서로 활용될 수 있다[26, 27].

트랜스포머 등장 이후, 언어모델은 대규모 텍스트 데이터와 주의집중 기법을 접목함으로써 고도의 추론이 필요한 텍스트 생성에서 뛰어난 능력을 발휘하게 되었고, 이는 LLM의 발전으로 이어졌다[28-30]. LLM은 인코더 기반, 디코

더 기반, 인코더-디코더 기반 모델로 분류된다. 먼저, 인코더 기반 LLM은 입력 텍스트의 문맥적 표현을 자기 주의집중을 통해 벡터로 표현(representation)하여 텍스트를 이해하는 것에 특화되어 단순한 문맥 이해 과업에서 높은 수행 능력을 보이는 LLM이다[23]. 다음으로, 인코더-디코더 기반 LLM은 인코더에서 입력 텍스트의 의미를 먼저 파악한 뒤 디코더가 이를 참조하여 텍스트를 출력하는 방식으로 작동하여 교차 주의집중(cross-attention)을 통해 인코더와 디코더의 관계 파악에 집중하도록 설계된 모델로 번역 과업에 특화되어 있다[31]. 이때 디코더란 자기회기적 방식(autoregressive)으로 토큰을 생성하는 모듈로, 입력된 이전 토큰을 조건으로 높은 추론 능력을 활용하여 다음 토큰을 순차적으로 생성한다. 구체적으로, 'How are you?'라는 문장을 입력받아 'I'm'까지 디코더가 생성하였다면 그 다음으로 올 수 있는 토큰인 'fine'을 생성하기 위해 입력된 문장과 이전까지 생성된 토큰을 연결한 'How are you? I'm' 부분을 모두 참고하며 이 과정에서 중요한 토큰을 더 많이 참고하는 주의집중 기법이 활용된다. 마지막으로, 디코더 기반 모델은 입력된 텍스트에 대한 자기 주의집중을 통해 이해한 내용을 기반으로 새로운 텍스트를 생성한다. 대규모 텍스트 데이터로 사전학습된 GPT, Gemini, LLaMA 등의 고수준 LLM은 이러한 디코더 기반 구조를 채택하고 있으며 입력된 텍스트로부터 고도의 추론이 필요한 과업에서 높은 성능을 달성한다[28, 29, 32].

본 연구는 트랜스포머 구조의 주의집중 기법을 활용하여 다양한 유형의 LLM이 국내 초등 5학년 수학 문항 난이도 예측 과정에서 어느 정도의 성능을 달성하며, 어떠한 단어와 개념에 주의집중하는지 정량적·정성적으로 분석하고자 한다. 이를 통해 단순히 LLM의 성능을 제시하는 데 그치지 않고, 모델의 내부 판단 과정을 2022 개정 초등 수학 교육과정과 연결지어 설명 가능성의 관점에서 분석한다는 점에서 기존 LLM 활용 수학 문항 난이도 예측 연구와 차별화된다.

3. 연구 방법

3.1 문항 난이도 예측 모델

본 연구에서는 초등학교 5학년 수학 문항 난이도를 예측하는 모델의 성능과 특성을 분석하기 위해, 기존 선행연구[5, 8, 9]에서 가장 널리 사용되며 안정적인 성능을 보이는 두 가지 기계학습 기반 앙상블 모델인 XGBoost[19]와 Random Forest[18], 그리고 다양한 유형의 한국어 지원 LLM을 활용하였다. LLM으로는 인코더 기반 모델에 해당하는 KLUE-RoBERTa-Large(0.3B)[33], 인코더-디코더 기반 모델에 해당하는 Pko-T5-Large(0.8B)[34], 그리고 디코더 기반 모델에 해당하는 Qwen3-0.6B[35]와 LLaMA3-Instruct 계열의 3B 및 8B 모델[28]을 활용하였다. 각 LLM 모델명에 붙은 '-B'는 모델의 파라미터 수를 나타내며 십억 단위(billions)를 나타낸다. 표현의 편의상 KLUE-RoBERTa-Large와 Pko-T5-Large는 각각 RoBERTa와 T5로 본 논문에서 표기하였다. 디코더 기반 LLM으로 Qwen과 LLaMA를 선택한 이유는 오픈소스 모델로서 사전학습 시 한국어를 포함한 다국어 데이터셋(multilingual datasets)이 활용되어 영어 이외의 다국어 이해 능력이 뛰어나 국내 초등 수학 문항의 난이도 예측에 적합하고, 데이터 노출 및 보안에 민감한 교육 분야 고유의 특성과 연산 자원이 부족한 교육현장을 고려하였을 때 Open AI의 GPT[29]나 구글의 Gemini[32]와 다르게 파인튜닝 후 모델의 자체 관리가 가능하여 비교적 안전하기 때문이다.

XGBoost[19]는 그래디언트 부스팅 기반의 앙상블 모델로, 여러 약한 결정 트리를 순차적으로 학습하여 이전 단계의 오류를 보완하는 방식으로 높은 예측 성능을 달성한다. 특히 테이블 형태의 데이터에서 안정적인 성능을 보이며, 다양한 특성 간의 비선형 관계를 효과적으로 모델링할 수 있어 교육 데이터 분석 분야에서 널리 활용되어 왔다[9].

Random Forest[18]는 다수의 결정 트리를 독립적으로 학습한 뒤 예측 결과를 종합하는 배깅 기반 앙상블 모델이다. 개별 트리의 과적합을 완화하고 분산을 줄이는 데 강점을 가지며, 특성 수와 데이터 규모가 제한된 환경에서도 비교적 견고한 성능을 보인다.

LLM은 다양한 유형을 활용하되, 디코더 기반 LLM은 다른 인코더 기반 및 인코더-디코더 기반 LLM과 유사한 규모의 파라미터 수를 가진 Qwen3-0.6B를 함께 비교군에 포함하여 비슷한 파라미터 수에서의 모델 비교도 수행하였다. LLaMA3 모델은 3B와 8B 수준의 파라미터 규모에 따른 성능 차이 또한 비교·분석하였으며, 데이터 기반 파인튜닝을 적용한 경우와 적용하지 않은 경우도 추가로 비교하여 파인튜닝이 LLM의 문항 난이도 예측에 미치는 영향 또한 검증하고자 하였다.

3.2 분석 데이터셋

본 연구에서 활용한 데이터셋은 AI Hub(https://aihub.or.kr)에서 공개된 '수학 과목 문제 생성 데이터셋'으로 ㈜대교에서 문제은행을 기반으로 문제 생성 모델 학습을 위해 3만 건 이상의 학년 및 유형별 수학 문제 및 풀이과정 이미지와 대응되는 텍스트 데이터를 구축한 문항 데이터셋이다. 본 연구는 초등 5학년을 연구 대상으로 하기 때문에 관련 데이터만 NAN 값 제거 후 전처리하여 활용하였고, 데이터 정보는 Table 1과 같다. 데이터는 총 8,565개의 문항 데이터를 활용하였고, 문항 난이도는 교육과정의 내용영역별 목표, 내용요소, 성취기준 등을 기준으로 수학 교과 전문가에 의해 3단계로 평가되어 '상(1,706, 19.92%)', '중(4,508, 52.63%)', '하(2,351, 27.45%)'로 데이터셋에서 제공되는 것으로 활용하였다. 본 연구에서 활용한 데이터셋 변수는 난이도를 포함하여 Table 2와 같이 총 10개이며, 변수명(예: question_term)은 데이터셋에 명시된 것을 활용하였다.

Table 1. Statistics of Dataset

Level of School	Grade	Number of Item	Proportion (%)		
			High	Middle	Low
Elementary School	5	8,565	19.92% (1,706)	52.63% (4,508)	27.45% (2,351)

Table 2. Description of Features

Feature	Description	Type	Range
question_term	Semester	Number	1, 2
question_unit	Item Unit Title	String	
question_topic_name	Item Topic	String	
question_type1	Item Type	String	Short Answer, Multiple Choice
question_step	Learning Stage	String	Basic, Application
question_sector1	Assessment Domain	String	Calculation, Understanding, Reasoning, Problem Solving
question_sector2	Content Domain	String	Numbers and Operations, Geometry and Measurement, Change and Relationships, Data and Chance
question_difficulty	Item Difficulty Level	Number	1=high, 2=medium, 3=low
question_text	Item Text	String	
answer_text	Model Answer Text	String	

본 연구는 서론에서 제시한 바와 같이 대규모 평가 문항 구성과 학습자의 학습 지원을 위한 문항 배열 목적의 난이도 평가 효율화를 위해서 LLM이 개념별 인지적 난이도를 예측하는 정도를 분석하는 것을 목표로 한다. 이를 위해 교육과정의 내용요소와 성취기준을 기반으로 전문가가 분류한 인지적 난이도를 레이블로 활용하였다. 이러한 방법은 사전 검사 맥락 및 대상 집단에 의존적인 경험적 난이도의 한계를 완화할 수 있지만 동시에 전문가의 주관이 개입되며 실제 학생을 대상으로 한 실증이 이루어지지 않았다는 점에서 또 다른 한계가 존재한다[2, 3]. 따라서, 본 연구의 결과물을 바탕으로 국내의 다양한 초등학생 집단을 대상으로 경험적 난이도에 대한 후속 연구가 필요하다.

