Paper Proposal

논문 발표

GPT 기반 정보 교과 프로그래밍 영역을 위한 객관식 문제 출제 프로그램 개발

7팀	2025.06.13

목차

Content

Part 1

서론 & 이론적 배경

정보 과목에서의 비효율적인 시험 출제 방식

1) 시간과 노력 소모

교사가 수동으로 문제를 출제하는
과정은 많은 시간과 노력이 소요

정보 과목은 이론보다 실습위주
→ 적합한 지필평가 문항 개발 어려움

2) 신뢰성 및 효율성 부족

정보 과목은 타 과목에 비해
기출문제 부족으로
문제 출제 가이드라인 취약

→ 시험의 신뢰성↓
→ (교사입장) 시간적 효율성↓

3) 학습 동기 저하

학생 개개인의
이해도와 성취도 반영 X

획일화된 평가 방식?
→ 학습 동기↓

정보 과목에서의 비효율적인 시험 출제 방식

교사에만 의존하는 출제 방식?
AI 활용한 새로운 평가 패러다임

1) 시간과 노력 소요

교사가 수동으로 문제를 출제하는
과정은 많은 시간과 노력이 소요

정보 과목은 이론보다 실습위주
→ 적합한 지필평가 문항 개발 어려움

2) 신뢰성 및 효율성 부족

정보 과목은 타 과목에 비해
기출문제 부족으로
[illegible]약

→ 시험의 신뢰성↓
→ (교사입장) 시간적 효율성↓

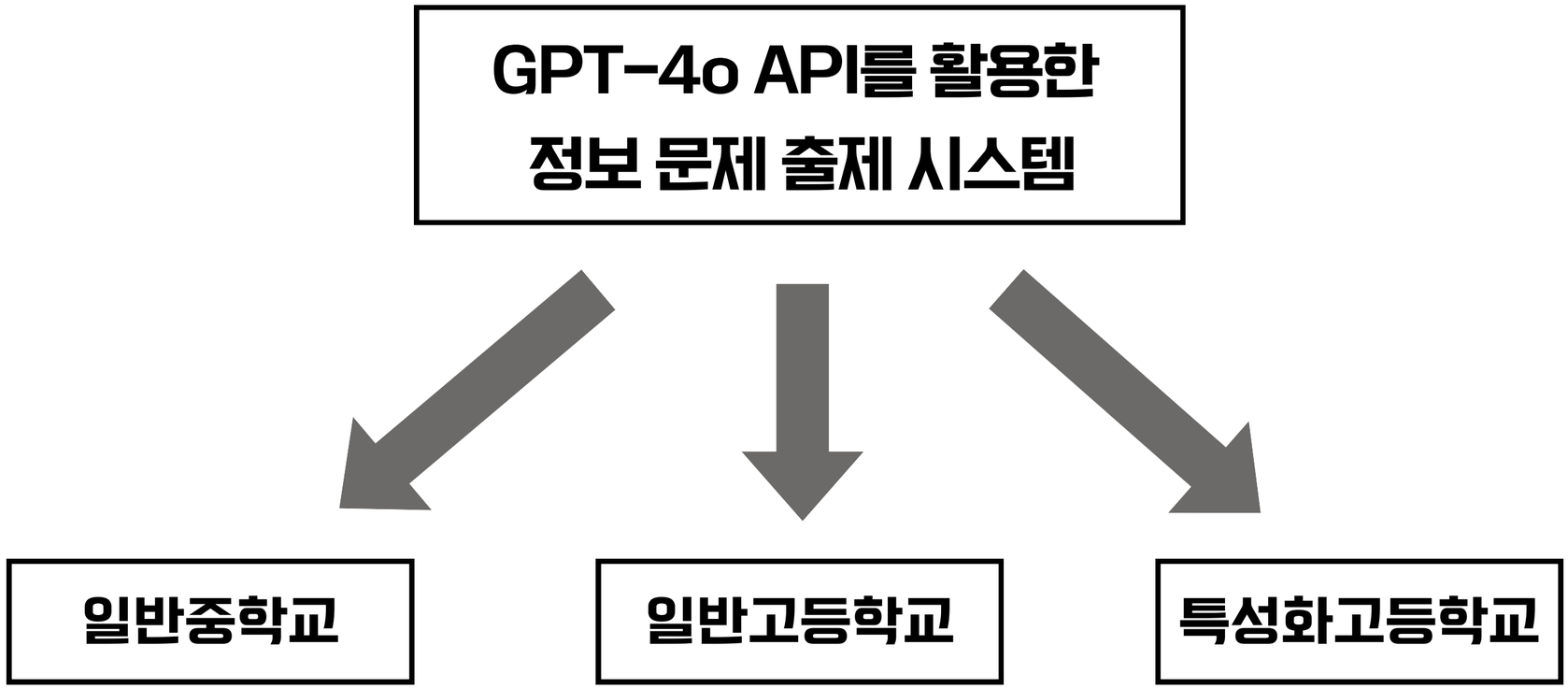
3) 학습 동기 저하

학생 개개인의
이해도와 성취도 반영 X

획일화된 평가 방식?
→ 학습 동기↓

"개인화된 문제 출제"
"자동으로 채점하는 시스템"

