석 사 학 위 논 문

수학 학습에서 질문 명료화를 지원하는 AI에이전트 설계 및 개발: 고등학교 2학년 수학적 귀납법 단원 중심으로

김규봉

부산대학교 교육대학원

AI융합교육전공

2026년 2월

수학 학습에서 질문 명료화를 지원하는 AI 에이전트 설계 및 개발: 고등학교 2학년 수학적 귀납법 단원 중심으로

김규봉

2026년 2월

수학 학습에서 질문 명료화를 지원하는 AI에이전트 설계 및 개발: 고등학교 2학년 수학적 귀납법 단원 중심으로

이 논문을 교육학석사 학위논문으로 제출함

김규봉

부산대학교 교육대학원

AI융합교육전공

지 도 교 수 박성호

김규봉의 교육학석사 학위논문을 인준함

2025년 12월 27일

위원장	박성호	인
위 원	송길태	인
위 원	남윤경	인

차 례

표 차례

그림 차례

수학 학습에서 질문 명료화를 지원하는 AI에이전트 설계 및 개발 : 고등학교 2학년 수학적 귀납법 단원 중심으로

김규봉

부산대학교 교육대학원 AI융합교육전공

요약

생성형 AI의 교육 현장 확산에도 불구하고 학생 질문의 낮은 품질이 효과적 학습을 저해하고 있다. 예비조사(385건)에서 학생 질문의 72.3%가 학습 맥락 정보를 제공하지 않았으며, 대부분의 AI 도구가 채택한 즉시 답변 방식은 학생의 사고 과정 확장을 지원하지 못하는 한계를 보였다.

본 연구는 Dewey의 반성적 사고 이론을 기반으로 질문 명료화 프로세스를 핵심으로 하는 AI 에이전트 시스템 MAICE를 설계·개발하고, 고등학교 2학년 수학적 귀납법 단원에 적용하여 학습 효과를 검증하였다. 교사는 도미노 모델, 귀납 과정 연결, 등식/부등식 전략 등의 핵심 개념을 매 수업마다 반복 강조하여 학생들의 공통 언어를 형성하였으며, 이는 AI 학습의 기반이 되었다.

연구 방법: 고등학교 2학년 58명을 명료화 우선 모드(n=28)와 즉시 답변 모드(n=30)에 무작위 배정하여 3주간 A/B 테스트를 수행하였다(284개 유효 세션). 시스템은 Gemini 2.5 Flash를 사용하여 운영되었으며, 평가 방법의 한계를 상호 보완하기 위해 LLM 평가(N=284)와 교사 예비 평가(N=100)를 병행하였다. 질문-답변-맥락 평가 체크리스트를 개발하여 3개 독립 AI 모델(Gemini 2.5 Flash, Claude 4.5 Haiku, GPT-5 mini)이 대규모 패턴 탐색을, 외부 수학 교사 2명이 교육적 타당성 검증을 수행하였다. 평가자 간 신뢰도는 LLM 간 $\alpha=0.868$, 교사 간 $r=0.644$, LLM-교사 간 $r=0.743$(3개 모델 평균)이었다.

주요 결과: LLM-교사 이중 평가를 통해 명료화 효과를 상호 검증하였다. LLM 평가(N=284, 3개 모델 평균)에서 명료화 모드는 학습 지원에서 유의하게 우수하였고(C2: p=0.002, d=0.376), 특히 하위권 학생에게 효과를 보였다(Q1 전체: +2.46점, p=0.033, d=0.511; Q1 C2: p=0.001, d=0.840). 교사 예비 평가(N=100)에서도 동일한 방향성이 관찰되었으며(+2.25점, p=0.031, d=0.307), 하위권에서 큰 효과(+6.91점, p=0.009, d=1.117)가 일치하였다. LLM-교사 간 높은 상관(r=0.743, 3개 모델 평균)으로 상호 검증에 성공하였다. 학생 설문(n=40)에서 60%가 명료화 방식을 선호하여 객관적 평가와 일치하였다. LLM 평가는 충분한 표본으로 패턴을 발견하였고, 교사 평가는 예비 수준(평가자 2명)이나 패턴의 교육적 타당성을 확인하였다. 후속 연구에서 교사 평가 확대(평가자 10명 이상, 표본 300개 이상)가 필요하다.

결론: 질문 명료화 프로세스는 학습 지원을 향상시키며, 특히 하위권 학생의 교육 격차 해소에 기여할 가능성을 확인하였다. 본 연구는 Dewey의 이론을 AI 교육 도구로 구현하였으며, LLM-교사 이중 평가 상호 검증 모델을 제시하여 대규모 객관적 평가와 전문가 타당성 검증을 조합하였다. 교사 주도 프롬프팅 설계가 가능한 확장 가능한 연구 플랫폼을 제시하였다.

주요어: 질문 명료화, AI 에이전트, 반성적 사고, 수학적 귀납법, 멀티에이전트 시스템, Dewey, 교육 격차 해소, 교사 주도 연구, 프롬프팅 설계

I. 서론

1. 연구의 필요성

가. 수학 교육에서 생성형 AI 활용의 확산과 한계

대규모 언어 모델(Large Language Model, LLM)의 비약적 발전에 따라 ChatGPT, Claude 등 생성형 AI 도구가 교육 현장에 빠르게 도입되고 있다. 특히 수학 교육 영역에서는 학생들이 개념 이해, 문제 풀이, 추가 학습 등의 목적으로 LLM을 활용하는 빈도가 급증하고 있으며, 일부 교사들은 이를 보조 학습 도구로 활용하려는 시도를 하고 있다.

그러나 실제 교실 현장에서 LLM을 보조교사로 활용하려는 시도 과정에서, 교육적 효과성에 대한 근본적인 문제들이 지속적으로 관찰되었다. 현재 대부분의 LLM 기반 AI 도구(ChatGPT, Claude, Gemini 등)는 학생의 질문에 대해 질문의 질이나 맥락과 무관하게 즉각적으로 답변을 제공한다. 이러한 즉시 답변 제공 방식은 신속한 정보 제공이라는 장점이 있으나, 학습자가 스스로 질문을 구조화하고 사고를 확장하는 과정을 지원하지 못하여 깊이 있는 학습을 저해할 가능성이 제기되고 있다.

나. 학생 질문의 질적 문제: 실증 분석

본 연구에서는 MAICE 시스템 설계에 앞서, 현재 일반적인 LLM 사용 방식의 교육적 문제점을 실증적으로 파악하기 위해 2024년 5월 중순(학기 초)에 예비 조사를 실시하였다. 예비 조사는 다음의 순차적 과정으로 진행되었다:

1단계: 데이터 수집 - 고등학교 수학 수업 환경에서 학생들에게 ChatGPT와 동일한 사용자 인터페이스를 제공하는 간단한 웹앱을 배포하였다. 수업 시간 동안 학생들이 자유롭게 질문하고 AI 답변을 받을 수 있도록 하여 총 385건의 실제 질문-답변 쌍을 수집하였다.

2단계: 문제점 파악 및 루브릭 구성 - 수집된 데이터를 분석한 결과, 학생 질문과 AI 답변에서 반복적으로 나타나는 질적 문제 패턴들이 관찰되었다. 이러한 문제점들을 명료화하고 그 존재 여부를 객관적으로 판단하기 위해, 전통적 교육학 이론(Dewey의 반성적 사고, Bloom의 교육목표분류학 등)과 최신 AI 교육 평가 연구를 종합하여 체계적인 분

석적 채점 기준(analytic rubric)을 개발하였다.

관찰된 구체적 문제 패턴에 대응하여 6개 평가 영역을 구성하였다.

[표 1-1] 예비조사 루브릭 6개 평가 영역 구성

구분	평가 영역	코드	평가 내용
질문 평가	수학적 전문성	A1	수학 개념의 정확성, 교과과정 내 위치, 용어 사용의 적절성
	질문 구조화	A2	질문 대상·범위·초점의 명확성, 구체적 어려움 제시
	학습 맥락 적용	A3	학습자 수준, 선수 지식, 학습 목적, 이해 상태 명시
AI 답변 평가	학습자 맞춤도	B1	학습자 수준 파악, 선수 지식 연계, 난이도 조절, 개인화 된 피드백
	설명의 체계성	B2	개념의 위계적 설명, 단계별 논리 전개, 핵심 요소 강조, 예시 활용
	학습 내용 확장성	B3	심화학습 방향 제시, 응용문제 연계, 오개념 교정, 자기주도 학습 유도

각 영역은 5점 리커트 척도(1=매우 부족 ~ 5=매우 우수)로 평가되며, 교사가 6개 영역 각각에 대해 전체적인 판단을 바탕으로 1~5점을 직접 부여한다. 평가를 돕기 위해 각 영역별로 4개의 세부 평가 요소(총 24개)를 참고 가이드로 제시하였으나, 최종 점수는 교사의 종합적 판단에 따른다. 질문 영역 15점, 답변 영역 15점으로 총 30점 만점이다.

3단계: 동료교사 평가 실시 - 중등 수학교사 4명(교직 경력: 1년차 2명, 5년차 1명, 7년차 1명, 평균 3.5년)이 독립적으로 385건의 질문-답변을 6개 영역 각각에 대해 1~5점으로 평가하였다.

4단계: 평가 결과 분석 및 주요 발견 - 교사 루브릭 평가 결과, 일반적인 LLM 사용 방식의 구체적 문제점들이 실증적으로 확인되었다:

[표 1-2] 예비조사 루브릭 평가 결과 (질문 및 AI 답변)

구분	평가 영역	평균 점수 (5점 만점)	최저점(1점) 비율	중앙값
질문 평가	수학적 전문성(A1)	2.277	45.50%	2
	질문 구조화(A2)	2.049	45.80%	2
	학습 맥락 적용(A3)	1.5	72.30%	1
AI 답변 평가	학습자 맞춤도	2.474	27.60%	3
	설명의 체계성	2.765	23.90%	3
	학습 내용 확장성	1.832	48.90%	2

[표 1-3] 학생 질문 문제 유형별 분석

문제 유형	해당 영역	발생 비율	사례 수	주요 문제점
맥락 없는 질문	학습 맥락	72.30%	184건	학습자 수준·목적 미제시
모호한 질문	질문 구조화	45.80%	25건	질문 대상·범위 불명확
수학 무관 질문	수학적 전문성	45.50%	26건	학습 집중도 저하

1) 맥락 없는 개념 설명 요청

○ 실제 사례 (ID 294): "근데 어떻게 증명한거야?"

- 교사 평가: 수학적전문성 1점, 질문구조 1점, 학습맥락 1점 (질문 총점 3.0/15점)
- AI 답변: "어떤 것을 증명했는지 구체적으로 말씀해 주시면 더 자세히 답변드리겠습니다. 일반적으로 수학적 또는 과학적 증명은 논리적 추론과 증거를 사용하여..."