3.3 연구 절차

3.3.1 데이터 전처리

데이터 전처리는 다음과 같이 진행되었으며 전처리된 데이터의 정보는 Table 3과 같다. 먼저, 초등 5학년 데이터셋을 '1학기 수와 연산', '2학기 수와 연산', '1학기 도형과 측정', '2학기 자료와 가능성'에 해당하는 데이터셋으로 분류한다. 다음으로, 모델 학습에 활용되는 1학기 수와 연산 데이터의 적은 수를 고려하여 5-폴드 교차 검증 기법을 적용하고자 학습 데이터와 테스트 데이터를 80:20 비율로 분할하여 활용하였다. 그리고 학습 데이터 중 테스트 데이터와 동일한 수만큼 분할하여 모델 학습 시 조기 종료(early stopping)를 위한 검증 데이터로 활용하였다. 이때 각 분할에는 난이도 수준별 비율이 유지되도록 설정하였다. 타 학기 및 영역에서의 문항 난이도 예측 일반화 성능 확인에 사용되는 데이터셋은 테스트에 활용되기 때문에 별도의 분할 없이 모든 데이터가 활용되었고, 각 폴드별로 다르게 분할된 1학기 수와 연산 학습 데이터셋으로 학습된 모델로 테스트하였다.

Table 3. Statistics of Preprocessed Data (N=8,565)

Semester	Content	Train	830	Test
1	Number and Operations	2,493	–	830
2	Number and Operations	–	–	2,744
1	Geometry and Measurement	–	–	1,151
2	Data and Chance	–		517

XGBoost와 Random Forest 모델 학습의 경우는 각 데이터셋에 대해 Table 2에 제시된 변수 중 정답 레이블에 해당하는 question_difficulty를 제외한 9가지 변수가 입력 특성으로 활용되었다. 이때, 숫자가 아닌 string 타입의 명목형 변수는 원-핫 인코딩(one-hot encoding)으로 변환하여 활용되었다. 추가적으로, 'question_topic_name + question_text + answer_text' 형태로 Table 4와 같이 문항의 주제, 문항 텍스트, 그리고 모범답안 텍스트를 이어 붙여 구성한 문항 관련 텍스트의 단어 수, 문장 수, 평균 문장 길이, 그리고 수식의 수를 추출하여 변수로 함께 활용하였다. 반면, LLM은 Table 4와 같이 문항 주제 및 텍스트, 그리고 모범답안 텍스트를 이어 붙인 문항 관련 텍스트만 입력받아 문항의 난이도를 예측하도록 설계되었다.

Table 4. Example of Concatenated Item Text

Item Text Example
● Topic (question_topic_name): Understanding Multiples ● Problem (question_text): A number from 1 to 60 is placed in □ so that the result of the calculation becomes a multiple of 6. How many numbers can be correctly placed in □? 312 + □ ● Sample Answer (answer_text): Since 312 ÷ 6 = 52, 312 is a multiple of 6, and thus □ must also be a multiple of 6. Among the numbers from 1 to 60, the multiples of 6 are 6 × 1 = 6, 6 × 2 = 12, ⋯, 6 × 10 = 60. Therefore, there are 10 such numbers.

3.3.2 실험 설정

본 연구는 초등학교 5학년 수학 문항 난이도 예측을 다중 범주 분류 문제(multi-class classification)로 설정하였으며, 기계학습 기반 모델과 LLM의 성능을 비교·분석하고,

LLM의 일반화 성능 및 예측 과정에서의 주의집중 양상을 다각도로 검증하기 위해 다음과 같은 세 가지 실험 설정을 구성하였다.

첫 번째 실험에서는 동일 학기 및 내용영역(1학기 수와 연산)의 데이터를 활용하여, 기계학습 기반 모델과 LLM의 문항 난이도 예측 성능을 비교하였다. 기계학습 기반 모델은 문항에서 추출한 사전 정의된 특성을 입력으로 하여 지도학습 방식으로 학습하였으며, LLM은 문항 관련 텍스트를 입력하여 지도학습 방식의 파인튜닝을 적용 및 미적용한 경우를 구분하여 실험하였다.

두 번째 실험에서는 모델 학습에 사용되지 않은 영역의 문항에 대한 모델의 일반화 성능을 평가하기 위해, 학습 데이터와 다른 학기 및 내용영역의 문항을 테스트 데이터로 활용하였다. 구체적으로, 1학기 수와 연산 영역 문항으로 학습한 모델을 대상으로, (1) 동일한 내용영역의 다른 학기 문항(2학기 수와 연산 영역)과 (2) 다른 내용영역의 문항(도형과 측정, 자료와 가능성)에 대한 난이도 예측 성능을 각각 분석하였다.

세 번째 실험에서는 LLM이 문항 난이도를 예측하는 과정에서 어떠한 단어에 주의집중하는지 분석하기 위해, 모델의 주의집중 가중치를 활용하였다. 구체적으로, 파인튜닝을 적용 및 미적용한 LLM을 대상으로 입력 문항 텍스트에서 주의집중 가중치가 높은 상위 K개의 단어(Top-K 단어군)를 추출하고, 이들 단어 중 2022 개정 초등학교 수학 교육과정의 5-6학년군 수와 연산 영역 내용체계와 성취기준에 제시된 핵심어가 차지하는 비율과 전체 핵심어 유형 중 Top-K 단어군에 등장한 핵심어 유형의 비율을 분석하였고, 추가로 사례 분석을 진행하였다. 이때, 분석의 용이성을 고려하여 LLM이 이해하는 하나의 토큰을 단어로 간주하였고, 교육과정 핵심어는 내용 체계와 성취기준에 명시된 핵심 내용요소로 Table 5(부록 Table 10 한국어 버전 참조)와 같이 선정하여 모든 문항에 일관되게 적용하였다.

Table 5. Key Curriculum Terms

Key Terms
number, natural number, fraction, numerator, denominator, decimal, addition, subtraction, multiplication, division, factor, multiple, round up, round down, round, number range, mixed operations, calculation, +, −, ×, ÷, =, 〉, 〈

3.4 평가지표

본 연구에서는 수학 문항 난이도 예측 모델의 성능을 정량적으로 평가하기 위해, 이차 가중 카파 계수(quadratic weighted kappa, QWK)와 F1 score를 사용하였다. QWK는 예측값과 실제값 간의 일치도를 측정하는 지표로, 범주 간 순서 정보와 오차의 크기를 반영하여 불일치 정도를 가중하여 계산한다. 본 연구는 순서가 존재하는 3가지 단계(상, 중, 하)로 난이도가 평가되기 때문에 이를 고려하기 위해 해당 지표가 사용되었다. F1 score는 정밀도(precision)와 재현율(recall)의 조화 평균으로, 난이도 간 분포 불균형이 존재하는 상황에서 모델의 성능을 보다 균형있게 평가하는 지표이다. 추가적으로, 혼동행렬(confusion matrix)을 활용하여 각 모델들이 실제 예측한 결과에 대한 분석을 진행하였고, 일부 모델들 간 성능 차이의 통계적 유의성을 분석하기 위해 대응표본 t-검정(paired t-test)을 수행하여 유의확률(p)과 효과크기(d)를 제시하였다.

3.5 모델 학습 환경 및 구현

실험은 Linux(Ubuntu 22.04) 기반 서버 환경(RAM 32GB)에서 수행되었으며, LLM 학습 및 추론에는 NVIDIA A100 GPU를 활용하였다. LLM은 Python 3.9 환경에서 PyTorch 2.8 프레임워크(https://pytorch.org/)와 Hugging Face Transformers 5.0 라이브러리(https://huggingface.co/)로 구현하였으며, 기계학습 기반 모델은 scikit-learn 1.5.1(https://scikit-learn.org) 및 XGBoost 2.1.4(https://xgboost.readthedocs.io) 라이브러리를 사용하여 구현하였다.

XGBoost와 Random Forest 모델은 학습을 위해 Estimator의 수(1000~5000, 500 단위), Max Depth(1~5), 학습률(0.01~0.00001, ×0.1 단위)을 조정하며 검증 데이터셋으로 그리드 서치(grid search)를 통해 가장 좋은 성능을 보이는 모델을 최종 모델(Estimator: 5000, Max Depth: 5, 학습률: 0.001)로 선택하였다. Random Forest 모델은 학습률이 사용되지 않았고, 나머지 파라미터는 XGBoost 모델과 동일하게 설정되었다.

LLM은 Hugging Face Transformers 라이브러리에서 제공되는 사전학습된 가중치와 SFTTrainer 모듈을 활용하여 문항 난이도 예측 과제에 맞게 파인튜닝하였다. 파인튜닝은 토큰화(tokenizing)된 문항 관련 텍스트를 모델 입력으로 사용하고, 난이도 수준을 출력 토큰으로 예측하는 방식으로 수행되었다. 학습 과정에서는 AdamW 옵티마이저(optimizer)를 사용하였으며, 에폭(epoch)은 5, 배치(batch) 크기는 4, 학습률은 0.00005, 웜업 비율(warmup ratio)은 0.03으로 설정하여, 전체 학습 데이터 수 중 3% 만큼 학습될 때까지 학습률을 0에서 최대값까지 선형적으로 증가시켜 안정된 학습이 되도록 하는 선형 스케줄링(linear scheduling) 방식을 활용하였다. 또한, 과적합 방지를 위해 500 스텝마다 검증 후 검증 데이터에서의 성능을 기준으로 2회 연속 손실함수의 값이 감소하지 않으면 조기 종료되도록 하였다. 이외 나머지 하이퍼파라미터는 라이브러리에서 제공하는 기본값을 활용하였다.