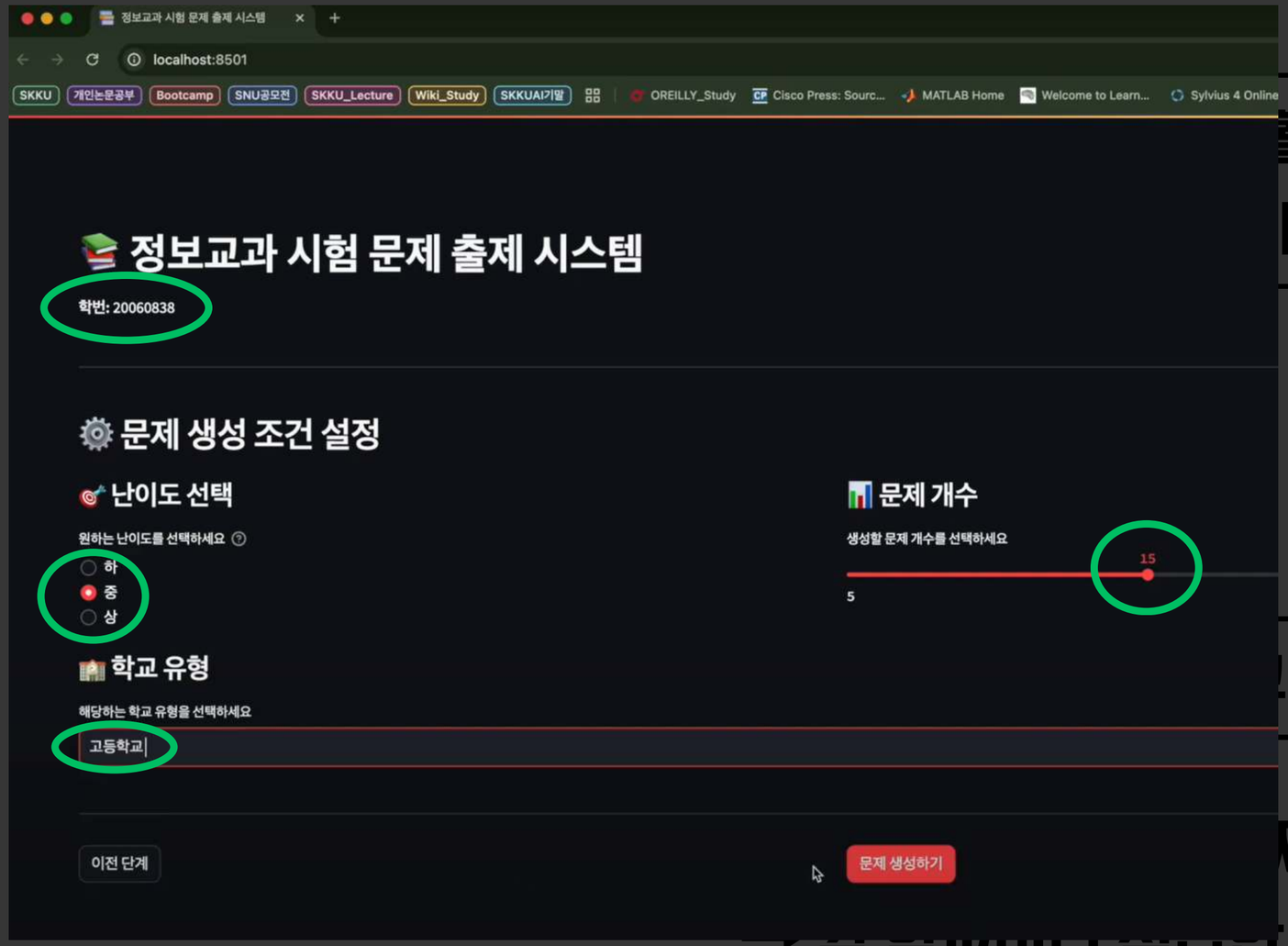
서론 – 연구 목적 및 범위



GNN 모델을 활용하여 문제 난이도 분류

→ 각 학교마다 적합한 문제 선별

서론 – 연구 목적 및 범위



활용한

스템

Streamlit으로 만든 웹 애플리케이션 UI

1) 교사 : 문제 출력 및 평가

2) 학생 : 피드백 제공

문제 난이도 평가가 정확한가?

교사 & 학생의 피드백 순환 구조 적절?

현장에서 적용이 가능한가?

제 난이도 분류

문제 선별

선행 연구 및 배경

김슬기(2023)

Chat GPT를 개념, 예제, 연습문제,
코드 검토 등에 적용

단순히 코드가 일치하는지 비교
프로그래밍 결과,
입출력이 같은지 판별

특히, 정보 교과 학습에서 유의미한
보조 역할 할 수 있음을 증명

김창석(2024)

교육용 AI 플랫폼으로