○ 문제점: 어떤 증명을 지칭하는지, 어떤 부분이 이해되지 않는지, 학습자가 어느 수준인지 전혀 파악할 수 없어 AI도 추가 정보를 요청할 수밖에 없는 구조. 이러한 질문에서는 AI가 적절한 난이도와 설명 방식을 선택할 수 있는 근거가 부족하여, 학습자 맞춤형 답변 생성이 구조적으로 불가능하다.

2) 모호하고 불명확한 질문

◯ 실제 사례들:

- (ID 142) "내 증명에대해서 피드백해줘" - 어떤 증명을, 어느 부분을 평가받고 싶은지 불명확
- (ID 47) "a+b = c+d" - 단순 수식 제시로 질문 의도 파악 불가, AI는 영어로 수식 설명만 제공
- (ID 315) "토푸리아 마카체프 5분 5라운드 라이트급에서 챔피언 벨트를 두고 싸우면 누가 이길 것 같아? 승자를 말해줘" - 수학과 무관하며 질문 형태도 불명확

◯ 문제점: 질문의 대상, 범위, 초점이 불명확하여 AI가 학생이 실제로 무엇을 알고 싶어하는지 파악할 수 없음

3) 수학과 무관한 질문

◯ 실제 사례들:

- (ID 128) "반갑다" - 단순 인사, AI도 "어떻게 도와드릴까요?" 응답만 제공 (평균 질문 총점 3.0/15점)
- (ID 119) "발해의 건국과 발전을 정리해줘" - 역사 질문으로 수학과 완전 무관
- (ID 344) "수학으로 된 노래가 있어?" - 수학 교육과 관련 없는 호기심 질문

◯ 문제점: 수학 학습 목적에서 완전히 이탈하여 수업 집중도를 저하시키며, 제한된 수업 시간에서 AI 교육 도구의 효과성을 훼손

4) 질문 품질의 극명한 편차

흥미롭게도, 매우 불량한 질문들과 대조적으로 일부 학생들은 매우 구조화되고 맥락이 풍부한 질문을 제시하기도 하였다.

○ 우수 질문 실제 사례 (ID 358):

- "나는 소프트웨어마이스터고에 다니는 고등학교 2학년이야 현재 수학을 공부하고 있는데 0!이 1인 이유를 알기 쉽게 알려줘."
- 교사 평가: 수학적전문성 4점, 질문구조 3점, 학습맥락 5점

○ 우수한 점:

- 학습자의 신분과 수준 명시 (소프트웨어마이스터고 2학년)
- 현재 학습 상황 설명 (수학 공부 중)
- 구체적 질문 (0!이 1인 이유)
- 설명 방식 요청 (알기 쉽게)

이러한 질문 품질의 큰 편차(최저 3.0점 vs 최고 10.0점)는 학생들이 AI에게 효과적으로 질문하는 방법에 대한 명시적 교육이 필요함을 시사한다.

다. AI 답변의 질적 문제: 교사 루브릭 평가 결과

학생 질문에 대한 LLM의 답변을 동일한 교사 패널이 개발한 루브릭으로 평가한 결과, 심각한 교육적 한계가 확인되었다.

[표 1-4] AI 답변 문제 유형별 분석

문제 유형	해당 영역	평균 점수	최저점 비율	주요 문제점
학습 확장 부족	학습 확장성	1.832	48.90%	심화·응용·자기주도 유도 실패
맞춤화 실패	학습자 맞춤도	2.474	27.60%	수준 파악·난이도 조절 실패
체계성 부족	설명 체계성	2.765	23.90%	교육과정 용어·위계 무시

각 루브릭 영역별로 확인된 구체적인 AI 답변의 문제점은 다음과 같다.

1) 학습자 맞춤도 문제: 맥락 파악 실패

루브릭의 '학습자 맞춤도' 영역은 AI 답변이 학습자 수준별 접근, 선수지식 연계성, 학습 난이도 조절, 개인화된 피드백을 제공하는지 평가한다. 그러나 프리패스 방식 AI는 질문에 학습 맥락이 부족할 때 학습자 수준을 파악하지 못하거나, 심지어 질문의 의미 자체를 잘못 해석하는 문제를 보였다.

◯ 실제 사례 (ID 130): "지수의 확장에 대해 설명해줘"

- AI 답변: "지수의 확장에 대해 설명하자면, 이는 지수가 영역을 확장하고 더 많은 지역이나 새로운 시장으로 진출하는 과정을 의미합니다. 이러한 확장은 일반적으로 비즈니스의 성장 및 발전 전략의 일환으로 이루어지며... 1. 지리적 확장: 새로운 도시, 지역, 국가로 사업을 확장하여..."
- 교사 평가: 4명 모두 학습자적합성 1점 (최저점)

◯ 루브릭 위반 항목:

- 학습자 수준별 접근 실패 (고등학교 수학인지 비즈니스인지 파악 못함)
- 선수지식 연계성 부재 (수학 개념과 무관)
- 학습 난이도 조절 불가능 (맥락 오해로 잘못된 답변)

◯ 문제점: 수학의 '지수(exponent)'와 비즈니스 '지점 확장'을 혼동하여 완전히 무관한 답변 제공

2) 학습자 맞춤도 문제: 학습 난이도 조절 실패

◯ 실제 사례: "로지스틱 회귀가 뭐야?" (고등학교 1학년)

- AI 답변: "로지스틱 회귀는... 고등학교 1학년 수준에서 이해할 수 있도록 조금 더 쉽게 설명하면... odds ratio... 최대우도추정..."

◯ 루브릭 위반 항목

- 학습 난이도 조절 실패 (대학 수준 통계 개념 사용)
- 개인화된 피드백 부재 (고1 수준 고려 없음)

◯ 문제점: "고등학교 1학년 수준"이라고 언급하면서도 실제 내용은 대학 수준의 통계학 개념 포함

3) 설명의 체계성 문제: 교육과정 용어 사용의 부적절성

루브릭의 '설명의 체계성' 영역은 개념 설명의 위계화, 단계별 논리 전개, 핵심 요소 강조, 예시 활용의 적절성을 평가한다. 한국 교육과정에 맞지 않는 용어 사용은 학생의 개념 형성에 혼란을 초래한다.

◯ 실제 사례: "수열이 뭐야?"

- AI 답변: "수열은... 명시적 공식(Explicit Formula) (a_n)을 (n)에 대한 함수로 표현합니다..."

◯ 루브릭 위반 항목:

- 핵심 요소 강조 부적절 (교육과정 비표준 용어)
- 개념 설명의 위계화 실패 (한국 교육과정 체계 무시)

◯ 문제점: 대한민국 수학 교육과정 표준 용어 '일반항' 대신 '명시적 공식(Explicit Formula)' 사용으로 교사-학생 간 의사소통 혼란 초래

4) 학습 내용 확장성 문제: 심화학습 방향 제시 실패

루브릭의 '학습 내용 확장성' 영역은 심화학습 방향 제시, 응용문제 연계성, 오개념 교정 전략, 자기주도 학습 유도를 평가한다.

◯ 실제 사례 (ID 145): "수1의 핵심 부분이 뭐야?"

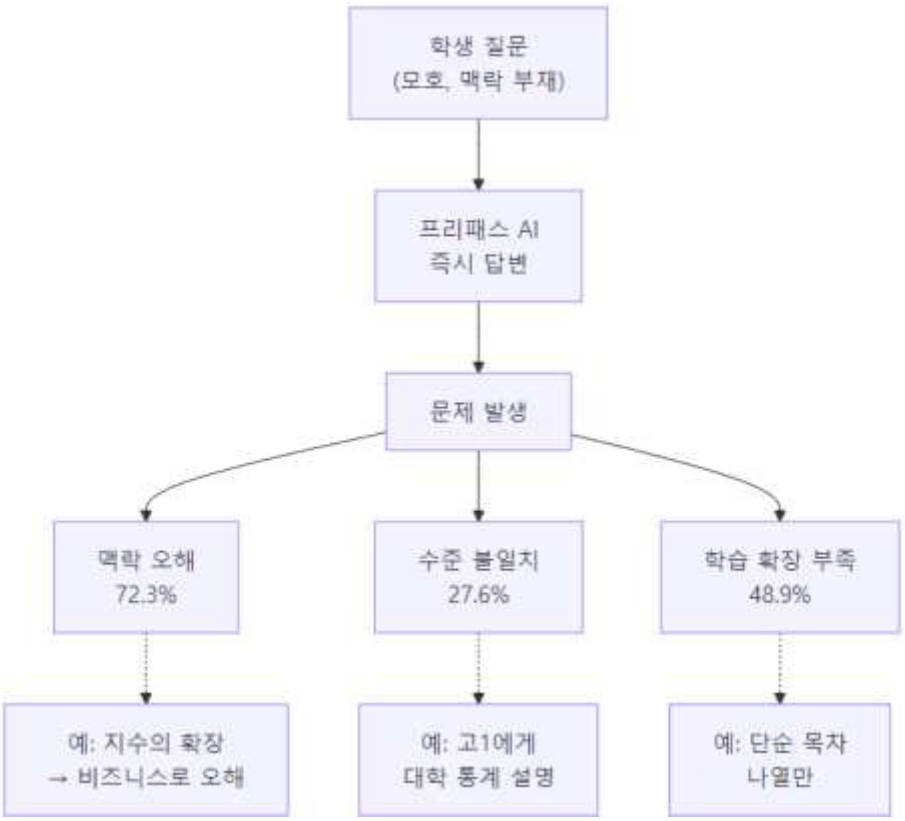
- AI 답변: "수1의 핵심 부분은 고등학교 수학 1에서 다루는 주요 개념들을 의미합니다. 주요 내용은 다음과 같습니다: 1. 다항식: 다항식의 사칙연산, 인수분해, 나머지 정리... 2. 방정식과 부등식: 일차방정식, 이차방정식..."
- 교사 평가: 학습자적합성 3점, 설명체계성 2점, 학습확장성 1점

◯ 루브릭 위반 항목:

- 심화학습 방향 제시 부재 (단순 나열만)
- 응용문제 연계성 없음
- 자기주도 학습 유도 실패 (학생이 스스로 탐구할 방향 제시 없음)

◯ 문제점: 학생이 왜 이 질문을 했는지 파악 없이 관련 개념만 나열. 과도한 정보 제시로 인지 부담 가중

이러한 문제들은 프리패스 방식 AI가 학습자의 현재 이해 수준과 필요성을 고려하지 않고, 심화학습을 유도하거나 자기주도적 학습을 촉진하는 교육적 역할을 수행하지 못함을 보여준다.