추가적으로, 앞서 설명된 바와 같이 본 연구에서 활용된 모델 학습 데이터셋에는 난이도 수준별로 불균형이 존재(상: 18%, 중: 53%, 하: 29%)하기 때문에 이러한 불균형을 모델 학습 과정에서 통제할 수 있는지 확인하고자 상, 하 수준의 레이블에 대해 오버샘플링(oversampling)을 적용하여 중 수준의 레이블과 동일한 비율이 되도록 재구성한 후 모델을 학습시킴으로써 오버샘플링 적용 전 데이터(2,493)

와 적용 후 데이터(3,963)로 학습한 경우를 비교하였다.

4. 연구 결과

4.1 학습 데이터 범위의 수학 문항 난이도 예측

4.1.1 수학 문항 난이도 예측 성능 비교

Table 6은 기계학습 기반 모델과 다양한 LLM의 QWK 및 F1 점수를 제시한다. 본 실험은 모델의 학습에 활용된 문항 난이도 데이터와 동일한 학기(1학기) 및 내용영역(수와 연산)에서 모델 학습에 사용되지 않은 문항에 대한 난이도 예측 결과를 비교하였다.

기계학습 기반 모델은 모두 0.61 이상의 성능을 보였으며 특히 Random Forest가 XGBoost보다 더 우수한 성능을 보였다. 이는 해당 데이터에서 수작업으로 추출된 문항 특성을 활용한 전통적인 기계학습이 일정 수준의 예측 성능을 확보할 수 있음을 보여준다. 또한, 데이터의 수가 적고, 모델 활용 변수의 수가 많지 않을 때는 배깅 기반 앙상블 모델이 그래디언트 부스팅 기반 앙상블 모델보다 더욱 효과적이라는 선행연구를 고려하였을 때[19, 36, 37], 본 연구의 초등 수학 문항 난이도 예측 학습 데이터셋은 동일한 학기와 내용영역의 문항 난이도를 예측할 때 그 규모와 특성 변수의 수가 Random Forest 모델을 학습시키기에 더 적합함을 확인할 수 있다. 그러나, 이러한 모델들은 문항의 언어적 정보를 제한적으로 반영하는 구조적 한계를 지닌다.

Table 6. Performance of Math Item Difficulty Prediction in Terms of Quadratic Weighted Kappa (QWK) and F1 Score. Values Represent the Mean over 5 Folds, with Standard Deviations in Parentheses.

Category		Model	QWK	F1 Score
Machine Learning		XGBoost	0.6051 (0.017)	0.6734 (0.021)
		Random Forest	0.6249 (0.020)	0.7018 (0.017)
LLM	Encoder-based	RoBERTa (0.3B)	0.0000 (0.000)	0.2310 (0.000)
	Encoder-Decoder-based	T5 (0.8B)	0.7853 (0.022)	0.7780 (0.029)
	Decoder-based	Qwen-0.6B (W/O Fine-tuning)	0.0085 (0.029)	0.0049 (0.041)
		Qwen-0.6B (W Fine-tuning)	0.8142 (0.027)	0.8091 (0.021)
		LLaMA-3B (W/O Fine-tuning)	0.0139 (0.044)	0.0088 (0.035)
		LLaMA-3B (W Fine-tuning)	0.8184 (0.025)	0.8173 (0.021)
		LLaMA-8B (W/O Fine-tuning)	0.3132 (0.034)	0.3039 (0.029)
		LLaMA-8B (W Fine-tuning)	0.8416 (0.031)	0.8376 (0.016)

이에 비해 파인튜닝을 한 LLM은 RoBERTa를 제외하고 전반적으로 기계학습 기반 모델보다 우수한 성능을 보였다. 이는 문항의 메타데이터보다 문항 텍스트가 난이도 예측에 더 유의미한 변수로 작동함을 시사한다. 인코더 기반 모델인 RoBERTa는 모든 문항의 난이도를 '중' 수준으로 예측하는 편향을 보이며 기계학습 기반 모델보다 낮은 성능을 보인 반면, 인코더-디코더 기반 언어모델(T5)과 디코더 기반 언어모델(Qwen, LLaMA)은 두 평가지표 모두에서 전반적으로 다른 비교모델을 능가하였다. 대표적으로, 인코더-디코더 기반 LLM인 T5 및 디코더 기반 LLM 중 최저 성능을 보인 Qwen이 기계학습 모델 중 최고 성능을 보인 Random Forest와 통계적으로 유의미한 차이를 보였다(T5 vs. Random Forest: $p<0.001$, dqwk=0.2789, df1=0.2477, Qwen vs. Random Forest: $p<0.001$, dqwk=0.2912, df1=0.2643). 특히, 디코더 기반 언어모델이 인코더-디코더 기반 언어모델보다 높은 예측 성능을 보였다.

인코더 기반 언어모델의 성능이 상대적으로 매우 낮은 것은 본 연구에서 모든 LLM에 동일하게 설정한 하이퍼파라미터 셋팅 혹은 LLM에 입력하기 위한 문항 텍스트 구성 방식이 인코더 기반 언어모델에 최적화되지 않았고, 데이터셋에 존재하는 불균형으로 인해 예측 편향이 나타났을 가능성이 높다. 따라서 각 언어모델별로 최적화된 하이퍼파라미터 셋팅 혹은 입력 데이터 구성 방식에 대해 탐구하는 것은 유의미한 연구이며, 본 연구가 다양한 LLM 모델의 수학 문항 난이도 예측의 경향을 비교 · 분석하는 것에 초점을 두는 것을 고려하여 후속 연구로 남겨둔다. 반면, 텍스트 생성 과정에서 발현되는 디코더의 추론 능력은 본 연구의 데이터셋의 한계를 잘 극복하여 수학 문항 난이도 예측에 중요한 역할을 하는 것으로 해석할 수 있다.

텍스트 생성을 위한 추론에 특화된 Qwen-0.6B를 파인튜닝한 경우가 인코더와 디코더가 모두 활용되지만 추론보다 인코더 및 디코더 간의 정보 교류에 더 특화된 T5(0.8B)에 비해 적은 파라미터에도 불구하고 더 높은 성능을 보이며 그 차이가 통계적으로도 유의한 점($p<0.01$, dqwk=0.2046, df1=0.2165)은, 모델 파라미터 수보다 디코더 중심의 구조적 특성과 대규모 사전학습 과정에서 확보된 고수준의 추론 능력[28, 29, 32]이 난이도 예측에 필요한 정보를 효과적으로 포착하도록 작동함을 시사한다.

파인튜닝을 적용한 LLaMA-3B와 LLaMA-8B 모델은 각각 QWK 0.8184 및 0.8416, F1 점수 0.8173 및 0.8376를 기록하여, 다른 모든 비교 모델을 일관되게 능가하였다. LLaMA-8B가 LLaMA-3B보다 난이도 예측을 더 정확하게 하는 것으로 보아($p<0.05$, dqwk=0.1933 df1=0.1813), scaling law 현상에 의해 디코더 기반 LLM의 파라미터 규모가 클수록 초등 수학 문항 난이도 예측 능력이 향상됨을 확인할 수 있다[30, 38].

추가적으로, 본 연구의 학습 데이터셋에 존재하는 레이블 간 불균형 문제를 해결하기 위해서 Figure 1과 같이 오버샘플링을 적용한 경우를 기존 학습 데이터셋만 활용한 경우와 비교하였다. 그 결과, 오버샘플링을 적용한 경우 모델에 따라 성능 향상 여부가 달랐는데 그 차이는 크지 않았고,

RoBERTa의 예측 편향은 일부 완화되었으나 여전히 낮은 예측 성능을 보였으며, 모델별 성능의 경향 또한 동일하였다. 따라서, 이후의 실험은 모두 기존 데이터셋만 활용하여 학습한 모델의 결과를 비교하였다.

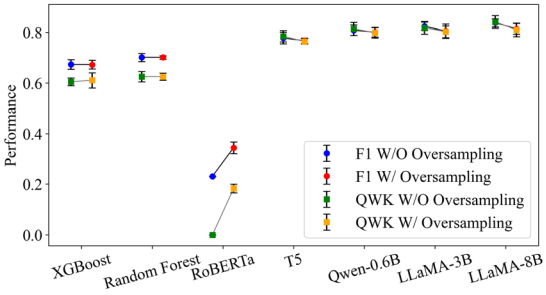

Figure 1. Model-wise Mean Performance over 5-Fold Cross-Validation under Two Training Conditions: without Oversampling and with Oversampling.

종합하면, 본 실험 결과는 파인튜닝을 적용한 디코더 기반 LLM이 수학 문항 난이도 예측 과제에서 다른 비교모델에 비해 더 높은 예측 성능을 달성함을 보여준다. 이는 디코더 기반 LLM이 대규모 사전학습을 통해 획득한 언어 이해 및 추론 능력을 교육 데이터셋을 활용한 파인튜닝을 통해 교육 맥락에 효과적으로 정렬시킬 수 있으며, 이를 통해 문항 난이도를 정확하게 예측할 수 있는 잠재력을 지니고 있음을 시사한다.

4.1.2 혼동행렬 기반 난이도 예측 결과 분석

각 모델의 실제 예측 결과를 분석하기 위해 기계학습 기반, 인코더 기반, 인코더-디코더 기반, 디코더 기반 유형별로 가장 높은 성능을 보인 모델의 혼동행렬을 Figure 2와 같이 비교하였다.