유의미한 효과 보임

AI 기술 활용 교육이
실제로 적용 가능함 증명

학생별 배경지식을 고려한
맞춤형 AI 교육 플랫폼 제공

학년별로 학습 효과 다름 증명

선행 연구 및 배경

김슬기(2023)

Chat GPT를 [illegible]
코드 검토 등에 적용

단순히 코드가 일치하는지 비교
프로그래밍 결과,
[illegible]

특히, 정보 교과 학습에서 유의미한
보조 역할 할 수 있음을 증명

김창석(2024)

교육용 AI 플랫폼으로
유의미한 효과 보임

AI 기술 활용 교육이
실제로 적용 가능함 증명

[illegible]
맞춤형 AI 교육 플랫폼 제공

학년별로 학습 효과 다름 증명

AI 를 적절한 방향으로 활용?

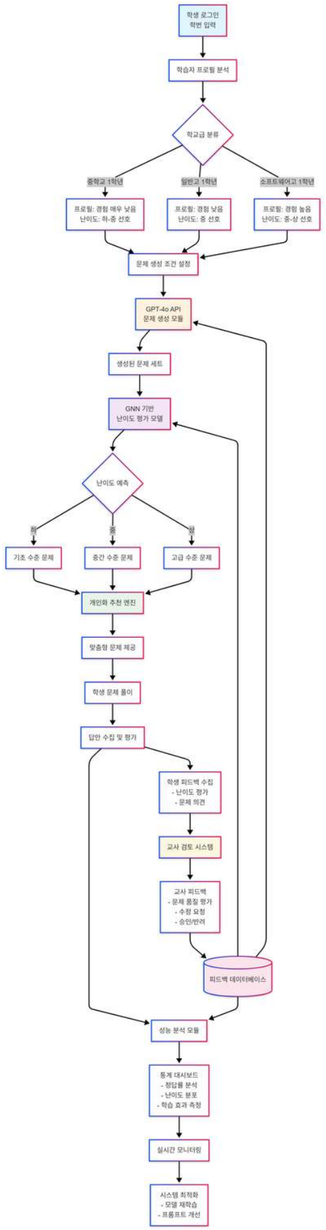
→ 문제 출제 이외에 평가까지 가능

이러한 배경 활용?