[그림 1-1] 프리패스 방식의 문제점 예시

라. 프리패스 방식의 구조적 한계와 MAICE 시스템 설계의 필요성

[표 1-5] 예비조사 종합 결과 및 질문-답변 품질 상관관계

구분	질문 품질	AI 답변 품질	상관관계
평균 점수 (15점)	5.826 (SD 2.925)	7.071 (SD 2.800)	r=0.691***
중앙값	6	7	p<0.001
백분율	38.80%	47.10%	-

추가 분석 결과, 질문 품질과 답변 품질 간에는 강한 양의 상관관계($r=0.691$, $p<0.001$)가 확인되어, 질문이 명료하고 맥락이 풍부할수록 AI 답변의 질도 향상되는 것으로 나타났다.[1)]

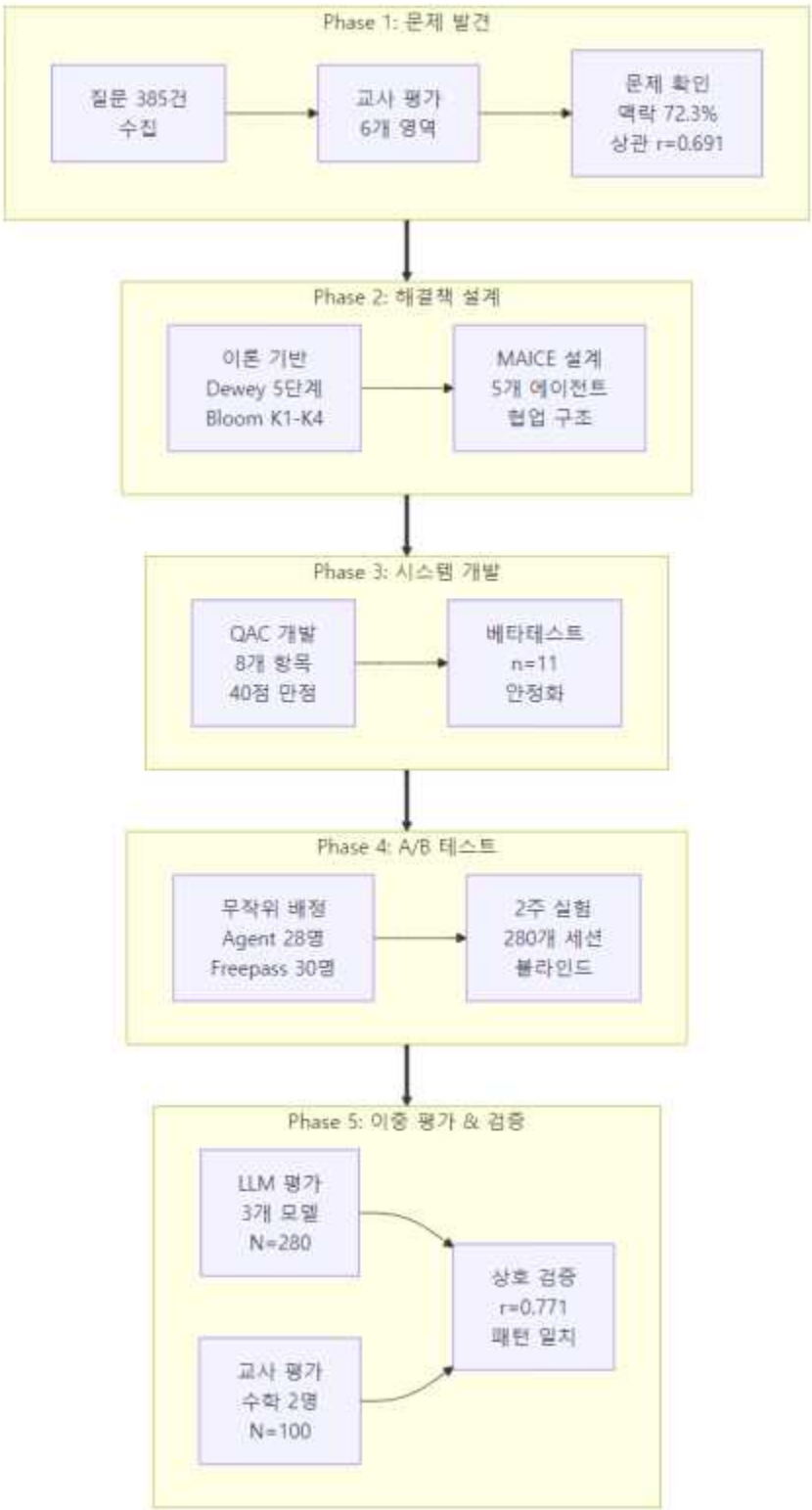
이러한 문제점들은 단순히 개별 사례의 오류가 아니라, 현재 LLM이 채택하고 있는 즉시 답변 제공 방식의 구조적 한계에서 비롯된다. 이 방식은 질문의 질이나 맥락과 무관하게 즉각적인 답변을 제공함으로써, 학습자가 스스로 질문을 명료화하고 구조화하는 과정을 경험할 기회를 제공하지 않는다. 예비 조사에서 관찰된 바와 같이, 이는 학생들의 LLM 활용 만족도를 저하시킬 뿐만 아니라, 실질적인 학습 효과를 얻지 못하는 결과를 초래한다.

특히 수학적 귀납법과 같이 논리적 사고의 구조화와 단계적 추론이 핵심인 단원에서는, 즉시 답변 제공 방식이 학생의 사고 과정을 충분히 유도하지 못하고 피상적 이해에 머무르게 하는 한계가 있다. 수학적 귀납법은 "n=1일 때 성립함을 보이고, n=k일 때 성립한다고 가정하면 n=k+1일 때도 성립함을 보인다"는 명확한 논리 구조를 요구하는데, 단순한 결론 제시만으로는 학생의 개념 이해와 적용 능력을 효과적으로 향상시키기 어렵다.

이에 본 연구는 예비 조사에서 확인된 프리패스 방식의 구조적 한계를 극복하기 위해 질문 명료화 기반 AI agent 시스템 MAICE(Mathematical AI Chatbot for Education)를 개발하고, 무작위 배정 A/B 테스트를 통해 그 효과를 검증하였다. MAICE는 Dewey의 반성적 사고 이론과 Bloom의 지식 분류 체계에 기반하여 학습자의 사고 과정을 깊이 있게 지원하며, 특히 학습 맥락 부재(72.3%)와 학습 확장성 결여(48.9%) 문제를 개선하는 데 중점을 두었다.

---

1) 예비 조사 385건 질문-답변 쌍에서 질문 총점(A1+A2+A3, 15점 만점)과 답변 총점(B1+B2+B3, 15점 만점) 간 Pearson 상관계수를 계산한 결과이다. 특히 질문의 학습맥락 점수와 답변의 학습확장성 점수 간 상관계수는 0.521로, 학생이 자신의 학습 상황을 명확히 제시할수록 AI가 학습을 확장하는 답변을 제공할 가능성이 높아짐을 보여준다.


Phase 1: 문제 발견
질문 385건 수집
교사 평가 6개 영역
문제 확인 맥락 72.3% 상관 r=0.691
Phase 2: 해결책 설계
이론 기반 Dewey 5단계 Bloom K1-K4
MAICE 설계 5개 에이전트 협업 구조
Phase 3: 시스템 개발
QAC 개발 8개 항목 40점 만점
베타테스트 n=11 안정화
Phase 4: A/B 테스트
무작위 배정 Agent 28명 Freepass 30명
2주 실험 280개 세션 블라인드
Phase 5: 이중 평가 & 검증
LLM 평가 3개 모델 N=280
교사 평가 수학 2명 N=100
상호 검증 r=0.771 패턴 일치

[그림 1-2] 연구 진행 절차

마. 예비 조사 결과가 시사하는 두 가지 설계 방향

예비 조사 결과는 MAICE 시스템 설계에 두 가지 중요한 방향을 제시하였다

1) 시스템 설계 방향: 질문 명료화 프로세스의 필요성

질문 품질과 답변 품질 간 강한 상관관계(r=0.691)는 질문 개선이 학습 효과 향상의 핵심임을 보여준다. 특히 학습 맥락 부재(72.3%)와 학습 확장성 결여(48.9%)가 가장 심각한 문제로 확인되어, MAICE는 이 두 영역을 집중적으로 개선하는 질문 명료화 프로세스를 핵심 기능으로 설계하였다.

가) 평가 도구 개선 방향: 객관적 측정 도구의 필요성

예비 조사 루브릭의 낮은 교사 간 일치도(r=0.28, ICC=0.29)는 5점 척도 종합 판단 방식의 주관성 문제를 드러냈다. 이는 본 실험에서 더 객관적이고 신뢰성 높은 평가 도구가 필요함을 시사하였다.

이러한 필요성에 따라 본 연구는 QAC 체크리스트(Question-Answer-Context Checklist)를 개발하였다(6.4절 상세). QAC는 예비 조사 루브릭의 6개 영역에 대화 맥락 영역(C)을 추가하고, 각 항목을 4개의 명확한 체크리스트 요소로 분해하여 이진 판단(0/1)으로 평가하는 개선된 도구이다(8개 항목, 32개 체크리스트, 40점 만점).

본 실험에서는 평가의 일관성과 대규모 평가 가능성을 위해 3개 독립 AI 모델(GPT-5-mini, Claude-4.5-Haiku, Gemini-2.5-Flash)을 평가자로 활용하였으며, 이들 AI 평가자 간 신뢰도는 ICC(2,1)=0.595, Cronbach's α=0.840으로 확인되었다. 추가로 연구 객관성 확보를 위해 외부 현직 수학교사 2명(평가자 96, 97)이 동일한 100개 세션을 독립적으로 평가하여(총 200개 평가 레코드) AI 평가의 타당성을 검증하였으며, 교사 간 신뢰도는 Pearson 상관계수 r=0.644로 나타났다(7장 참조).

2. 연구 목적

예비 조사에서 확인된 핵심 발견은 질문 품질과 답변 품질 간 강한 상관관계(r=0.691)로, 이는 질문 개선이 학습 효과 향상의 핵심 메커니즘임을 의미한다. 따라서 본 연구는 고등학교 2학년 수학적 귀납법 단원을 중심으로 질문 명료화와 지식 유형별 맞춤 답변을 통합한 멀티 에이전트 AI 학습 시스템 MAICE(Mathematical AI Chatbot for Education)를 설계·개발한다.

MAICE는 다음 3가지 핵심 기능을 통합한다

- ○ 질문 명료화(QuestionImprover: Dewey 5단계 기반 사고 구조화 지원)
- ○ 지식 유형별 맞춤 답변 (QuestionClassifier + AnswerGenerator: Bloom K1-K4 분류 및 차별화된 답변 전략)
- ○ 대화 맥락 관리(LearningObserver: 학습 흐름 유지 및 누적 지원).

MAICE의 교육적 효과를 검증하기 위해, 본 연구는 무작위 배정 A/B 테스트를 설계하였다. 학생들을 두 그룹으로 나누어 한 그룹은 MAICE의 통합 학습 지원 기능을 제공받는 방식(본 연구에서 "Agent 모드"로 명명)을 사용하고, 다른 그룹은 일반적인 LLM처럼 즉시 답변만 제공하는 방식(본 연구에서 "Freepass 모드"로 명명)을 사용하도록 하여, MAICE의 명료화 중심 멀티 에이전트 설계가 학습 지원에 미치는 효과를 비교 분석한다.