결과를 보면, 모델의 QWK와 F1 점수가 높을수록 '중' 수준으로 편향되는 경향이 완화되는 것을 확인할 수 있다. 이는 난이도 예측 성능이 향상될수록, '중'레이블이 과반을 차지하는 불균형 학습 데이터의 한계를 보다 효과적으로 극복함을 의미한다. 또한, '상-하'오류는 난이도 예측 성능이 높은 모델일수록 그 빈도가 감소하였는데, 두 수준이 난이도의 양극단에 위치해 있어 오류의 빈도도 다른 오류에 비해 상대적으로 적었다. 반면, '상-중'및'중-하'오류 유형의 빈도는 전반적으로 높았는데 이것은 인접 난이도를 구분하는 것이 상대적으로 어렵기 때문으로 해석된다. 그럼에도 불구하고, 실제 '중'수준을 '상'으로 예측한 경우의 T5를 제외하면, 성능이 높은 LLM일수록 인접 난이도 오류에서 그 빈도가 감소하였다. 이는 성능이 우수한 모델일수록 난이도 구분이 미묘한 경우에도 보다 정확한 예측을 수행함을 시사한다.

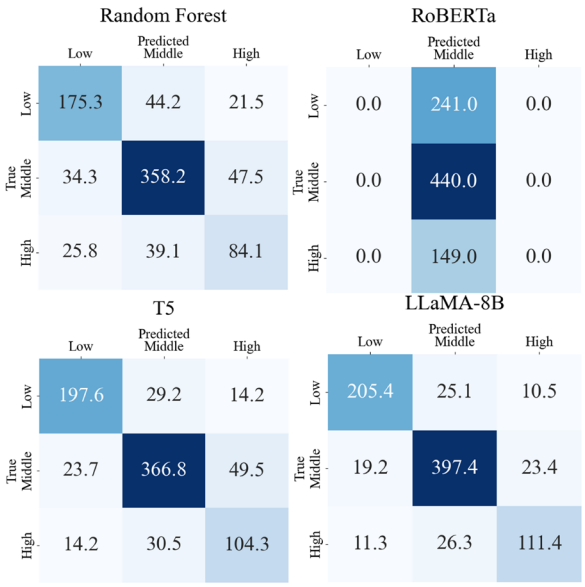

Figure 2. Confusion Matrices Showing the Predicted Labels versus True Labels. Values Represent the Mean over 5 Folds and Darker Shades of Blue Indicate Higher Counts.

4.1.3 디코더 기반 LLM의 파인튜닝 여부에 따른 문항 난이도 예측 결과 비교

Table 6을 보면, 파인튜닝을 하지 않은 디코더 기반 LLM은 다른 비교모델에 비해 매우 낮은 성능을 보였다. 디코더 기반 LLM의 zero-shot 설정에서는 QWK와 F1 점수가 크게 하락하였으며, 이는 사전학습된 지식만으로는 국내 초등 수학 문항 난이도 예측이 제한된다는 점을 보여준다. 이는 LLM의 수학 문항 난이도 예측에 있어서 도메인 특화 파인튜닝을 통한 도메인 정렬(domain alignment)의 중요성을 보여준다[39, 40].

추가적으로, 파인튜닝을 미적용한 경우에도 모델의 파라미터 사이즈가 클수록 문항 난이도 예측을 더욱 정확하게 수행하였다(8B>3B>0.6B). 이것은 파인튜닝을 하지 않더라도, scaling law 현상에 의해 모델의 파라미터 규모가 클수록 초등 수학 문항 난이도 예측 능력이 상대적으로 향상됨을 시사한다[30, 38]. 그러나, 파인튜닝을 미적용한 디코더 기반 LLM은 일관되게 매우 낮은 성능을 보였기 때문에 이후 실험에서는 제외하였다.

4.2 수학 문항 난이도 예측 일반화 성능 분석

4.2.1 동일 내용영역 내 타 학기 문항에 대한 난이도 예측 결과 분석

Table 7은 1학기 수와 연산 영역 문항 난이도 데이터셋으로 학습한 모델이, 동일한 내용영역에 속하는 2학기 문항들의 난이도를 예측한 성능을 보여준다. 이는 학습에 포함되지 않은 학기에 대한 모델의 일반화 성능을 분석하기 위한 설정이다.

실험 결과, 학습에 활용된 문항과 동일한 학기 및 내용영

역 문항에 비해, 타 학기 문항에 대한 난이도 예측 성능이 전반적으로 하락하였다. 특히, QWK 성능은 F1 점수에 비해 더욱 하락하였다. 이것은 모델들이 각 난이도 수준별로는 어느 정도 잘 예측하지만, 한 단계 차이보다 두 단계 차이가 나는 '상-하' 수준에서의 예측 오류를 상대적으로 많이 보이기 때문으로 해석할 수 있다. 따라서 타 학기에서의 난이도 예측 상황은 동일한 내용영역에 속하더라도 학기별로 상이한 세부 성취기준과 내용요소가 모델로 하여금 난이도 수준 간의 서열 관계를 안정적으로 학습하는 데 어려움을 유발하며, 특히 상-하 난이도 수준 간의 구분에서 추가적인 개선이 필요함을 시사한다.

Table 7. Mean Performance of Math Item Difficulty Prediction in a Different Semester in Terms of Quadratic Weighted Kappa (QWK) and F1 Score over 5-Fold Cross-Validation with Standard Deviations in Parentheses

Category		Model	QWK	F1 Score
Machine Learning		XGBoost	0.4752 (0.015)	0.5337 (0.023)
		Random Forest	0.4411 (0.020)	0.5109 (0.018)
LLM	Encoder-based	RoBERTa (0.3B)	0.0000 (0.000)	0.2299 (0.000)
	Encoder-Decoder-based	T5 (0.8B)	0.5522 (0.017)	0.6254 (0.015)
	Decoder-based	Qwen-0.6B	0.5841 (0.023)	0.6491 (0.019)
		LLaMA-3B	0.6052 (0.014)	0.6591 (0.016)
		LLaMA-8B	0.6513 (0.020)	0.7211 (0.017)

모델별 성능을 비교하면, 데이터 기반 파인튜닝 적용 LLM이 RoBERTa모델을 제외하고 모두 기계학습 기반 모델에 비해 높은 난이도 예측 수행을 보였다. 기계학습 기반 모델과 달리 파인튜닝을 한 인코더-디코더 기반 및 디코더 기반 LLM은 모든 평가지표에서 0.55 이상을 유지하였다. 대표적으로 인코더-디코더 LLM인 T5 및 디코더 기반 LLM 중 최저 성능을 보인 Qwen이 기계학습 모델 중 최고 성능을 보인 XGBoost와 통계적으로 유의미한 차이를 보였다(T5 vs. XGBoost: $p<0.001$, $d_{qwk}=0.2498$, $d_{f1}=0.2557$, Qwen vs. XGBoost: $p<0.001$, $d_{qwk}=0.2423$, $d_{f1}=0.2692$). 그러나, 인코더 기반 LLM은 여전히 '중'수준 예측 편향이 관찰되었다. 이는 앞서 4.1.1절에서 논의된 것처럼 본 연구의 하이퍼파라미터 셋팅과 입력 데이터 구성이 인코더 기반 모델에 최적화되지 않음에 기인했을 가능성이 있으며, 디코더의 추론 능력이 본 연구의 데이터셋이 갖는 한계를 잘 극복하고 LLM의 난이도 예측에 중요한 역할을 하며 학기 간 변화에도 강건하게(robust) 대응하도록 함을 시사한다. 또한, Qwen-0.6B 모델이 T5 모델보다 높은 성능을 보이는 것($p_{qwk}<0.01$, $d_{qwk}=0.2234$, $p_{f1}<0.05$, $d_{f1}=0.2009$)은 텍스트 생성 및 추론에 특화된 디코더 기반 모델이 텍스트 이해와 생성의 교차로 추론 능력이 상대적으로 약한 인코더-디코더 기반 모델에 비해 타 학기 일반화가 더욱 잘 작동함을 보여준다. 또한, LLaMA-3B와 8B에서 성능이 점진적으로 증가하는 것으로 보아($p<0.01$, $d_{qwk}=0.2297$, $d_{f1}=0.2376$), 디코더 기반 LLM 성능이 파라미터 증가에 따라 향상됨을 확인할 수 있다.

종합하면, 본 결과는 동일한 내용영역에서도 학기 변화가 문항 난이도 예측 성능에 영향을 미친다는 점을 보여주는 동시에, 디코더를 활용하는 LLM이 파인튜닝을 통해 이러한 변화에 강건하게 작동함을 시사한다.

4.2.2 타 내용영역 문항에 대한 난이도 예측 분석

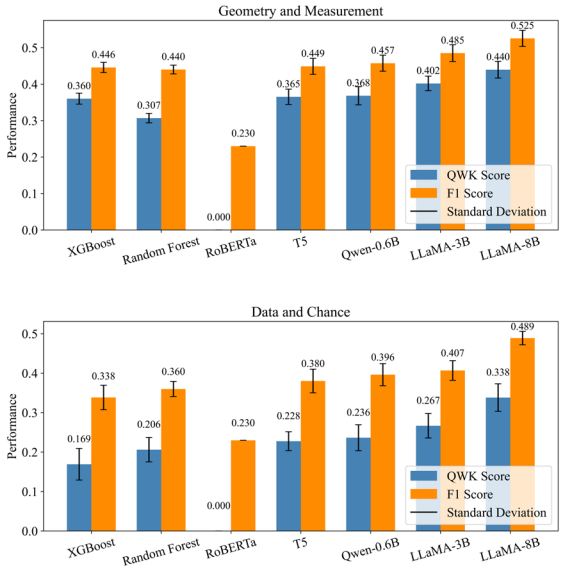

Figure 3. Mean Performance of Math Item Difficulty Prediction on Different Topics in Terms of Quadratic Weighted Kappa (QWK) and F1 Score over 5-Fold Cross-Validation.