→ 자동평가와 개인화 기능 수행하는 시스템

Part 2

핵심 기술



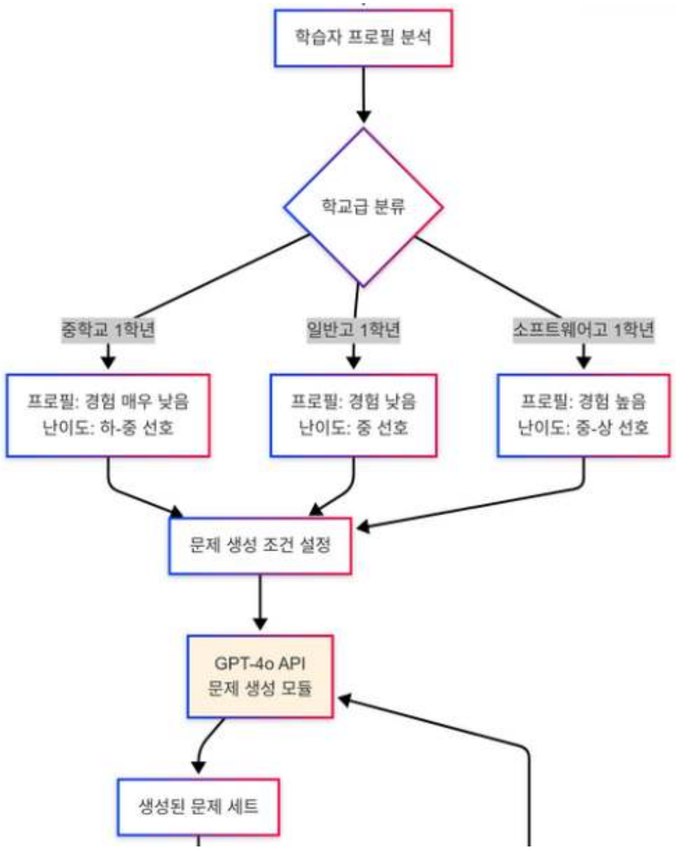
시스템 전체 흐름

단계 1) 학습자 특성을 분석하여 문제 조건 설정

단계 2) GPT-4o API를 통한 문제 생성

단계 3) GNN 기반 모델로 생성된 문제의 난이도 평가

단계 4) 학습자 특성에 적합한 문제 제공 및 피드백



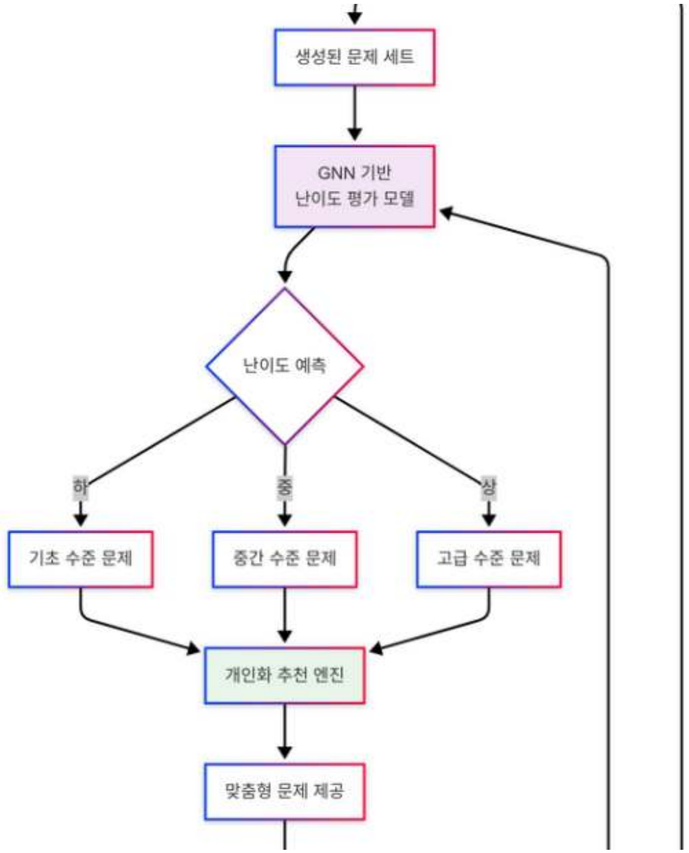
프롬프트 엔지니어링

프롬프트 :

"다음 조건에 맞는 정보 과목 문제를 {num_questions}개 생성해줘.
- 학교 유형: {school_type}
- 난이도: {difficulty}
- 대상 학습자 특성: {learner_profile}
- 형식: JSON 배열 구조"

추가 조건 :

- 중학교: "프로그래밍 첫 접촉, 기본 개념 위주"
- 일반고: "기초 프로그래밍 경험, 개념과 응용 혼합"
- 특성화고: "풍부한 경험, 심화 알고리즘 중심"



GNN 기반 난이도 평가 모델

GNN?

- GNN이란 Graph Neural Network으로, 그래프 형태의 데이터를 처리하는 데에 적합한 인공신경망임.
- 그래프 형태로 노드와 엣지를 가지고 있음.