가. 학술적 연구 목적

1) 질문 명료화 기반 멀티 에이전트 시스템 개발

예비 조사에서 확인된 질문의 질적 문제(학습맥락 부재 72.3%, 질문구조 불명확 45.8%, 수학적전문성 결여 45.5%)를 개선하는 질문 명료화 프로세스를 핵심으로 하는 멀티 에이전트 시스템 개발

2) 질문 품질 개선 효과 검증

MAICE의 질문 명료화 프로세스가 학생의 질문 품질(학습맥락, 질문구조, 수학적전문성)을 일반적인 즉시 답변 방식 대비 실질적으로 개선함을 입증

3) 학습 경험 향상 증명

질문 명료화 프로세스를 통한 구조화된 학습 경험이 즉시 답변 방식보다 학습 만족도와 메타인지를 향상시킴을 입증

4) 수학적 귀납법 단원 적용 검증

MAICE 에이전트 시스템이 수학적 귀납법 단원(논리 구조 이해, 단계별 증명 필요)에서 효과적으로 작동하는지 검증

나. 교육적 실천 목적

본 연구는 학술적 검증과 함께, 교육 현장에서 활용 가능한 실용적 기여를 다음과 같이 목표한다

1) 메타인지 발달 지원 검증

베타테스트(n=11)에서 학생들은 명료화 과정을 통해 "모호한 질문을 구체화하는 방법"(4.0/5점), "무엇을 모르는지 스스로 규정"(4.1/5점) 능력이 향상됨을 보고하였다. 본 실험에서는 더 큰 표본(N=280)을 통해 메타인지 발달 효과를 정량적으로 검증한다.

2) 교사 맞춤형 프롬프팅 연구 환경

MAICE는 5개 에이전트의 프롬프트를 모두 공개하고 수정 가능하도록 설계하였다. 교사는 자신의 교육 목표(예: 수학적 귀납법의 특정 단계 강조)에 맞춰 명료화 질문 템플릿, K1-K4 분류 기준, 답변 전략을 직접 수정하고 그 효과를 측정할 수 있다.

3) 확장 가능한 연구 플랫폼 제공

본 연구의 QAC 체크리스트, LLM-교사 이중 평가 모델, A/B 테스트 설계는 다른 교과·단원에 적용 가능하다. 시스템 코드와 평가 도구를 공개하여 후속 연구자와 교사가 자신의 맥락에서 재현·확장할 수 있도록 한다.

3. 연구 문제

예비 조사를 통해 확인된 일반적인 즉시 답변 제공 방식의 핵심 문제는 학생들의 질문 품질 부족(학습맥락 72.3% 최저점)이며, 질문 품질과 답변 품질 간 강한 상관관계(r=0.691)는 질문 개선이 학습 효과 향상의 핵심 메커니즘임을 시사한다.

앞서 1.2절에서 설명한 바와 같이, 본 연구는 MAICE 시스템의 두 가지 작동 방식—질문 명료화를 제공하는 Agent 모드와 일반적인 LLM처럼 즉시 답변만 제공하는 Freepass 모드—를 무작위 배정 A/B 테스트로 비교한다. MAICE 시스템의 효과성을 검증하기 위한 연구 문제를 다음과 같이 설정한다:

가. RQ1. 질문 품질 개선 검증 (핵심 메커니즘):

MAICE의 질문 명료화 프로세스(Agent 모드)가 학생의 질문 품질을 실질적으로 개선하는가? 특히 예비 조사에서 가장 심각한 문제로 확인된 학습 맥락 적용, 질문 구조화, 수학적 전문성 영역에서 일반적인 즉시 답변 방식(Freepass 모드) 대비 유의미한 개선이 나타나는가?

나. RQ2. 학습 경험 향상 검증 (교육적 가치):

질문 명료화 프로세스를 통한 구조화된 학습 경험(Agent 모드)이 즉시 답변 방식(Freepass 모드)보다 학생의 학습 만족도와 메타인지를 향상시키는가?

4. 용어의 정의

가. 예비 조사(Pilot Study):

본 연구의 MAICE 시스템 설계에 앞서, 2024년 5월 중순 실제 고등학교 수학 수업 환경에서 프리패스 방식 LLM의 교육적 문제점을 실증적으로 파악하기 위해 실시한 질문-답변 수집 및 교사 루브릭 평가 연구 (385건 질문, 1,012건 교사 평가)

나. MAICE (Mathematical AI Chatbot for Education):

예비 조사에서 확인된 일반적인 LLM 사용 방식의 한계를 극복하기 위해 개발된 질문 명료화 기반 AI agent 시스템. 본 연구에서는 MAICE를 두 가지 모드로 운영하여 비교 실험을 수행한다.

다. Agent 모드:

MAICE 시스템의 질문 명료화 프로세스를 제공하는 실험군 조건. QuestionImprover 에이전트가 학생의 초기 질문을 받아 명료화 질문을 통해 문제를 구체화한 후, 명료화된 질문에 대해 맞춤형 답변을 제공한다.

라. Freepass 모드:

MAICE 시스템의 대조군 조건으로, 일반적인 LLM(ChatGPT, Claude 등)처럼 질문 명료화 과정 없이 학생의 질문에 즉시 답변만 제공하는 방식. "Freepass"는 "명료화 과정을 거치지 않고(free) 바로 통과(pass)하여 답변"이라는 의미로 본 연구에서 명명하였다.

마. 질문 명료화:

모호하거나 불완전한 질문을 구체적이고 명확한 질문으로 개선하는 과정으로, 학습자가 스스로 질문을 구조화하고 사고 과정을 명료화하도록 지원하는 교육적 개입.

바. AI 에이전트(AI Agent):

특정 목표를 달성하기 위해 환경을 인식하고, 자율적으로 의사결정을 내려 행동하는 소프트웨어 시스템. 본 연구에서는 LLM을 기반으로 특정 교육적 역할(질문 분류, 명료화 지원, 답변 생성 등)을 수행하도록 설계된 독립적 구성 요소를 의미함.

사. 멀티 에이전트 시스템:

질문 분류, 명료화 지원, 답변 생성, 학습 관찰 등 각기 다른 역할을 수행하는 여러 개의 독립적인 AI agent가 협업하여 학습 과정을 지원하는 시스템

아. 학습 효과:

수학적 이해도, 문제해결 능력, 질문 품질 개선, 메타인지 향상 등을 포함하는 학습 성과와 학습 경험의 종합적 측정

II. 이론적 배경

1장의 예비 조사에서 프리패스 방식 LLM의 교육적 문제점이 실증적으로 확인되었다. 학생 질문의 72.3%가 학습맥락 최저점을 받았고, AI 답변의 48.9%가 학습확장성 최저점을 받았으며, 질문 품질과 답변 품질 간 강한 상관관계(r=0.691)가 발견되었다. 이는 질문 개선이 학습 효과 향상의 핵심 메커니즘임을 시사한다. 본 장에서는 이러한 문제를 해결하기 위해 개발한 질문 명료화 기반 AI agent 시스템 MAICE의 이론적 기반을 체계적으로 제시한다.

MAICE(Mathematical AI Chatbot for Education)는 다음 5개의 독립적 에이전트로 구성된 멀티 에이전트 시스템이다

[표 2-1] MAICE 멀티 에이전트 시스템 구성

약자	에이전트 명칭	핵심 역할	이론적 기반
QC	Question Classifier	질문 품질 진단 및 K1-K4 분류	Bloom 분류학 (2.1절)
QI	Question Improver	명료화 질문으로 문제 구체화	Dewey 반성적 사고 (2.2절)
AG	Answer Generator	K1-K4 유형별 맞춤 답변 생성	Bloom + AI 피드백 (2.4절)
LO	Learning Observer	대화 요약 및 컨텍스트 관리	세션 연속성 유지 (3.3.4절)
FT	Free Talker	대조군: 명료화 없는 즉시 답변	(비교 기준선)

1. 이론적 기반의 구조

MAICE 시스템의 핵심 설계 철학은 Dewey의 반성적 사고 5단계 중 2단계(문제 정의)를 명료화 프로세스로 구현하는 것이다. 다음 다이어그램은 프리패스 방식의 문제점과 MAICE의 해결 방안을 비교한다

1장의 예비 조사에서 발견된 질문 품질과 답변 품질의 강한 상관관계(r=0.691)는 Dewey의 2단계(문제 정의)가 학습 효과의 핵심 메커니즘임을 실증적으로 뒷받침한다.

[그림 2-1] 프리패스 방식과 MAICE 명료화 방식의 비교

[표 2-2] 이론적 기반과 MAICE 구현의 연결

이론 영역	이론 (2장)	핵심 개념	1장 문제점 해결	MAICE 구현
교육학적 토대	2.1 Bloom 지식 분류	K1→K4 4단계 분류	질문구조 불명확 45.8%	QC질문 분류 → AGK1-K4 맞춤 답변
	2.2 Dewey 반성적 사고	2단계 문제 정의	학습맥락 부재 72.3%	QI 명료화 프로세스
	2.3 질문 생성 이론	구조·완결·의도	수학적전문성 45.5%	QC품질 진단 → QI개선 유도
기술적 구현	2.4 AI 피드백 시스템	즉시성·맞춤성	일방향 답변	실시간 대화형 피드백
	2.5 멀티에이전트	역할 분담·협업	단일 AI의 한계	QC/QI/AG/LO/FT 5개 협업
	2.6 수학적 귀납법	단원 특성 반영	일반적 설명	단원 맥락 적응형 명료화
평가 방법론	2.7 루브릭 개발	질문-답변 평가	평가 기준 부재	QAC 체크리스트

표 2-1과 그림 2-1에서 보듯이, 각 이론은 1장에서 발견된 문제점을 해결하기 위한 이론적 해결책을 제시한다. 이어지는 절에서는 각 이론의 핵심 개념과 본 연구에의 적용을 상세히 검토한다.

2. 블룸의 지식 분류: 4가지 지식 차원

블룸의 개정 분류(Anderson & Krathwohl, 2001)에서 지식 차원(Knowledge Dimension)은 다음 네 범주로 구성된다. 본 연구에서는 질문 명료화의 목표를 학습자의 지식 차원에 정렬하여, 질문 품질과 학습 효과를 동시에 개선한다.

[표 2-3] Bloom의 지식 차원 분류 (K1~K4)

본 연구 표기	지식 차원	원문 표기	정의	수학적 귀납법 맥락의 예시 질문
K1	사실적 지식	Factual Knowledge	용어, 기본 사실, 기호·규칙의 기억	"귀납법의 기본 단계와 귀납 단계의 정의는 무엇인가요?"
K2	개념적 지식	Conceptual Knowledge	관계, 분류, 원리·법칙의 이해	"귀납 가정이 증명 논리에서 맡는 역할을 설명해 주세요."
K3	절차적 지식	Procedural Knowledge	방법·알고리즘·기법의 적용 절차	"부등식 증명에서 k→k+1 전개는 어떤 순서로 진행하나요?"
K4	메타인지적 지식	Metacognitive Knowledge	자기 인지·전략 선택·오류 진단	"내 전개에서 누락한 가정은 무엇이며, 어떤 전략을 선택해야 하나요?"