Figure 3은 1학기 수와 연산 영역 문항 난이도 데이터셋으로 학습한 모델이, 타 내용영역인 도형과 측정 및 자료와 가능성의 문항 난이도를 얼마나 정확하게 예측하는지 보여준다. 이는 학습 데이터와 내용영역이 상이한 상황에서의 일반화 성능을 분석하기 위한 설정으로, 4.2.1절의 동일 내용영역 내 학기 간 일반화보다 더 높은 수준의 일반화 성능을 확인할 수 있다.

실험 결과, 모든 모델에서 동일 학기·내용영역 문항에 대한 예측 성능뿐만 아니라, 동일 내용영역 내 타 학기 문항에 대한 예측 성능과 비교하더라도 QWK와 F1 점수가 크게 하락하였다. 이는 내용영역의 변화가 문항의 개념적 지식, 문제 해결 방식, 언어적 표현에 실질적인 차이를 유발하며, 이러한 차이가 난이도 예측 일반화를 더욱 어렵게 하는 요인임을 시사한다.

모델별로 살펴보면, 모든 내용영역에서 RoBERTa는 여전히 '중'수준으로의 예측 편향을 보였고, 파인튜닝을 적용한 다른 LLM은 기계학습 기반 모델을 모두 능가하였다. 도

형과 측정 영역에서 T5와 Qwen 모델 모두 기계학습 모델과의 차이가 통계적으로 유의하지 않았고, LLaMA-3B와 8B 모델은 기계학습 모델 중 최고 성능을 보인 XGBoost 모델과 통계적으로 유의함을 보였다(LLaMA-3B vs. XGBoost: $p<0.01$, $d_{qwk}=0.2294$, $d_{f1}=0.2251$, LLaMA-8B vs. XGBoost: $p<0.001$, $d_{qwk}=0.2529$, $d_{f1}=0.2503$). 또한, 자료와 가능성 영역에서는 대표적으로 인코더-디코더 기반 LLM인 T5 및 디코더 기반 LLM 중 최저 성능을 보인 Qwen과 기계학습 모델 중 최고 성능을 보인 Random Forest의 차이가 통계적으로 유의미한 차이를 보였다(T5 vs. Random Forest: $p<0.05$, $d_{qwk}=0.1899$, $d_{f1}=0.1808$, Qwen vs. Random Forest: $p<0.05$, $d_{qwk}=0.2114$, $d_{f1}=0.2193$).

또한, 이전 결과들과 일관되게 디코더 기반 LLM이 다른 비교모델을 능가하여 높은 추론 능력이 타 내용영역 문항에서의 강건한 예측에도 중요하게 작동하는 것을 확인하였다. 그러나, 이는 상대적으로 높은 성능이며, 인코더-디코더 기반과 디코더 기반 LLM이 모든 내용영역에서 QWK 기준 0.44, F1 점수 기준 0.53 이하의 낮은 결과를 보였다. 이것은 디코더를 활용하는 LLM이 학습 데이터와 동일한 학기 및 내용영역, 또는 타 학기로의 일반화 상황에서는 실제 교육 현장의 1차 문항 평가 단계에서 활용될 수 있는 가능성을 보여주지만, 내용 간 일반화 상황에서는 '상-중'혹은'상-하' 구분과 같은 난이도 차이 구분이 어렵고 레이블 불균형 영향도 많이 받기 때문에 관련 문제 극복을 위한 추가 연구가 필요함을 시사한다.

종합하면, 내용영역의 변화는 문항 난이도 예측에서 일반화의 어려움을 크게 증가시키는 요인으로 작용하며, 이러한 조건에서도 파인튜닝을 적용한 디코더 기반 LLM이 상대적으로 높은 예측 성능을 보였다. 그러나, 여전히 교육 현장에서 활용되기에는 성능이 제한되며, 이를 향상시키기 위한 후속 연구가 필요하다.

4.3 디코더 기반 LLM의 수학 문항 난이도 예측 과정에서의 주의집중 양상

4.3.1 디코더 기반 LLM의 주의집중 단어에 대한 교육과정 내용체계 핵심어 포함 양상 분석

Figure 4는 앞서 수학 문항 난이도 예측에서 가장 높은 성능을 보인 디코더 기반 LLM이 파인튜닝 전후 입력된 수학 문항 텍스트에서 주의집중한 가중치(0.0-1.0)를 기준으로 내림차순으로 정렬 후 Top-K개 단어 가중치 합의 누적 분포를 제시한다. Figure 4에 따르면 상위 15개 단어부터는 그 누적 가중치가 0.75를 넘고, 상위 20개 단어부터는 누적 가중치의 증가가 크지 않은 것을 확인할 수 있다. 이것은 파인튜닝된 LLM이 상위 20개 단어에 특히 집중하는 것을 나타내며, 이에 따라 본 절에서는 상위 10개, 15개, 20개의 단어를 추출하였을 때, 초등학교 5학년 수학 교육과정의 수와 연산 영역 내용체계와 성취기준에 제시된 핵심어가 포함된 단어의 평균 비율(막대 그래프)과 Table 5에 제시된 전체 핵심어 유형 중 각 Top-K 단어군에 나타나는 핵심어 유형의 평균 비율(꺾은선 그래프)을 Figure 5와 같이 파인튜닝 전후 LLM별로 비교하였다. 이를 통해 모델이 문항 난이도 예측 과정에서 교육과정의 핵심 개념을 참조하는 정도를 분석하였다. 이는 초등학교 수학 문항의 난이도와 문항 특성 간 관계를 분석한 선행연구에서도 교육과정의 내용체계와 성취기준을 바탕으로 각 문항에 나타나는 개념별 내용요소를 문항 특성으로 정의하여 교과 전문가가 문항 텍스트 내 각 문항 특성의 출현 빈도를 코딩하여 난이도와의 관계를 분석한 방법에 착안하여 진행하였다[11, 12].

분석 결과, Figure 5에서 확인할 수 있듯이, 추출된 Top-K 단어의 개수가 증가할수록 교육과정 핵심어 포함 비율과 Top-K 단어군에 나타나는 핵심어 유형의 수가 상승하는 경향이 관찰되었다. 이는 디코더 기반 LLM이 더 많은 단어를 참조할수록, 교육과정의 핵심 개념을 더욱 포괄적으로 참조하고 있음을 시사한다. 또한, 수학 문항 난이도 예측 성능이 높아질수록 모든 Top-K 설정에서 교육과정 핵심어를 포함한 비율과 주의집중한 핵심어 유형의 다양성이 일관되게 높았고, 파인튜닝을 적용한 경우가 미적용한 경우에 비해 교육과정 핵심어에 주의집중하는 비율이 더욱 높고, 다양한 유형의 핵심어에 주의집중하는 경향이 나타났다.

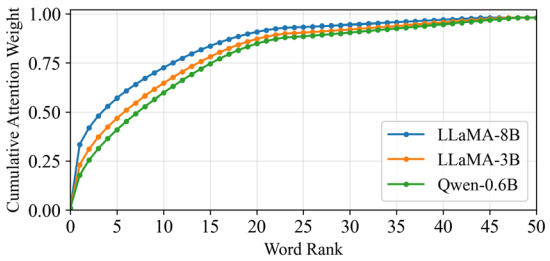

Figure 4. Cumulative Attention Weight Distribution over Token Ranks for LLMs

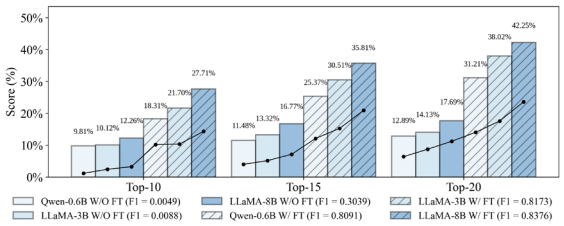

Figure 5. Item Difficulty Prediction Performance (F1) and Average Coverage of Key Curriculum Terms within the Top-K Attention Words of LLMs. Bars Represent the Proportion of Key Terms among the Top-K Attention Words, while Lines Indicate the Ratio of Distinct Key Terms Covered among Key Terms Attended by LLMs.

해당 결과는 선행연구에서 수학 데이터를 활용한 파인튜닝을 통해 LLM의 수학 문제 풀이 능력이 향상되고[41, 42], 수학 문제 풀이 관련 피드백 생성 능력이 개선된다는 보고를 확장한다[43]. 즉, 수학 문항 난이도 데이터 기반 파인튜닝은 LLM이 난이도 예측 관련 단서에 더욱 집중하여 난이도를 정확하게 예측하게 한다.

다만, 이러한 LLM의 주의집중 분석은 난이도 예측 과정에 대한 부분적인 단서를 제공하는 것으로, LLM이 교과교육 관점에서 어떠한 근거에 의해 수학 문항의 난이도를 예측하는지 체계적인 분석을 위해서는 주의집중 데이터와 함께 각 문항에 내재된 형태적 특성 및 핵심역량과 같은 다양한 요소를 바탕으로 전문가 검토에 의한 분석이 후속 연구로 진행될 필요가 있다.