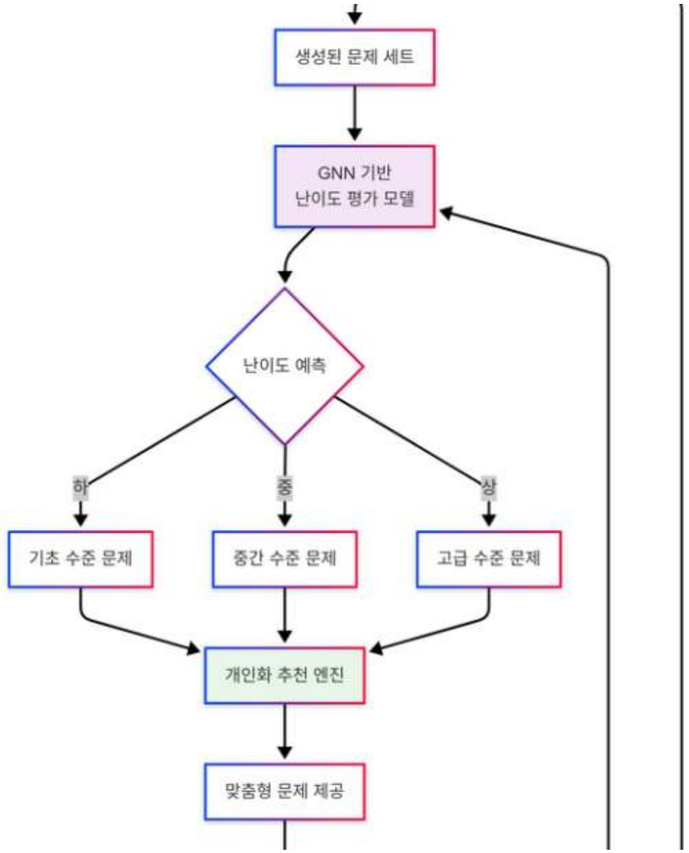
그래프 구조(무방향 그래프)

노드 생성:

- 생성한 문제를 노드로 지정함
- 노드는 문제에서 특성을 추출하여 특성 벡터의 값을 가지고 있음
- 단일 노드의 경우 자기 자신으로의 루프를 추가하여 그래프 구조를 유지하도록 함

엣지 생성:

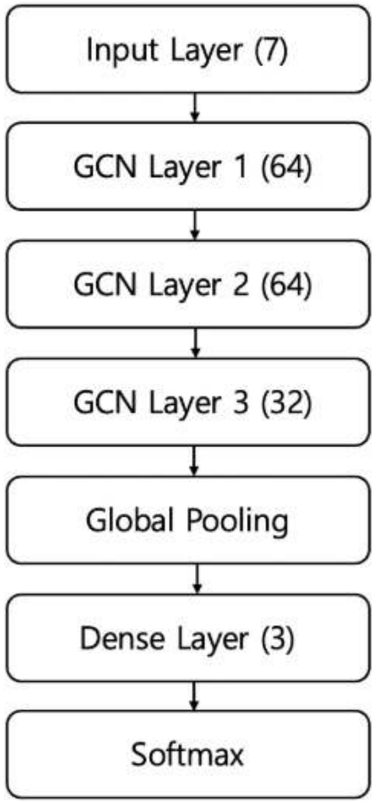
- 코사인 유사도를 기반으로 함
- 유사도가 0.7보다 크면 엣지를 생성함



GNN 기반 난이도 평가 모델

노드의 특성 벡터(8차원으로 구성)

- 텍스트 길이 (정규화: /100)
- 단어 수 (정규화: /50)
- 프로그래밍 키워드 밀도
- 평균 선택지 길이 (정규화: /20)
- 단원별 가중치 (정규화: /3)
- 코드 포함 여부 (0 또는 1)
- 수식 포함 여부 (0 또는 1)
- 복잡도 점수 (정규화: /10)



GNN 기반 난이도 평가 모델

훈련 데이터 :
기존의 JSON 파일 기반으로 한 데이터에 피드백 데이터를 점진적으로 추가함
하이퍼 파라미터 :

- 에포크 수: 200회
- 최적화 알고리즘: Adam Optimizer (learning_rate=0.01)
- 손실 함수: Negative Log-Likelihood Loss (NLLLoss)
- 정규화: Dropout (rate=0.2, 과적합 방지)
- 배치 처리: 전체 그래프 단일 배치 방식
- 성능 모니터링: 50 에포크마다 손실값 출력 및 수렴 확인