본 연구에서는 시스템 구현과 서술의 편의를 위해 4가지 지식 차원을 다음과 같이 약어로 표기한다: K1(사실적 지식), K2(개념적 지식), K3(절차적 지식), K4(메타인지적 지식). 이는 Anderson & Krathwohl (2001) 원문의 표기법은 아니며, 본 연구가 에이전트 명명 및 분류 체계를 위해 도입한 것이다. 이후 본문에서는 주로 K1-K4 표기를 사용한다.

이러한 지식 차원 분류는 MAICE의 명료화 프로세스와 다음과 같이 정렬된다. 사실적·개념적 지식(K1-K2)의 경우 용어 혼동이나 개념적 오해를 명료화 질문으로 선제 교정하며, 절차적 지식(K3)은 절차적 막힘을 단계화된 프롬프트로 유도한다. 메타인지적 지식(K4)은 학습자가 스스로의 오류와 전략을 메타인지적으로 진단하도록 반문을 통해 유도한다.

본 연구의 QC(Question Classifier)는 질문 분류 단계에서 학습자의 발화를 이 4가지 지식 차원으로 분류하고, AG(Answer Generator)가 각 차원에 정렬된 맞춤 답변을 제공한다. 이를 통해 질문의 질적 향상과 문제 해결 과정을 지원한다.

3. 듀이의 반성적 사고 이론: 명료화 프로세스의 철학적 기반

가. 반성적 사고의 정의와 5단계

듀이(1910)는 반성적 사고(reflective thinking)를 "어떤 믿음이나 지식의 형태에 대한 능동적이고 지속적이며 신중한 고려로서, 그것을 뒷받침하는 근거와 그것이 이끄는 결론들을 검토하는 것"으로 정의하였다.[3] 원문에서 Dewey는 "Active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it, and the further conclusions to which it tends" 라고 표현하였다(p.6). 이는 단순한 정보 습득이나 수동적 수용을 넘어, 학습자가 능동적으로 믿음의 근거를 탐구하고 그 함의를 추론하는 과정을 강조한다.

Dewey(1910, p.13)는 "사고의 기원은 어떤 당혹감, 혼란, 또는 의심(the origin of thinking is some perplexity, confusion, or doubt)"이라고 강조하며, 반성적 사고가 단순한 제안의 수용이 아닌 추가 증거를 탐색하는 능동적 과정임을 명확히 하였다.

듀이가 제시한 반성적 사고의 5단계는 다음과 같다(Dewey, 1910, p.72):

[표 2-4] Dewey의 반성적 사고 5단계와 수학적 귀납법 학습 예시

단계	Dewey 원문 표현	본 연구의 한국어 명칭	수학적 귀납법 학습 예시
1단계	a felt difficulty	문제 상황 인식	"귀납법 문제를 풀려는데 어디서부터 시작해야 할지 모르겠다"
2단계	its location and definition	문제의 위치 파악 및 정의	"귀납 단계에서 n=k+1을 증명할 때 귀납 가정을 어떻게 사용하는지 모르겠다"
3단계	suggestion of possible solution	가능한 해결책의 제안	"귀납 가정 P(k)를 식에 대입해보면 될까?"
4단계	development by reasoning of the bearings of the suggestion	제안된 해결책의 추론적 전개	P(k)를 실제로 대입하여 P(k+1) 유도 시도
5단계	further observation and experiment leading to its acceptance or rejection	관찰과 실험을 통한 수용 또는 거부	"귀납 가정을 대입하고 정리하니 P(k+1)이 증명되었다"

위 표는 Dewey (1910)의 원문 표현을 정확히 제시하며, 한국어 명칭은 본 연구에서 교육적 맥락에 맞게 번역한 것이다.

나. 수학 학습에서 반성적 사고의 중요성

수학 학습에서 반성적 사고는 특히 중요하다. Schoenfeld(1985)는 수학적 문제해결 과정이 단순한 알고리즘 적용이 아니라, 문제 상황 분석, 전략 선택, 실행, 검증의 메타인지적 과정임을 강조하였다.[4] 학생들이 수학적 문제에 직면했을 때, 자신의 기존 지식과 새로운 문제 상황 사이의 간극을 인식하고, 이를 해결하기 위해 체계적으로 탐구하는 과정이 바로 반성적 사고이다.

수학적 귀납법 학습에서는 다음과 같은 반성적 사고가 특히 중요하다:

- 문제 인식: "n=k+1 단계가 n=k 단계와 어떻게 다른가?"
- 문제 정의: "귀납 가정을 '언제', '어디에' 사용하는가?"
- 가설 형성: "식을 전개하면 귀납 가정과 연결될까?"
- 검증: 실제로 전개하여 논리적 연결 확인
- 성찰: "이 방법을 다른 귀납법 문제에도 적용할 수 있을까?“

다. 기존 LLM의 한계: 2단계(문제 정의) 생략

1장의 예비 조사에서 확인된 프리패스 방식의 근본적 문제는 Dewey의 2단계(문제 정의)를 완전히 생략한다는 것이다(그림 2-1 상단 참조).

프리패스 방식에서 학생이 "귀납법 어려워요"라고 질문하면(1단계: 문제 상황 인식), AI는 2단계(문제 정의) 없이 즉시 3-5단계로 진행하여 "귀납법은 다음과 같이 증명합니다... 1. 기본 단계: n=1일 때... 2. 귀납 단계: n=k일 때 가정하면..."과 같은 일반적 설명을 제공한다. 그 결과 학생은 자신이 정확히 무엇을 모르는지 인식하지 못한 채 AI의 설명을 수동적으로 받아들이게 된다.

이는 1장에서 관찰된 학습맥락 부재 72.3%의 근본 원인이다. 학생이 자신의 어려움을 명확히 정의하지 못한 상태에서 AI가 일방적으로 답변하므로, 학습자 수준 파악 실패(27.6% 최저점)와 학습확장성 결여(48.9% 최저점)로 이어진다.

라. MAICE의 해결 방안: 2단계 명료화 프로세스

MAICE는 Dewey의 2단계(문제 정의)를 명료화 프로세스로 구현하여, 학생이 스스로 질문을 구조화하도록 돕는다(그림 2-1 하단 참조).

MAICE 방식에서는 학생이 "귀납법 어려워요"라고 질문하면(1단계: 문제 상황 인식), QI 에이전트가 2단계(문제 정의)를 위해 명료화 질문을 제시한다: "귀납법 중에서 어떤 부분이 가장 어렵거나 궁금하신가요? • 기본 단계 증명? • 귀납 가정 이해? • 귀납 단계 전개?" 학생이 "귀납 단계에서 식을 전개하는 과정이요"라고 응답하면 2단계가 완성되고, AG 에이전트는 학생이 정의한 문제에 맞춘 맞춤형 답변을 생성한다: "n=k+1 대입 후 식 전개 과정을 단계별로 설명해드릴게요..."

그 결과 학생이 자신의 어려움을 명확히 인식하고 표현하게 되며, 이는 메타인지 능력 향상과 맞춤형 답변 수신으로 학습 효과 증대로 이어진다.

마. 본 연구에의 적용

듀이의 반성적 사고 이론은 MAICE의 명료화 프로세스 설계에 적용된다. Dewey 5단계는 MAICE의 5개 에이전트 구조와 직접 매핑되며(상세 매핑은 2.5절 라 참조), 특히 2단계(문제 정의)가 QI 에이전트의 명료화 프로세스로 구현된다(구현 상세는 3.3.2절 참조).

2단계(문제 정의)의 구현인 명료화 프로세스는 Dewey의 5단계를 실제 대화형 질문으로 변환한다:

○ Dewey 5단계 기반 명료화 전략:

- 1단계 (문제 인식): "어떤 부분이 가장 어렵거나 궁금하셨나요? "
- 2단계 (문제 정의): "이해한 부분과 아직 헷갈리는 부분을 나누어볼까요?"
- 3단계 (연결 탐색): "알고 있는 개념과 비교하면 어떤 점이 비슷하거나 다른가요?"
- 4단계 (사고 전개): "왜 이 부분이 궁금하신지 조금 더 설명해주실 수 있나요?"
- 5단계 (이해 검증): "어디서부터 막히셨는지 말씀해주실 수 있나요?"

○ 실제 구현 특징:

- K1-K4 질문 유형별로 맞춤형 명료화 질문 생성
- 매우 모호한 질문("수열 알려줘")에는 구체적인 선택지 제공
- 각 학생 응답마다 PASS/NEED_MORE 평가로 충분성 판단
- 최대 3회 시도 제한으로 과도한 명료화 방지

이는 단순히 정보를 전달하는 것이 아니라, 학생이 스스로 질문을 구조화하고 사고를 명료화하는 과정을 경험하도록 설계된 것이다. 이러한 명료화 프로세스는 Dewey의 반성적 사고 이론에 따르면 메타인지 발달과 학습 효과 향상을 가져올 것으로 기대된다. 본 연구에서는 8장에서 이를 실증적으로 검증한다.

4. 질문 생성 및 개선 이론: 질문 품질의 구조화

가. 질문 생성의 교육적 가치

질문 생성(question generation)은 학습자의 메타인지 및 비판적 사고를 유도하는 핵심적 활동으로, 학습자 주도적인 문제 해결과 이해 중심 수업을 가능하게 한다.

King(1994)은 초등학교 4-5학년 58명을 대상으로 한 실험 연구에서, 질문 생성 훈련을 받은 학생들이 대조군에 비해 이해도 검사(literal comprehension: $F(2,21)=3.56$, $p<.05$; inferential/integrative comprehension: $F(2,21)=4.17$, $p<.05$)와 7일 후 파지 검사(retention test: $F(2,21)=16.34$, $p<.001$)에서 유의미하게 높은 성취를 보였다. 특히 경험 기반 질문(experience-based questions)을 생성한 학생들이 수업 내 질문(lesson-based questions)만 한 학생들보다 장기 기억에서 더 우수한 성과를 나타냈다. 이는 학생 생성 질문이 단순 정보 습득을 넘어 복잡한 지식 구성(complex knowledge construction)을 유도하며, 이것이 학습 효과를 증진시킨다는 것을 실증적으로 보여준다.

특히 수학 학습에서 학생이 스스로 질문을 생성하는 것은 다음과 같은 교육적 효과를 갖는다:

- 능동적 학습: 교사의 일방적 설명을 수동적으로 받아들이는 것이 아니라, 학습자가 능동적으로 지식을 구성
- 메타인지 발달: 자신이 무엇을 아는지, 무엇을 모르는지 스스로 인식하는 능력 향상
- 개념 이해 심화: 질문을 구성하는 과정에서 개념 간 관계를 탐색하고 논리적 연결을 시도
- 학습 동기 증진: 자신의 궁금증에서 시작한 질문은 학습 몰입도를 높임

나. 효과적인 질문의 특징

King(1994)은 학생 생성 질문이 깊이 있는 이해와 장기 기억에 긍정적 영향을 미친다는 것을 실증하였으며, Graesser & Person(1994)은 튜터링 과정에서의 질문 유형을 체계적으로 분류하였다.[7] 본 연구는 이러한 선행 연구들과 예비 조사에서 관찰된 질문 품질 문제를 종합하여, 효과적인 학습 질문의 3가지 핵심 특징을 다음과 같이 도출하였다:

다음 제시하는 효과적인 질문의 3가지 특징은 King(1994)과 Graesser & Person(1994)의 이론적 틀을 바탕으로, 본 연구의 예비 조사에서 발견된 질문 품질 문제(1장 1.1.2절)를 해결하기 위해 구체화한 것이다.