4.3.2 LLM의 주의집중 단어 사례 분석

Table 8과 9(부록 Table 11, 12 한국어 버전 참조)는 1학기 수와 연산 영역의 테스트 문항을 대상으로, 본 연구의 다양한 실험에서 가장 높은 성능을 보인 LLaMA 계열 모델에 파인튜닝을 적용 및 미적용한 경우 각 언어모델이 난이도 예측 과정에서 주의집중한 상위 20개 단어를 비교한 사례 분석 결과를 제시한다. 본 사례는 4.3.1절의 정량적 분석에서 관찰된 교육과정 핵심어 포함 비율의 차이가 실제 텍스트 수준에서 어떻게 나타나는지 정성적으로 확인하기 위한 목적에서 분석되었다. 모델의 토크나이저(tokenizer)는 모델이 사전학습된 방식에 따라 토큰화하기 때문에 주의집중된 토큰이 한국어 단어 단위와 맞지 않을 수 있지만, 각 토큰을 한국어 의미 단위에 연결시켜서 분석하였다.

Table 8. Case Study 1 of Top-20 Attention Words of LLMs (Topic: Number and Operations)

Category	Case
Item Text	● Topic: Understanding Multiples ● Problem: A number from 1 to 60 is placed in □ so that the result of the calculation becomes a multiple of 6. How many numbers can be correctly placed in □? 312 + □ ● Sample Answer: Since 312 ÷ 6 = 52, 312 is a multiple of 6, and thus □ must also be a multiple of 6. Among the numbers from 1 to 60, the multiples of 6 are 6 × 1 = 6, 6 × 2 = 12, ·····, 6 × 10 = 60. Therefore, there are 10 numbers in total.
LLaMA-3B (w/o fine-tuning)	**multiples of**, therefore, 6, sample answer, be, 312, and, **numbers**, how many, **multiple**, of 6, thus □, also, ·····, 312 is, so, among, 1 to, problem, topic
LLaMA-3B (w/ fine-tuning)	10, total, **312 + □**, ·····, **10 = 60**, topic, **6 = 52**, to 60, **multiple**, how many, correctly, □, thus □, among, **multiples of**, 312 is, and, ÷, **1 = 6**, 1 to
LLaMA-3B (w/ fine-tuning)	10, are, also, **10 = 60**, □, **312 + □**, 312, and, 312 is, **multiples of**, therefore, 6, correctly, sample answer, how many, to 60, 1 to, topic, problem, thus □
LLaMA-8B (w/ fine-tuning)	**6 = 52, 312 + □**, 10, total, **10 = 60**, ×, thus □, 312 is, ·····, ÷, 1 to, 312, **multiple**, **1 = 6, 2 = 12**, **multiple of**, to 60, how many, **multiples of**, **calculation**
Key Curriculum Terms	number, natural number, fraction, numerator, denominator, decimal, addition, subtraction, multiplication, division, factor, multiple, round up, round down, round, number range, mixed operations, calculation, +, −, ×, ÷, =, 〉, 〈

Table 9. Case Study 2 of Top-20 Attention Words of LLMs (Topic: Number and Operations)

Category	Case
Item Text	● Topic: Making Equivalent Fractions Using Multiplication ● Problem: Read the dialogue and write the name of the person who made an incorrect statement. Seonwoo: 2/6 and 1/3 have the same value. Yejin: If you multiply both the numerator and the denominator of 3/5 by 7, you can make an equivalent fraction, 21/35 . Sangsu: You can make an equivalent fraction by multiplying both the numerator and the denominator by the any one and the same number. ● Sample Answer: To make equivalent fractions, both the numerator and the denominator must be multiplied by the same non-zero number.
LLaMA-3B (w/o fine-tuning)	dialogue, **multiplication**, value, can, 2/6, Sangsu, Yejin, same, incorrect, name, using, topic, making, 21/35, any, read, Seonwoo, make, write, problem
LLaMA-3B (w/ fine-tuning)	to, make, write, can, dialogue, Seonwoo, 2/6, **multiplication**, incorrect, 21/35, non-zero, multiply, read, **denominator**, 1/3, 3/5, person, Sangsu, **fraction**, Yejin
LLaMA-8B (w/o fine-tuning)	any, value, 3/5, **multiply**, 2/6, **fraction**, can, read, 3/5, Sangsu, dialogue, incorrect, Seonwoo, problem, Yejin, person, 7, make, same, topic
LLaMA-8B (w/ fine-tuning)	2/6, 21/35, 3/5, non-zero, any, **multiply**, person, to, can, **denominator**, 1/3, **numerator**, value, **fraction**, **number**, Sangsu, **fractions**, write, **multiplying**, Yejin
Key Curriculum Terms	number, natural number, fraction, numerator, denominator, decimal, addition, subtraction, multiplication, division, factor, multiple, round up, round down, round, number range, mixed operations, calculation, +, −, ×, ÷, =, 〉, 〈

교육과정 내용체계의 핵심어 목록과 비교 결과, 파인튜닝을 적용한 LLaMA-8B의 주의집중 단어군에는 '배수(multiples)', '곱셈(multiplication)', '분수(fraction)', '분모(denominator)', '분자(numerator)', '+', '×', '÷', '=' 등 내용체계 핵심어들이 파인튜닝을 적용한 LLaMA-3B에 비해 더 많은 비중으로 나타나는 것으로 확인되었다. 또한, 8B 모델의 주의집중 단어군에 더 다양한 유형의 핵심어가 포함되는 것을 관찰되었다. 이러한 차이는 앞선 정량 분석에서 확인된 교육과정 핵심어 포함 비율의 차이와 일관되며, 난이도 예측 성능이 높은 모델일수록 교육과정 기반 개념을 더 많이 그리고 더 다양하게 참조함을 보여준다. 또한, 파인튜닝을 미적용한 LLaMA 모델은 파인튜닝을 적용한 경우에 비해 파라미터 규모에 관계없이 주의집중한 단어군에 교육과정 핵심어가 포함되는 비율과 그 다양성이 모두 낮았다. 이것은 앞서 논의한 바와 같이 수학 문항 난이도 예측 과업에서 단순히 대규모 언어 데이터로 사전학습되어 습득된 LLM의 언어 이해 및 추론 능력에만 의존하는 것이 아니라 관련된 교육 데이터로 파인튜닝을 적용할 필요가 있음을 시사한다.

본 사례 분석은 LLM의 국내 초등 5학년 수학 문항 난이

도 예측 과정에서, 관련 데이터로 파인튜닝을 적용한 LLM이 난이도 판단 과정에 유용한 교육과정 핵심어에 주의집중하는 경향이 강화됨을 확인시켜 준다.

5. 결론 및 논의

본 연구는 초등학교 5학년 수학 문항을 대상으로 LLM을 활용한 문항 난이도 자동화 예측의 가능성과 한계를 다각도로 분석하였다. 특히 기존의 기계학습 기반 접근과 LLM 기반 접근을 비교하고, 학기 및 내용영역 변화에 따른 일반화 성능과 LLM의 주의집중 양상을 함께 분석함으로써, 수학 문항 난이도 예측에서 LLM의 교육적 활용 가능성을 체계적으로 검증하였다.

연구 결과는 다음과 같다. 먼저, 연구 문제 1과 관련하여 파인튜닝을 적용한 LLM은 디코더를 활용하는 LLM이 기계학습 기반 모델과 인코더 기반 LLM에 비해 학습 데이터와 동일한 범위의 문항 난이도 예측 과업에서 높은 QWK와 F1 점수를 보였으며, 그중 높은 추론 능력을 활용하는 디코더 기반 LLM이 가장 높은 성능을 보였다. 또한, 성능이 높아질수록 학습 데이터에 내재된 레이블 불균형 문제를 더 잘 극복하며 한 단계 수준의 미세한 난이도 차이를 더 잘 구분하는 것이 관찰되었다. 다음으로, 연구 문제 2와 관련하여 학습에 포함되지 않은 타 학기 및 내용영역 문항을 대상으로 한 일반화 실험에서도 디코더가 모델 구조에 포함된 LLM을 파인튜닝하는 것이 기계학습 기반 모델과 인코더 기반 LLM에 비해 더 높은 난이도 예측 성능을 보였다. 마지막으로, 연구 문제 3과 관련하여 연구 문제 1과 2에서 가장 높은 성능을 보인 디코더 기반 LLM의 문항 텍스트 주의집중 양상을 분석한 결과, 난이도 예측 성능이 높을수록 주의집중하는 상위 단어군에 교육과정 핵심어가 포함되는 비율이 더 높았고 모델이 주의집중하는 핵심어의 다양성 또한 높았다. 이러한 결과는 도메인 데이터 기반 파인튜닝을 통해 주의집중 기법을 활용하는 LLM의 문항 난이도 예측 성능이 향상되고, 기존의 기계학습 기반 방법을 능가한다는 점에서 LLM 파인튜닝 효과를 분석한 선행연구[24, 25]의 결과와 일부 일치하지만, 이전 연구에서 높은 난이도 예측 정확도를 보인 인코더 기반 LLM은 본 연구의 데이터셋에서 심각한 예측 편향을 보였으며 이를 극복하기 위한 추가적인 연구가 필요하다.

본 연구의 결과가 수학 문항 난이도 자동화 예측과 관련하여 제공하는 시사점은 다음과 같다. 먼저, 수학 문항 난이도 데이터로 파인튜닝된 디코더 기반 LLM은 난이도 예측에서 단순한 성능 향상을 넘어, 교육적으로 의미 있는 개념 단서에 근거한 판단을 수행할 수 있음을 보여준다. 이는 디코더 기반 LLM이 교육 평가 자동화에서 신뢰 가능한 도구로 활용되기 위해 요구되는 교육적 타당성의 한 측면을 정량적으로 뒷받침한다. 또한 학습에 포함되지 않은 타 학기 문항에 대해서도 QWK와 F1 점수가 0.55 이상을 기록하였다. 이는 제한된 학기 데이터로 파인튜닝한 디코더 기반 LLM이 동일 내용영역 내의 다양한 상황으로 일반화될 수 있는 가능성을 시사한다. 그러나, 타 내용영역으로의 일반화 성능은 내용영역에 따라 상이한 결과가 관찰되었고 다른 실험 환경에 비해 상대적으로 낮았기 때문에 이에 대한 추가적인 방법론 탐색이 필요하다.