Part 4

연구 결과

개요

5.1. AI 문제 출제 및 난이도 분류의 신뢰도 평가

AI 생성 문제의 품질과 정답 정확성을 검증하고, GNN 기반 난이도 예측 모델의 성능을 평가

5.2. 학습 효과 분석

AI 개인화 문제를 활용한 학습 집단의 성취도, 만족도 및 학습 동기 변화를 분석

5.3. 실패 사례 및 문제점

AI 시스템의 문제 생성 오류 및 실제 사용자 피드백을 통해 한계점과 개선 방향을 도출

5.1 문제 생성 품질

문제 생성 품질: 교사 평가

평가 항목	평균 점수	표준편차	최고점	최저점
문제 명확성	4.3	0.6	5	3.2
난이도 적절성	4.1	0.7	5	2.8
교육과정 연계성	4	0.8	5	2.5
전체 품질	4.1	0.7	5	2.8

→ AI 생성 문제의 전반적 품질은 우수 (평균 4.1점)

→ 단, 교육과정 연계성 점수가 가장 낮아, 교사의 추가적인 검토 및 수정 필요성을 시사

5.1 GNN 기반 난이도 평가

GNN 모델의 안정성

- **모델 안정성: 약 150 epoch 지점에서 손실 값이 안정화되며 성공적으로 수렴 (과적합 방지)**
- **예측 효율성: 동일 문제 반복 예측 시 일관성 확보**
- **추론 속도: 문제당 평균 0.03초**

학교급	정확도	정밀도	재현율	F1-score
중학교	0.743	0.721	0.756	0.738
일반고	0.789	0.776	0.801	0.788
특성화고	0.834	0.825	0.843	0.834
전체	0.789	0.774	0.8	0.787

→ 프로그래밍 경험이 많은 특성화고에서 가장 높은 예측 정확도(83.4%) 달성

→ 중학교에서 상대적 오류가 발생, 초급 학습자의 비일률적 문제 인식 패턴이 원인으로 추정

5. 2 학습 효과 분석

정답률 비교

그룹	사전 평가	사후 평가	향상도	t-value	p-value
중학교 통제군	48.20%	52.10%	3.90%	1.34	0.189
중학교 실험군	47.80%	56.30%	8.50%	2.87	0.008
일반고 통제군	62.10%	65.40%	3.30%	1.21	0.234
일반고 실험군	61.70%	69.80%	8.10%	2.94	0.006
특성화고 통제군	78.30%	80.10%	1.80%	0.89	0.378
특성화고 실험군	77.90%	82.40%	4.50%	2.12	0.041

→ 프로그래밍 경험이 많은 특성화고에서 가장 높은 예측 정확도(83.4%) 달성

→ 중학교에서 상대적 오류가 발생, 초급 학습자의 비일률적 문제 인식 패턴이 원인으로 추정

5. 2 학습 효과 분석

정답률 비교

학교급	난이도	사전 정답률	사후 정답률	상승폭
중학생	중	43.20%	58.70%	+15.5%p
일반고	중	62.80%	76.20%	+13.4%p
특성화고	상	64.10%	76.80%	+12.7%p

→ 중학생과 일반고생은 '중' 난이도 문제에서 가장 큰 성취도 향상을 보임

→ 특성화고생은 '상' 난이도 문제에서 가장 높은 성장률을 보여, 도전적인 과제가 효과적이었음을 시사

5. 2 학습 효과 분석

정답률 비교

학교급	문제 풀이 시간	재도전 의욕	자기효능감 (5점)
중학교	평균 31% 증가	52% → 78% (+26%p)	2.8 → 3.7 (+0.9)
일반고	평균 22% 증가	68% → 84% (+16%p)	3.4 → 4.1 (+0.7)
특성화고	평균 18% 증가	81% → 92% (+11%p)	4.1 → 4.4 (+0.3)

→ 학습 기초 단계인 중학생에게서 가장 두드러진 동기 부여 효과가 나타남 (풀이 시간, 재도전, 효능감 모두 최고 상승)

→ AI가 제공하는 맞춤형 난이도가 학습 몰입과 자신감 향상에 긍정적 영향을 미침

5.3 실패 사례 및 문제점

문제 생성 실패 사례 및 유형

- **논리적 오류 사례: 파이썬 변수 선언 문제에서 문법적으로 틀린 int x = 10을 정답으로 처리하는 오류 발생**
  - **생성된 문제: "다음 중 파이썬에서 변수를 선언하는 올바른 방법은?"**
  - **문제점: int x = 10 (오답)이 정답으로 채점됨**