1) 구조화 (Structuring)

질문이 명확한 구조를 가지고 있어야 한다:

[표 2-5] 효과적인 질문의 구조 요소

질문 의도	기대 답변 유형	예시
정의 확인	간결한 정의	"귀납 가정이 정확히 뭐예요?"
개념 이해	관계 설명	"귀납 가정은 왜 필요한가요?"
절차 학습	단계별 안내	"귀납 가정을 어떻게 사용하나요?"
오류 진단	메타인지적 피드백	"제 풀이에서 뭐가 틀렸나요?"

○ 좋은 예시: "귀납 단계에서 n=k+1을 증명할 때, 귀납 가정 P(k)를 어떻게 사용하나요?"

- 대상: 귀납 가정 P(k)
- 범위: n=k+1 증명 단계
- 초점: 사용 방법

○ 나쁜 예시: "귀납법 어려워요"

- 대상 불명확 (기본 단계? 귀납 단계? 개념?)
- 범위 불명확 (어떤 문제 유형?)
- 초점 부재 (무엇이 어려운지 불명확)

2) 완결성 (Completeness)

질문에 필요한 조건과 정보가 모두 제시되어야 한다:

[표 2-6] 효과적인 질문의 완결성 조건

조건 유형	정의	예시
학습 수준	현재 어디까지 배웠는가	"수열까지는 배웠어요"
선수 지식	어떤 배경 지식을 가지고 있는가	"등비수열은 알아요"
시도한 방법	무엇을 시도해봤는가	"n=k+1을 대입했는데..."
막힌 지점	어디서 막혔는가	"식 정리 과정에서"

○ 완결성 높은 질문 예시:
"고2 수학에서 수학적 귀납법을 배우고 있는데, 등비수열 합 공식을 귀납법으로 증명하는 문제에서 n=k+1 단계의 식을 전개했을 때 귀납 가정을 어떻게 대입해야 하는지 모르겠어요.“

이 질문은 학습 수준(고2), 학습 주제(귀납법), 구체적 문제(등비수열 합), 시도한 방법(n=k+1 대입), 막힌 지점(귀납 가정 대입)을 모두 제시하여 AI가 맞춤형 답변을 생성할 수 있게 한다.

3) 의도의 명시성 (Explicitness)

질문의 목적과 기대하는 답변 유형이 명확해야 한다:

[표 2-7] 효과적인 질문의 의도 명시성

질문 의도	기대 답변 유형	예시
정의 확인	간결한 정의	"귀납 가정이 정확히 뭐예요?"
개념 이해	관계 설명	"귀납 가정은 왜 필요한가요?"
절차 학습	단계별 안내	"귀납 가정을 어떻게 사용하나요?"
오류 진단	메타인지적 피드백	"제 풀이에서 뭐가 틀렸나요?"

◯ 의도가 명확한 질문:

- "귀납 단계의 정의를 간단히 설명해주세요" → 간결한 정의 기대
- "귀납 가정을 단계별로 어떻게 사용하는지 예시를 들어 설명해주세요" → 상세한 절차 기대

다. 기존 연구의 한계: 피드백 시스템 부재

대부분의 선행연구들은 질문 생성 기법을 가르치는 데 초점을 맞추었으며, 질문의 질적 평가와 개선을 위한 체계적인 피드백 시스템에 대한 연구는 제한적이다. 교사가 좋은 질문 만드는 방법을 교육하더라도, 학생이 실제 질문을 생성할 때 질문 품질을 평가하고 개선 피드백을 제공하는 시스템이 없다면 다음과 같은 문제가 발생한다:

1) 모호한 질문 → 모호한 답변: 질문 품질이 낮으면 AI 답변도 일반적이고 맥락 없는 설명에 그침

2) 나쁜 습관 고착: 피드백 없이 반복되면 모호한 질문 습관이 고착됨

3) 질문 개선 기회 상실: 질문을 구조화하고 명료화하는 메타인지 능력 발달 기회를 놓침

특히 AI 학습 환경에서는 즉각적인 질문 개선 피드백이 더욱 중요하다. 학생이 질문하는 순간이 바로 질문 품질을 개선할 수 있는 교육적 기회이기 때문이다.

라. 본 연구의 접근: 명료화 기반 질문 개선

MAICE는 질문 생성 이론을 실시간 피드백 시스템으로 구현하여, 학생이 질문하는 순간 질문 품질을 개선할 수 있도록 돕는다:

[그림 2-3] MAICE의 질문 개선 메커니즘: 실시간 명료화

이는 단순히 질문 방법을 가르치는 것이 아니라, 질문하는 순간 실시간으로 개선 과정을 경험하게 하여 메타인지 능력을 체득하도록 한다.

마. 1장 문제점과의 연결

예비 조사에서 발견된 질문 품질 문제는 질문 생성 이론의 3가지 특징 결여와 정확히 일치한다. 질문구조 불명확(45.8%)은 구조화 부족, 학습맥락 부재(72.3%)는 완결성 부족, 수학적전문성 결여(45.5%)는 의도 명시성 부족에서 기인한다(상세 매핑은 표 2-2 참조).

King(1994)의 이론에 따르면 좋은 질문이 좋은 학습을 만든다.

바. 본 연구에의 적용

○ 질문 생성 및 개선 이론은 MAICE에 다음과 같이 적용된다:

- 질문 품질 평가 기준 (3.4절): 구조화·완결성·의도를 측정하는 QAC 체크리스트 A영역 설계
- 명료화 전략 (3.3.2절): 3가지 특징을 유도하는 단계별 질문 설계
- 실시간 피드백 (3.3.1절): QC가 질문 품질을 진단하고 개선 방향 제시
- 효과성 검증 계획 (8장): 명료화 프로세스가 질문 품질 향상과 메타인지 발달로 이어지는지 검증 예정

5. AI 기반 피드백 시스템: 실시간 상호작용의 설계 원리

가. 피드백의 교육적 효과

피드백은 학습 과정에서 핵심적 역할을 하며, 특히 즉시성(immediacy)과 맞춤성(adaptiveness)이 학습 효과를 결정한다. Hattie와 Timperley(2007)의 메타분석 연구에 따르면, 효과적인 피드백은 학습자의 성취도에 평균 효과 크기 d=0.79의 큰 영향을 미친다.

[표 2-8] 효과적인 피드백의 핵심 요소

요소	정의	수학 학습 예시
즉시성	학습자 행동 직후 제공	질문 직후 명료화 질문 제시
구체성	일반적이 아닌 구체적 지적	"귀납 가정 사용 부분에서..."
건설성	오류 지적 + 개선 방향	"n=k+1 대신 귀납 가정을 먼저 확인해보세요"
맞춤성	학습자 수준에 적합	하위권에게는 더 세분화된 단계 제공

나. 질문 명료화를 통한 답변 범위 제한

MAICE의 핵심 전략은 질문을 구체화하여 답변 범위를 자동으로 제한하는 것이다:

◯ 기존 AI의 문제:

- 모호한 질문에 대해 모든 내용을 포괄적으로 설명
- 학생이 실제로 필요한 부분을 찾기 어려움
- 인지 과부하 발생 (정보량 과다)

◯ MAICE의 접근:

[표 2-9] MAICE의 질문 명료화를 통한 답변 범위 제한 메커니즘

단계	역할	활동	이론적 기반
1	QC	"어떤 부분이 어렵나요?"	Dewey 1단계: 문제 인식
2	학생	"귀납 단계요"	문제 구체화
3	QC	"귀납 단계 중에서?"	Dewey 2단계: 문제 정의
4	학생	"식 전개 과정"	최종 명료화
5	AG3	식 전개만 설명	Bloom K3: 절차적 지식

◯ 핵심 원리:

- 질문이 구체화되면 → 답변 범위가 자동으로 제한됨 → 인지 부담 최소화
- 학생은 "귀납법 전체"가 아니라 "식 전개 과정"만 학습하게 되어, 필요한 정보에 집중할 수 있다.

다. 실시간 상호작용의 기술적 구현

MAICE는 실시간 피드백의 즉시성을 보장하기 위해 다음 기술을 활용한다:

1) 스트리밍 응답: AI 답변을 타이핑하듯 실시간으로 전송 (Server-Sent Events, 4장 4.2.1절)

2) 비동기 처리: FastAPI의 async/await로 여러 학생의 질문을 동시에 처리 (4장 4.1.1절)

3) 멀티프로세스 아키텍처: 5개 에이전트가 독립 프로세스로 실행되어 병렬 처리 (4장 4.2.3절)

4) 응답 시간 최적화: 평균 2초대 첫 응답 (베타테스트, 3장 3.6.2절)

라. 본 연구에의 적용

AI 피드백 이론은 MAICE에 다음과 같이 적용된다:

○ 즉시성: 학생 질문 직후 QC가 품질 진단, QI가 명료화 질문 제시
○ 구체성: 질문의 어떤 부분(구조화/완결성/의도)이 부족한지 명시적 피드백
○ 건설성: 단순 지적이 아닌 명료화 질문으로 개선 방향 제시
○ 맞춤성: K1-K4 유형별 차별화된 답변 전략 (2.1절 참조)
○ 기술적 구현(스트리밍 응답, 비동기 처리)은 4장에서, 효과성 검증은 QAC 루브릭의 C2(학습 과정 지원) 항목으로 측정한다(7-8장)

6. 멀티 에이전트 시스템: 역할 분담과 협업

가. 멀티 에이전트 시스템의 정의

멀티 에이전트 시스템(Multi-Agent System, MAS)은 여러 자율적 agent가 협력하여 복잡한 문제를 해결하는 분산 시스템이다.[9] 교육 분야에서 멀티 에이전트 시스템은 학습자의 다양한 요구에 대응하는 개인화된 학습 환경을 구축하는 데 활용되고 있다.