그럼에도 불구하고, 약 3,000개 규모의 상대적으로 작은 데이터로 파인튜닝된 디코더 기반 LLM이 예측한 수학 문항 난이도는 본 연구의 학습 데이터와 동일한 내용영역 및 실험 환경을 전제로 할 때, 학습 지원 과정에서 학생에게 제공할 문항의 사전 선별 단계에 활용될 수 있다. 이는 전문가가 심층적으로 분석·검토해야 할 문항의 범위를 부분적으로나마 축소하여, 검토 과정에서 요구되는 시간과 비용을 완화하는 데 기여할 가능성을 시사한다. 이는 최근 주목받고 있는 문항 생성 자동화 연구와의 결합 가능성을 보여주며, 생성된 문항의 질이 확인되지 않은 상황에서 본 연구의 결과물을 활용하면 전문가 검토 이전 단계에서 학습 데이터와 일치하는 내용영역 범위에서의 문항 난이도 기반 사전 필터링이 가능하다[44, 45].

나아가 교사가 학급 내의 학생들에게 개별화된 수준별 평가를 구성할 때, 디코더 기반 LLM이 자동으로 예측한 난이도를 기반으로 사전에 제공된 문제 후보군에서 교사 전문성을 바탕으로 학생 개개의 특성에 맞는 문항을 선별하여 평가지를 구성하는 방법은 본 연구에서 LLM의 난이도 예측 성능의 한계를 교사전문성으로 극복함과 동시에 교사가 모든 과정을 직접 수행해야 하는 현장의 어려움도 일부 완화할 수 있을 것이다.

위와 같은 시사점과 동시에 본 연구는 몇 가지 한계점을 지닌다. 먼저, 분석 대상이 초등학교 5학년 수학 문항으로 한정되어 있어 연구 결과를 다른 학년이나 다른 교과로 일반화하는 데에는 추가 검증이 필요하다. 이에 따라 향후 연구에서는 다양한 학년과 교과의 문항에 대해서 본 연구에서 제안한 접근법의 적용 가능성을 확장적으로 검증할 필요가 있다. 다음으로, 실험 과정에서 디코더 기반 LLM의 일반화 성능이 일부 내용영역에서는 충분히 높지 않은 것으로 나타났으며, 이는 내용영역의 차이에 관계없이 안정적인 성능을 확보할 수 있는 보다 강건한 디코더 기반 LLM 활용 방법의 필요성을 시사한다. 세번째로, 본 연구에서 활용된 데이터셋에는 난이도 간 불균형이 내재하는 한계가 있다. 이에 따라 파인튜닝을 적용한 다양한 LLM의 테스트 결과에서도 난이도 예측의 정량적 결과와 별개로 Figure 2와 같이 양극단 난이도 범주와 인접 난이도의 예측에서 높은 빈도의 오류를 보인다. 이것은 학습 데이터셋과 본 연구 대상이 된 LLM 모델들의 근본적인 한계로 인해 데이터 불균형으로 인한 예측 편향 문제가 완전히 해결되지는 않았음을 보여주며 이를 해결할 수 있는 보완된 방법론의 필요성을 시사한다. 네 번째로, 디코더 기반 LLM의 주의집중 가중치 분석은 모델의 내부 판단 과정을 해석하는 하나의 간접적 단서로서 의미를 가지지만, 주의집중이 곧 완전한 설명을 제공한다고 보기 어

렵다. 따라서 향후 연구에서는 주의집중 분석과 함께 근거 생성 또는 반사실적(counterfactual) 분석과 같은 다양한 해석 기법을 결합할 필요가 있다. 끝으로, 파인튜닝을 위한 소규모 데이터 구축에도 비용이 발생하는 문제를 해결하기 위해 최근 구글에서 발표한 교육용 LLM인 LearnLM[46]이나 거대추론모델(large reasoning model)에 사고사슬방식(chain-of-thought, CoT)이나 검색증강생성방식(retrieval augmented generation, RAG)의 고수준 프롬프팅을 적용하여 파인튜닝 없이 난이도를 정확하게 예측할 수 있는 방법에 대한 연구가 필요하다.

종합하면, 본 연구는 다양한 LLM의 수학 문항 난이도 예측 과업에서 파인튜닝의 효과를 분석하였고, 디코더 기반 LLM이 난이도 예측 과정에서 교육과정 핵심어에 주의집중하는 양상을 분석하여 LLM이 수학 교과 관련 개념을 일부 반영한 예측을 수행함을 확인하였다. 이는 문항 난이도 예측 자동화 연구에서 정확성과 함께 교육적 타당성과 설명 가능성을 동시에 고려한 연구 방향의 출발점으로, LLM 기반 평가 시스템의 실질적 활용을 위한 기초적 근거가 될 수 있을 것이다.

참고문헌

[1] AlKhuzaey, S., Grasso, F., Payne, T., & Tamma, V. (2024). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education, 34*(3), 862-914. https://doi.org/10.1007/s40593-023-00362-1

[2] Brennan, R. (2023). *Educational measurement*. Bloomsbury Publishing PLC.

[3] Cizek, G., & Earnest, D. (2015). Setting performance standards on tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (pp. 212-237). Routledge. https://doi.org/10.4324/9780203102961

[4] von Davier, M., Fishbein, B., & Kennedy, A. (2024). *TIMSS 2023 Technical Report (Methods and Procedures)*. Boston College, TIMSS & PIRLS International Study Center. https://timss2023.org/methods/

[5] Mojoyinola, M., Kehinde, O., & Tang, J. (2025). Enhancing Item Difficulty Prediction in Large-scale Assessment with Large Language Model. *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con)*, Pittsburgh, USA, 218-222.

[6] Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In R. Kadosh & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 1118–1134). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.014

[7] Gupta, U., & Zheng, R. (2020). Cognitive Load in Solving Mathematics Problems: Validating the Role of Motivation and the Interaction among Prior Knowledge, Worked Examples, and Task Difficulty. *European Journal of STEM Education, 5*(1), 5. https://10.20897/ejsteme/9252

[8] Štěpánek, L., Dlouhá, J., & Martinková, P. (2023). Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics, 11*(19), 4104. https://doi.org/10.3390/math11194104

[9] Yi, X., Sun, J., & Wu, X. (2024). Novel feature-based difficulty prediction method for mathematics items using XGBoost-based SHAP model. *Mathematics, 12*(10), 1455. https://doi.org/10.3390/math12101455

[10] Razavi, P., & Powers, S. (2025). *Estimating Item Difficulty Using Large Language Models and Tree-Based Machine Learning Algorithms.* https://doi.org/10.48550/arXiv.2504.08804

[11] Kim, S.,& Ban, J. (2025). Exploration of Item Features Affecting Item Difficulty and Item Discrimination in Elementary School Mathematics Diagnostic Tests of Basic Skills and Comparison of Prediction Models. *Journal of Educational Evaluation, 38*(1), 57-82. https://doi.org/10.31158/JEEV.2025.38.1.57

[12] Kim, S.,& Ban, J. (2024). An Exploratory Study on Factors Influencing Rasch Item Difficulty and Item Proportion Correct Using LLTM and LLTM-R models: Focusing on 3R's Math Test for 3rd Grade. *Journal of Educational Evaluation, 37*(1), 145-174. https://doi.org/10.31158/JEEV.2024.37.1.145

[13] Peters, S., Zhang, N., Jiao, H., Li, M., & Zhou, T. (2025). Review of Text-Based Approaches to Item Difficulty Modeling in Large-Scale Assessments. *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con)*, Pittsburgh, USA, 37-47.

[14] Webb, N., Alt, M., Ely, R., Cormier, M., & Vesperman, B. (2005). *The WEB alignment tool: Development, refinement, and dissemination*. Washington, DC: Council of Chief State School Officers. https://ccsso.org/

[15] Sheehan, K. (1997). A Tree-Based Approach to Proficiency Scaling and Diagnostic Assessment. *Journal of Educational Measurement, 34*(4), 333–352. http://www.jstor.org/stable/1435113

[16] Turner, R., Dossey, J., Blum, W., & Niss, M. (2013). *Using mathematical competencies to predict item difficulty in PISA: A MEG study*. PISA. https://www.oecd.org/en/about/programmes/pisa.html.

[17] Schneider, M., Huff, K., Egan, K., Gaines, M., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment, 18*(2), 99-121. https://doi.org/10.1080/10627197.2013.789296

[18] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

[19] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, 785–794. https://doi.

org/10.1145/2939672.2939785

[20] Trace, J., Brown, J., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Language Testing, 34*(2), 151-174. https://doi.org/10.1177/0265532215623581

[21] Hou, J., Maximilian, K., Quecedo, J., Stoyanova, N., & Yangarber, R. (2019). Modeling language learning using specialized Elo rating. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Italy, 494-506. https://doi.org/10.18653/v1/W19-4451

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems*, Long Beach, USA. https://doi.org/10.48550/arXiv.1706.03762

[23] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, USA, 4171-4186. https://doi.org/10.18653/v1/N19-1423

[24] Benedetto, L., Aradelli, G., Cremonesi, P., Cappelli, A., Giussani, A., & Turrin, R. (2021). On the application of transformers for estimating the difficulty of multiple-choice questions from text. *Proceedings of the 16th workshop on innovative use of NLP for building educational applications*, Portland, USA, 147-157.