- **오류 유형별 빈도:**
  - **한국어 표현 어색함: 12.4%**
  - **정답 자체의 오류: 7.3%**
  - **선택지 중복: 4.1%**
  - **문제 의도 불분명: 3.8%**

→ 언어적 자연스러움이 가장 시급한 개선 과제

→ 한국어 특화 AI 모델 활용 시 개선 기대 (하이퍼클로바X)

5. 3 실패 사례 및 문제점

학습자 부정 피드백

[중학교 학생]

"문제가 너무 어려워서 포기하고 싶어요" (23%)”

"설명이 이상해요" (19%)

"답이 틀린 것 같아요" (15%)

→ 초급자에겐 난이도가 높고, 문제 완성도가 부족함

[고등학교 학생]

"너무 쉬워서 시간 낭비 같아요" (34%)”

기존 문제은행이 더 나아요" (28%)

"AI가 만든 티가 너무 나요" (21%)"

→ 숙련자에겐 낮이도가 낮고, 인공적인 느낌이 강함

Part 4

결론

4-1
한계점 및 개선 방안

4-2
결론 및 시사점

남정훈

한계점

1. 데이터 신뢰성 및 현장성 부족
   - 더미 데이터 사용: 실제 학생 데이터를 수집하지 않고 "실제 학습자 특성을 반영한 더미 데이터"를 사용함. 이는 외적 타당도(일반화 가능성)에 한계.
2. API 의존성과 기술적 불안정성
   - GPT-4o API에 대한 높은 의존도 → 서비스 지연, 비용, 실패율(8.7%) 발생.
3. AI 문제 품질 편차
   - AI가 생성한 문제의 표현 오류, 중복, 문맥 불명확성 등 오류율이 최대 12.4% 발생.
4. 이론적·정책적 배경의 약점
   - Bloom, SOLO, UDL, 개인화 학습 이론 등과의 구체적 연결 부족.
   - 교육 정책 또는 AI 디지털교과서와의 실질적인 연계 및 적용 사례 제안 부족.
5. 분석 방법의 정교함 부족
   - 효과 검증에서 Cohen's d, ANOVA 등 통계 적용은 일부 있지만, 전체적으로 설명력 부족.
   - GNN 예측 정확도 외에 모델 비교나 성능 시각화 등 기계학습적 분석 다양성 부족.

개선 방향

1. 실제 학습자 기반 장기 연구 설계

2. 모델 고도화 및 독립성 확보 : GNN 외에도 XGBoost, LSTM, Transformer 등 다양한 분류 알고리즘과의 비교.

3. 문제 품질 자동 검증 체계 강화
   - 문제 생성 직후 문법 검사, 답안 유효성, 중복 탐지, 난이도 편차 필터링 등 자동 QA 모듈 탑재.
   - 한국어 어색함 개선을 위해 클로바 X, KSS 등 한국어 특화 언어모델 기반 후처리 도입.

4. 교육 이론 기반 재설계
   - 난이도 조절 기준을 Bloom's taxonomy (지식-이해-적용-분석-종합-평가)에 맞추어 다단계화.
   - 개인화 추천 로직을 Vygotsky의 ZPD(근접발달영역) 기반으로 재설계

5. 정책 연계 및 실용적 제안
   - 교육부 'AI 디지털 교과서' 시범학교와 연계한 실증연구로 확대.
   - 에듀테크 기업, 교육청과의 협업 실증 및 B2B 형태 제안서 작성으로 현실 적용력 강화.

결론 및 시사점

사용자 만족도 및 활용 가능성

- 정보 교사 및 예비 교원들이 시스템의 활용성 및 자동 프롬프트 기능에 대해 전반적으로 높은 만족도와 긍정적인 반응을 보였음.
- 다수의 교사(88%)가 실제 시험 출제에 활용할 의향을 나타냄.

사용자 요구사항 반영의 한계

- 반복문과 조건문 형태 등 복잡한 요구사항에 대한 충족도가 낮음.
- 정답의 정확성도 60~70% 수준으로, 실전 활용을 위해 추가적 개선 필요.

문제 출제 성능 개선을 위한 추가 연구 필요

- 향후 모델 성능 향상을 위해 다양한 프로그래밍 코드 데이터 추가 학습 필요.
- 반복문 및 조건문 형태, 사용자 정의 함수 개수 충족률 개선을 위한 학습 데이터 다양화 필요.