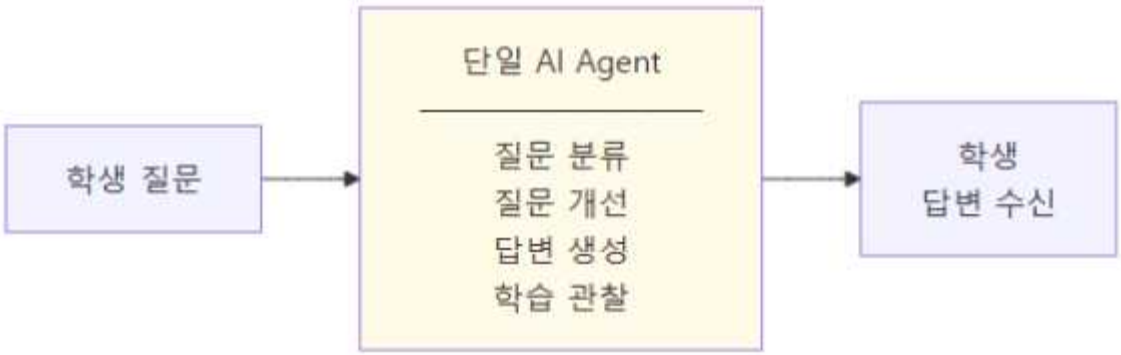
Wooldridge와 Jennings(1995)는 agent의 주요 특성으로 다음을 제시했다

[표 2-10] Wooldridge & Jennings의 Agent 특성과 교육적 응용

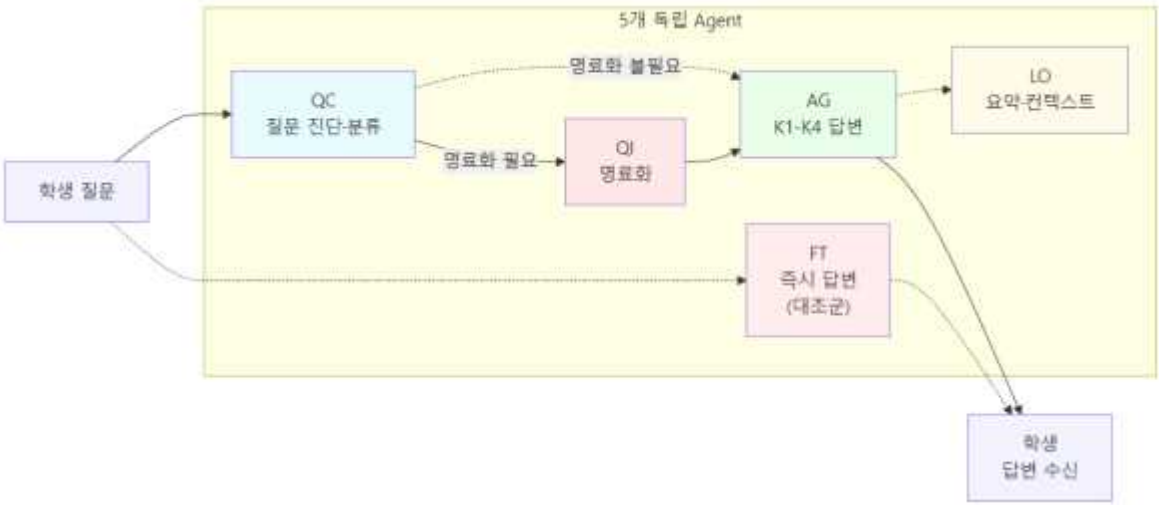
특성	정의	교육적 응용
자율성(Autonomy)	독립적으로 작동, 외부 개입 최소화	각 agent가 독자적 판단 및 실행
사회성(Social Ability)	다른 agent와 협력 및 정보 교환	agent 간 질문·답변 전달
반응성(Reactivity)	환경 변화를 감지하고 적절히 대응	학생 질문 패턴 변화 감지
능동성(Pro-activeness)	목표 지향적 행동, 선제적 조치	학습 어려움 조기 발견 및 개입

나. 단일 Agent vs 멀티 Agent의 차이

[그림 2-4] 단일 Agent 시스템



단일 Agent 시스템은 모든 역할을 하나의 Agent가 처리하므로 역할 간 충돌, 복잡도 증가, 전체 실패 위험, 유지보수 어려움 등의 문제가 발생한다.



[그림 2-5] 멀티 Agent 시스템의 역할 분담과 협업

멀티 Agent 시스템은 역할을 명확히 분담하여 각 agent를 독립적으로 최적화할 수 있고, 한 agent의 오류가 전체 시스템에 미치는 영향을 최소화하며, 새로운 agent 추가가 용이하다. MAICE는 Agent 모드(QC→QI→AG→LO)와 Freepass 모드(FT→LO)의 2가지 모드를 지원한다(상세 메커니즘은 3장 참조).

다. 교육용 멀티 에이전트 시스템의 설계 원칙

최근 Degen(2025)은 고등교육에서 Orchestrated Multi-Agent System (OMAS)의 개념을 제시하며, AI 소크라테스 튜터가 학생의 연구 질문 개발을 지원하는 방식을 연구하였다.[10] 이러한 조율된 멀티 에이전트 시스템은 각 agent가 독립적으로 작동하면서도 전체적으로 일관된 교육 목표를 달성하도록 설계된다.

교육 분야 멀티 에이전트 시스템은 다음 원칙을 따라야 한다:

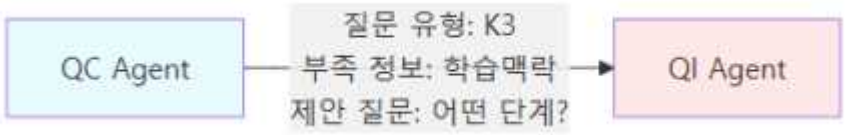
1) 명확한 역할 분담

◯ 각 agent의 책임 범위를 명확히 정의:

- 모호한 역할: "질문 처리 Agent" (너무 광범위)
- 명확한 역할: "질문 유형 분류 Agent" (구체적 책임)

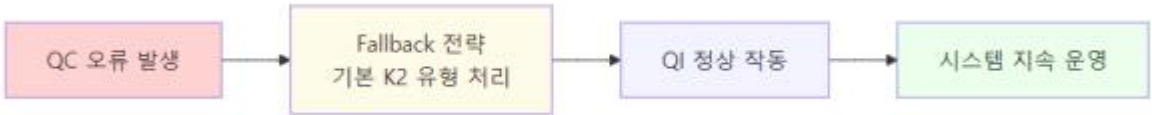
2) 효율적인 정보 교환

[그림 2-6] Agent 간 정보 교환 예시



3) 독립적 실패 처리

[그림 2-7] Agent 오류 시 Fallback 전략



라. 본 연구에의 적용

◯ 멀티 에이전트 이론은 MAICE에 다음과 같이 적용된다:

- 명확한 역할 분담: 5개 agent의 독립적 책임 정의 (3.3절)
- agent 간 협업 프로토콜: 정보 교환 구조 설계 (4.2절)
- 독립적 최적화: 각 agent의 프롬프트 독립 개선 가능 (4.2.3절)
- 시스템 안정성: 한 agent 오류 시 fallback 전략 (3.6.2절)
- 확장성: 새로운 agent 추가 용이한 아키텍처 (3.2절)

◯ 특히, MAICE는 교육적 프로세스를 agent 구조로 구현함으로써:

- Dewey 5단계 각각을 담당 agent가 지원 (QC → QI → AG → LO)
- 명료화(2단계)를 독립 agent(QI)로 분리하여 핵심 강조
- 대화 요약(LO)을 통해 장기 세션에서도 컨텍스트 관리 가능

이러한 멀티 에이전트 구조는 교육적 프로세스를 체계적으로 지원하며, 각 agent의 독립성과 협업을 통해 안정적인 시스템 운영이 가능할 것으로 기대된다. 본 연구에서는 8장에서 실제 학생 세션을 통해 시스템의 안정성과 교육적 효과를 검증한다.

7. 수학적 귀납법 단원 선정: 단원 맥락 반영 필요성

가. 연구 대상 단원으로 수학적 귀납법을 선정한 이유

본 연구는 고등학교 2학년 수학 I 과정의 수학적 귀납법 단원을 MAICE 시스템의 적용 대상으로 선정하였다. 이는 다음의 교육적 특성 때문이다:

1) 명료화가 특히 필요한 학습 내용

수학적 귀납법은 기본 단계(n=1)와 귀납 단계(n=k→k+1)라는 이중 구조를 가지며, 학생들은 자신의 어려움이 어느 단계에서 발생하는지 정확히 표현하기 어렵다. "귀납법 어려워요"와 같은 막연한 질문이 빈번하게 발생하며, 이는 명료화 프로세스의 효과를 검증하기에 적합하다.

2) Dewey 반성적 사고의 적용 가능성

수학적 귀납법의 핵심인 귀납 가정은 "증명하려는 것을 먼저 가정한다"는 역설적 특성을 가진다. 이는 학습자에게 당혹감과 혼란을 야기하며(Dewey의 1단계), 이를 정확히 정의하고(2단계) 해결하는 반성적 사고 과정이 필수적이다.

3) 유형별 전략 차이

등식 증명과 부등식 증명은 전개 전략이 크게 다르다(등식: 대입 후 정리, 부등식: 보조 부등식 필요). 학생이 어떤 유형의 문제에서 어려움을 겪는지에 따라 다른 설명이 필요하므로, 질문 명료화의 중요성이 더욱 부각된다.

나. LLM의 맥락 적응 능력 활용

MAICE는 수학적 귀납법에 특화된 별도의 코드를 작성하지 않는다. 대신 LLM의 맥락 이해 능력을 활용하여, 학생 질문에 포함된 "귀납법", "귀납 가정" 등의 단어로부터 단원 맥락을 자동으로 인식하고 적응한다:

- ○ QC 에이전트: 질문에서 "귀납 가정을 어떻게 써요?"를 보고 K3(절차적 지식)로 분류
- ○ QI 에이전트: "귀납법"이라는 맥락을 파악하여 "어느 단계가 어려운가요? 기본 단계? 귀납 단계?"와 같은 명료화 질문 동적 생성
- ○ AG 에이전트: K1-K4 분류에 따라 단원 맥락에 맞는 답변 생성

이러한 설계는 향후 다른 수학 단원(미적분, 삼각함수 등)으로 확장할 때도 시스템 구조 변경 없이 적용 가능함을 의미한다.

수학적 귀납법 단원의 구체적 특성과 MAICE의 실제 적용 사례는 5장에서 상세히 다룬다.

8. 평가 루브릭의 이론적 기반

본 연구는 MAICE 시스템의 효과성을 측정하기 위해 질문-답변 품질을 평가하는 QAC(Question-Answer-Context) 체크리스트를 개발하였다. 본 절에서는 루브릭의 개발 동기, 이론적 기반, 그리고 체크리스트 구조를 제시한다.

가. 루브릭 개발의 필요성과 이론적 기반

1) 개발 동기: 예비 조사에서 발견된 문제

본 연구는 MAICE 시스템 개발에 앞서 실제 중학생 385명이 AI와 나눈 수학 학습 대화를 분석하는 예비 조사를 수행하였다. 이 과정에서 학생 질문과 AI 답변에서 다음과 같은 체계적 문제가 발견되었다:

○ 질문 영역의 문제:

- 학습 맥락 부재: "수학적 귀납법 어려워요"와 같이 현재 학습 수준, 어려움의 구체적 지점, 학습 목표를 제시하지 않음
- 질문 구조 불명확: 한 질문에 여러 의도가 섞이거나, 필요한 조건이 누락됨
- 수학적 표현 부정확: 수학 용어의 오용 또는 개념 혼동

○ 답변 영역의 문제:

- 학습 확장성 결여: 단편적 답변 제공 후 심화 방향 제시 없음
- 학습자 맞춤도 부족: 학생의 수준이나 선수학습 상태를 고려하지 않은 일률적 설명
- 설명 체계성 부족: 논리적 흐름 없이 단편적 정보 나열

이러한 문제는 단순한 정답 제시를 넘어 교육적으로 효과적인 대화를 구성하는 요소

가 무엇인지 명확한 기준이 필요함을 보여준다. 기존 AI 평가 연구들은 주로 답변의 정확성만을 평가하였으나, 본 연구는 질문 품질도 학습 효과의 핵심 요소로 보고 이를 독립적으로 평가하는 체계를 개발하였다.