[25] Loginova, E., Benedetto, L., Benoit, D., & Cremonesi, P. (2021). Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Oline, 846-855. https://doi.org/10.26615/978-954-452-072-4_097

[26] Clark, K., Khandelwal, U., Levy, O., & Manning, C. (2019). What does bert look at? an analysis of bert's attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Italy, 276-286. https://doi.org/10.18653/v1/W19-4828

[27] Wiegreffe, S., & Pinter, Y. (2019). Attention is not not Explanation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, China, 11–20. https://doi.org/10.18653/v1/D19-1002

[28] Meta. (2024). *The llama 3 herd of models.* https://doi.org/10.48550/arXiv.2407.21783

[29] Open AI. (2023). *Gpt-4 technical report*. https://doi.org/10.48550/arXiv.2303.08774

[30] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent abilities of large language models*. https://doi.org/10.48550/arXiv.2206.07682

[31] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research, 21*(140), 1-67. https://dl.acm.org/doi/10.5555/3455716.3455856

[32] Google. (2023). *Gemini: a family of highly capable multimodal models*. https://doi.org/10.48550/arXiv.2312.11805

[33] Park, S., Moon, J., Kim, S., Cho, W., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J., & Cho, K. (2021). KLUE: Korean Language Understanding Evaluation. *Proceedings of the Advances in Neural Information Processing Systems*, Online, https://doi.org/10.48550/arXiv.2105.09680

[34] Park, D. (2022). *Pko-t5: PAUST Korean T5 for text-to-text unified framework* [computer software]. https://github.com/paust-team/pko-t5

[35] Qwen Team. (2023). *Qwen technical report*. https://doi.org/10.48550/arXiv.2309.16609

[36] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *An introduction to statistical learning. Springer* https://dl.acm.org/doi/book/10.5555/2517747

[37] Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research, 13*(1), 1063-1095. https://dl.acm.org/doi/10.5555/2188385.2343682

[38] Kaplan, J., McCandlish, S., Henighan, T., Brown, T., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models*. https://doi.org/10.48550/arXiv.2001.08361

[39] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, USA, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

[40] Wu, H., Xu, K., Song, L., Jin, L., Zhang, H., & Song, L. (2021). Domain-Adaptive Pretraining Methods for Dialogue Understanding. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Thailand, 665–669. https://doi.org/10.18653/v1/2021.acl-short.84

[41] Macina, J., Daheim, N., Hakimi, I., Kapur, M., Gurevych, I., & Sachan, M. (2025). Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, China, 204-221. https://doi.org/10.18653/v1/2025.emnlp-main.11

[42] Azerbayev, Z., Schoelkopf, H., Paster, K., Santos, M., McAleer, S., Jiang, A., Deng, J., Biderman, S., & Welleck, S. (2024). Llemma: An open language model for mathematics. *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Austria.

https://doi.org/10.48550/arXiv.2310.10631
[43] Dinucu-Jianu, D., Macina, J., Daheim, N., Hakimi, I., Gurevych, I., & Sachan, M. (2025). From Problem-Solving to Teaching Problem-Solving: Aligning LLMs with Pedagogy using Reinforcement Learning. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, China, 272-292. https://doi.org/10.18653/v1/2025.emnlp-main.15
[44] Yeung, C., Lee, J., & Tsou, B. (2019). Difficulty-aware distractor generation for gap-fill items. *Proceedings of the 17th annual workshop of the Australasian language technology association*, Australia, 159-164.
[45] Gao, Y., Bing, L., Chen, W., Lyu, M., & King, I. (2018). Difficulty controllable generation of reading comprehension questions. *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, China, 4968-4974. https://doi.org/10.48550/arXiv.1807.03586
[46] Google. (2024). *Learnlm: Improving gemini for learning*. https://doi.org/10.48550/arXiv.2412.16429

추성엽

· 2016년 청주교육대학교(교육학사, 수학심화)
· 2021년 청주교육대학교 초등영어교육전공(교육학석사)
· 2016년~2019년 충청북도교육청 공립 초등교사
· 2019년~2022년 서울특별시교육청 공립 초등교사
· 2022년~현재 한국과학기술원 데이터사이언스대학원 박사과정

관심분야 : AI 기반 교육, 데이터 중심 AI, LLM 활용 자동 평가, LLM 에이전트, 인간-컴퓨터 상호작용
chseye7@gmail.com

부 록

Table 5, Table 8, Table 9의 한국어 번역본

Table 5. Key Curriculum Terms

Key Terms
수, 자연수, 분수, 분자, 분모, 소수, 덧셈, 뺄셈, 곱셈, 나눗셈, 약수, 배수, 올림, 버림, 반올림, 수의 범위, 혼합 계산, 계산, +, −, ×, ÷, =, 〉, 〈

Table 9. Case Study 2 of Top-20 Attention Words of LLMs (Topic: Number and Operations)

Category	Case
Item Text	● 주제: 배수 알아보기 ● 문제: □ 안에 1에서 60까지의 자연수를 넣어 다음 계산 결과가 6의 배수가 되게 하려고 합니다. □ 안에 알맞은 수는 모두 몇 개입니까? 312 + □ ●모범답안: 312 ÷ 6 = 52에서 312가 6의 배수이므로 □도 또한 마찬가지로 6의 배수이어야 합니다. 따라서 1에서 60까지의 수 중에서 6의 배수는 6 × 1 = 6, 6 × 2 = 12, ……, 6 × 10 = 60 이므로 모두 10개입니다.
LLaMA-3B (w/o fine-tuning)	**배수이므로**, 따라서, 6, 모범답안, 합니다, 312, 또한, **수는**, 개입니까, **배수이어야**, 6의, □도, 마찬가지로, ……, 312가, 하려고, 중에서, 1 에서, 문제, 주제
LLaMA-3B (w/ fine-tuning)	10개입니다, 모두, **312+□**, ……, **10 = 60이므로**, 주제, **6 = 52에서**, 60까지의, 배수이어야, 개입니까, 알맞은, □, □도, 중에서, **배수이므로**, 312가, 또한, ÷, **1 = 6,** 1에서
LLaMA-8B (w/o fine-tuning)	10개입니다, 입니다, 마찬가지로, **10 = 60이므로**, □, **312 + □**, 312, 또한, 312가, **배수이므로**, 따라서, 6의, 알맞은, 모범답안, 개입니까?, 60까지의, 1에서, 주제, 문제, □도
LLaMA-8B (w/ fine-tuning)	**6 = 52에서**, **312+□**, 10개입니다, 모두, **10=60이므로**, **×**, □도, 312가, …… , ÷, 1에서, 312, **배수이어야**, **1 = 6**, **2 = 12**, **배수는**, 60까지의, 개입니까?, **배수이므로**, **계산**
Key Curriculum Terms	수, 자연수, 분수, 분자, 분모, 소수, 덧셈, 뺄셈, 곱셈, 나눗셈, 약수, 배수, 올림, 버림, 반올림, 수의 범위, 혼합 계산, 계산, +, −, ×, ÷, =, 〉, 〈

Table 8. Case Study 1 of Top-20 Attention Words of LLMs (Topic: Number and Operations)

Category	Case
Item Text	● 주제: 곱셈을 이용하여 크기가 같은 분수 만들기 ● 문제: 대화를 읽고 잘못 말한 사람의 이름을 쓰시오. - 선우 : 2/6와 1/3은 크기가 같아. - 예진 : 3/5의 분모와 분자에 7을 곱하면 크기가 같은 분수인 21/35을 만들 수 있어. - 상수 : 분모와 분자에 어떤 수든지 같은 수를 곱하면 크기가 같은 분수를 만들 수 있지.' ● 모범답안: 분모와 분자에 0이 아닌 같은 수를 곱해야 크기가 같은 분수를 만들 수 있습니다
LLaMA-3B (w/o fine-tuning)	대화를, **곱셈을**, 크기가, 있지, 2/6, 상수, 예진, 같은, 잘못, 이름을, 이용하여, 주제, 만들기, 21/35, 어떤, 읽고, 선우, 만들, 쓰시오, 문제
LLaMA-3B (w/ fine-tuning)	있습니다, 만들, 쓰시오, 있어, 대화를, 선우, 2/6, **곱셈을,** 잘못, 21/35, 0이, **곱해야**, 읽고, **분모**., 1/3, 3/5, 사람의, 상수, **분수**, 예진
LLaMA-3B (w/ fine-tuning)	어떤, 크기가, 3/5, **곱해야**, 2/6, **분수인**, 있지, 읽고, 3/5, 상수, 대화를, 잘못, 선우, 문제, 예진, 사람의, 7을, 만들, 같은, 주제
LLaMA-8B (w/ fine-tuning)	2/6, 21/35, 3/5, 0이, 어떤, **곱해야**, 사람의, 있습니다, 있지, **분모와**, 1/3, **분자에**, 크기가, **분수인**, **수를**, 상수, **분수를**, 쓰시오, **곱하**면, 예진
Key Curriculum Terms	수, 자연수, 분수, 분자, 분모, 소수, 덧셈, 뺄셈, 곱셈, 나눗셈, 약수, 배수, 올림, 버림, 반올림, 수의 범위, 혼합 계산, 계산, +, −, ×, ÷, =, 〉, 〈