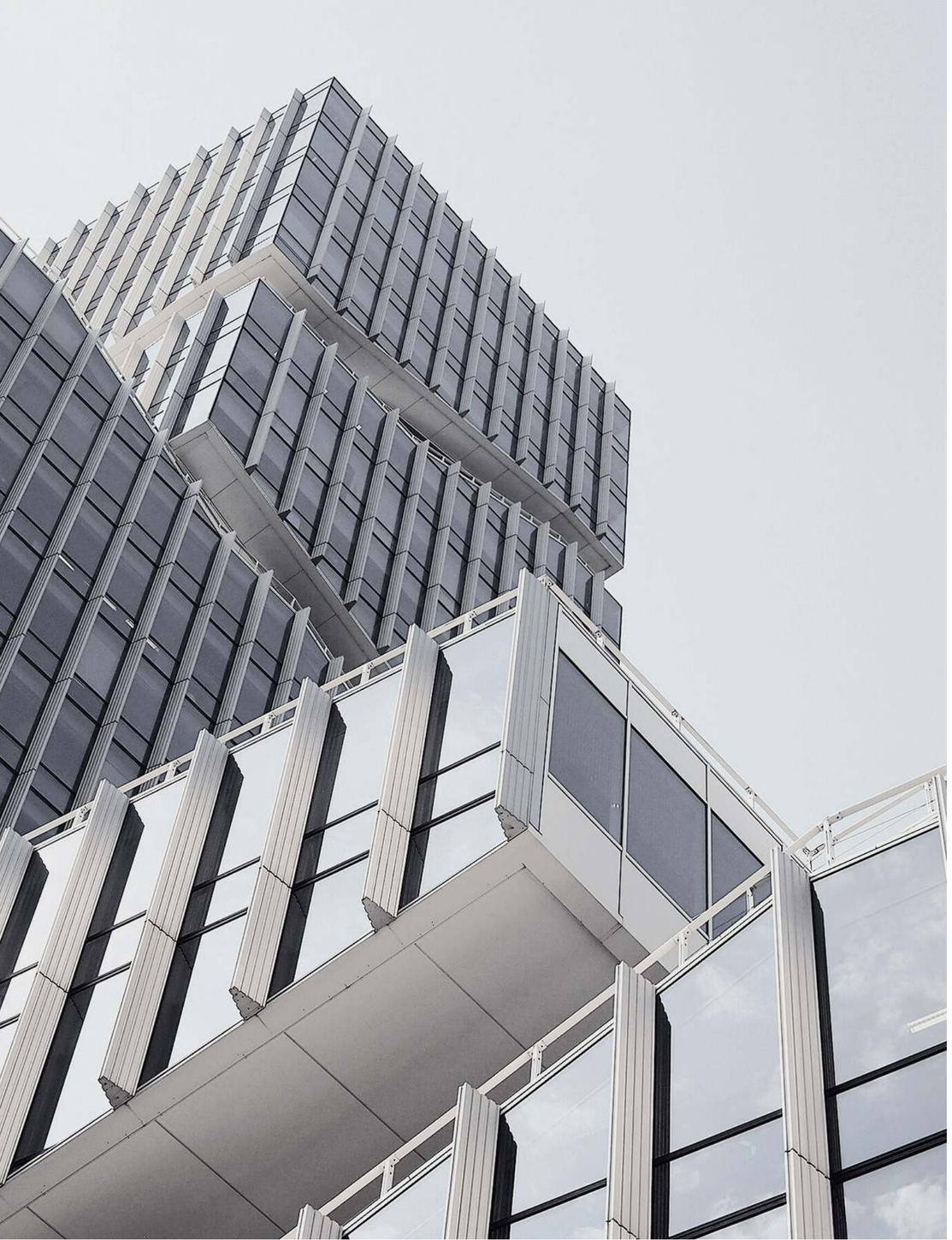
교사들의 업무 부담 경감 효과

- 시스템이 완벽하지는 않지만, 기본적인 문제 생성 과정 자동화를 통해 교사의 출제 부담을 줄이고 효율성을 높일 수 있음.
- 앞으로의 지속적인 개선을 통해 교육 현장에서 사용 가능한 문제 출제 도구로 자리 잡을 가능성이 존재

토론 주제

AI가 만든 문제 vs 선생님이 만든 문제, 어느 쪽이 더 좋을까?

남정훈



감사합니다