2) 이론적 기반: 교육학 이론의 통합

루브릭의 6개 평가 영역은 전통적 교육학 이론에 기반하여 설계되었다:

(1) A영역: 질문 평가 (15점)

◯ A1. 수학적 전문성 (5점)

- 이론적 기반: 내용지식(content knowledge)의 중요성
- 측정 대상: 수학 개념의 정확성, 용어 사용의 적절성, 교과과정 내 위계 파악
- 교육적 의의: 학습자가 문제의 본질을 정확히 파악하고 있는지 평가

◯ A2. 질문 구조화 (5점)

- 이론적 기반: King(1994) 질문 생성 이론, Graesser & Person(1994) 튜터링 질문 분류
- 측정 대상: 질문의 단일성, 조건 완결성, 문장 논리성, 의도 명확성
- 교육적 의의: 명확한 질문은 명확한 답변을 유도하며, 이는 효과적 학습 대화의 출발점

◯ A3. 학습 맥락 적용 (5점)

- 이론적 기반: 상황학습이론(situated learning), 근접발달영역(ZPD) 개념
- 측정 대상: 현재 학습 단계, 선수학습 내용, 구체적 어려움, 학습 목표
- 교육적 의의: 맥락 정보는 AI가 학생에게 적합한 수준의 설명을 제공하는 데 필수적

(2) B영역: 답변 평가 (15점)

○ B1. 학습자 적합성 (5점)

- 이론적 기반: 비계설정(scaffolding), 차별화 교수(differentiated instruction)
- 측정 대상: 수준별 접근, 선수지식 연계, 난이도 조절, 개인화 피드백
- 교육적 의의: 학생의 현재 수준에 맞는 설명이 학습 효과를 극대화

○ B2. 설명의 체계성 (5점)

- 이론적 기반: 인지부하이론(cognitive load theory), 멀티미디어 학습 원리
- 측정 대상: 개념 위계화, 단계별 논리, 핵심 강조, 예시 적절성
- 교육적 의의: 체계적 설명은 인지 부하를 줄이고 이해를 촉진

○ B3. 학습 내용 확장성 (5점)

- 이론적 기반: Bloom(1956, as cited in Anderson & Krathwohl, 2001) 교육목표분류학, Anderson & Krathwohl(2001) K1-K4 지식 차원
- 측정 대상: 심화 방향 제시, 응용 문제 연계, 오개념 교정, 자기주도 학습 유도
- 교육적 의의: 단편적 답변을 넘어 지속적 학습으로 연결

(3) C영역: 맥락 평가 (10점)

○ C1. 대화 일관성 및 연속성 (5점)

- 이론적 기반: 대화 일관성 이론, 공통기반이론(common ground theory)
- 측정 대상: 학습 목표 중심성, 대화 이력 참조, 주제 연속성, 턴 간 유기적 연결
- 교육적 의의: 일관된 대화 흐름은 학습 몰입과 이해 누적을 지원

○ C2. 학습 과정 지원성 (5점)

- 이론적 기반: Dewey(1910) 반성적 사고, 메타인지 이론
- 측정 대상: 사고 과정 유도, 이해도 확인, 메타인지 촉진, 깊이 있는 사고 유도
- 교육적 의의: 학습자가 자신의 사고 과정을 인식하고 조절하도록 지원

(4) 루브릭의 특징

본 루브릭은 기존 AI 평가 도구와 다음 점에서 차별화된다:

○ 질문 품질 독립 평가: 기존 연구는 AI 답변만 평가했으나, 본 연구는 질문 품질이 학습 효과의 선행 조건임을 인식

○ 이론-실증 통합: 교육학 이론을 실제 학생 데이터(385건)에서 발견된 문제와 연결

○ 체크리스트 방식: 각 영역을 4개의 구체적 요소로 세분화하여 평가 일관성 확보 (상세 체크리스트는 2.8)

○ MAICE 설계 근거 제공: 루브릭에서 발견된 A3(맥락 부재)와 B3(확장성 부족) 문제가 MAICE의 명료화 프로세스 설계 동기가 됨

나. 루브릭 구조: 체크리스트 기반 평가

예비 조사에서 발견된 문제 패턴과 교육 이론을 결합하여 8개 평가 항목, 32개 체크리스트 요소로 구성된 QAC 체크리스트를 개발하였다:

1) A영역: 질문 평가 (15점)

[표 2-13] QAC 루브릭 A영역: 질문 평가 체크리스트

항목	체크리스트 요소	평가 기준 (질문 형태)
A1. 수학적 전문성(5점)	① 개념 정확성	수학 용어를 정확하게 사용했는가?
	② 교과과정 위계	학년 수준에 맞는 개념인가?
	③ 용어 적절성	전문 용어를 적절히 사용했는가?
	④ 문제 방향 구체성	해결하려는 문제가 구체적인가?
A2.질문 구조화(5점)	① 질문 단일성	한 번에 하나의 명확한 질문을 하는가?
	② 조건 완결성	필요한 조건/정보를 모두 제시했는가?
	③ 문장 논리성	문장이 논리적으로 구성되었는가?
	④ 의도 명확성	무엇을 알고 싶은지 명확한가?
A3. 학습 맥락(5점)	① 학습 단계 설명	학년/단원/진도를 언급했는가?
	② 선수학습 언급	이전에 배운 내용을 언급했는가?
	③ 어려움 명시	어디서 막혔는지 구체적으로 말했는가?
	④ 학습 목표 제시	무엇을 배우고 싶은지 목표를 제시했는가?

2) B영역: 답변 평가 (15점)

[표 2-14] QAC 루브릭 B영역: 답변 평가 체크리스트

항목	체크리스트 요소	평가 기준 (질문 형태)
B1. 학습자 적합성(5점)	① 수준별 접근	학생 수준에 맞게 설명했는가?
	② 선수지식 연계	이미 배운 내용과 연결했는가?
	③ 난이도 조절	너무 어렵거나 쉽지 않은가?
	④ 개인화 피드백	학생 상황을 고려한 피드백인가?
B2. 설명의 체계성(5점)	① 개념 위계화	쉬운 것부터 어려운 것으로 단계적 설명?
	② 단계별 논리	각 단계가 논리적으로 연결되는가?
	③ 핵심 강조	중요한 내용을 명확히 강조했는가?
	④ 예시 적절성	이해를 돕는 적절한 예시 제공?
B3. 학습 확장성(5점)	① 심화 방향 제시	더 깊이 공부할 방향을 제시했는가?
	② 응용 문제 연계	관련된 응용 문제를 연결했는가?
	③ 오개념 교정	잘못된 이해를 바로잡았는가?
	④ 자기주도 유도	스스로 탐구하도록 유도했는가?

3) C영역: 맥락 평가 (10점)

[표 2-15] QAC 루브릭 C영역: 맥락 평가 체크리스트

항목	체크리스트 요소	평가 기준 (질문 형태)
C1. 대화 일관성(5점)	① 목표 중심성	학습 목표를 벗어나지 않고 진행?
	② 맥락 참조	전체 대화 이력을 기억하고 참조?
	③ 주제 연속성	주제가 자연스럽게 연결되는가?
	④ 턴 간 연결	각 발화가 직전 턴과 유기적 연결?
C2. 학습 과정 지원(5점)	① 사고 과정 유도	학생의 사고 과정을 유도하는가?
	② 이해도 확인	학생의 이해도를 확인하는가?
	③ 메타인지 촉진	학생이 학습 과정을 돌아보게 하는가?
	④ 깊이 있는 사고 유도	깊이 있는 사고를 유도하는가?

각 항목의 점수는 충족된 체크리스트 요소 개수로 자동 계산된다:

점수 = 충족 요소 개수 + 1

◯ 질문 영역 (A1+A2+A3): 최대 15점

◯ 답변 영역 (B1+B2+B3): 최대 15점

◯ 맥락 영역 (C1+C2): 최대 10점

◯ 전체 총점: 40점 만점

[표 2-16] 체크리스트 기반 평가 예시 (실제 데이터)

구분	낮은 품질 질문 예시	높은 품질 질문 예시
학생 질문	"수학적 귀납법이 뭐야"	"모든 자연수 에 대해, 임을 수학적 귀납법을 이용하여 증명하시오. 기저단계, 귀납단계를 사용하여 수식으로 간단히 풀어내시오."
A1. 수학적 전문성	4점 (3개 충족) ☑ 개념정확성, 용어적절성, 문제방향 ✕ 교과과정위계	5점 (4개 충족) ☑ 개념정확성, 교과과정위계, 용어적절성, 문제방향
A2. 질문 구조화	4점 (3개 충족) ☑ 질문단일성, 문장논리성, 의도명확성 ✕ 조건완결성	5점 (4개 충족) ☑ 질문단일성, 조건완결성, 문장논리성, 의도명확성
A3. 학습 맥락	1점 (0개 충족) ✕ 학습단계, 선수학습, 어려움, 목표 모두 부재	4점 (3개 충족) ☑ 학습단계, 선수학습, 어려움 ✕ 학습목표
질문 총점	9점/15점 (60%)	14점/15점 (93.3%)

체크리스트 방식은 이진 판단(충족/미충족)으로 평가자 간 일관성을 높이며, 어떤 요소가 부족한지 구체적으로 파악하여 개선 방향을 제시할 수 있다.

다. 본 연구에의 적용

본 루브릭은 기존의 검증된 측정 도구가 아닌, 본 연구가 개발하고 타당성을 탐색하는 평가 도구이다. 루브릭의 신뢰도 및 타당도 검증 과정, 교사 평가자 간 일치도, AI-교사 평가 일치도 등은 6장 6.4절에서 상세히 다룬다.

III. MAICE 교육 시스템 아키텍처

1. 설계 철학: "명료화 중심 학습"

가. 문제 인식: Freepass 방식의 근본적 한계

[표 3-1] Freepass 방식의 교육적 한계 (예비조사 결과)

한계 유형	발생 비율/사례	구체적 문제
질문 맥락 부재	72.3%	학습 수준·목적 정보 없음
맥락 오해	실증 사례	"지수의 확장"→비즈니스로 오해
수준 불일치	실증 사례	고1에게 대학 통계 개념 설명
인지 과부하	실증 사례	요청 없이 모든 개념 나열

1장에서 확인한 바와 같이, 일반적인 LLM의 Freepass 방식은 위와 같은 교육적 한계를 보였다.

이러한 문제들은 단순히 AI 성능의 문제가 아니라, 질문의 질이 낮을 때 아무리 좋은 AI도 적절한 답변을 생성할 수 없다는 구조적 한계에서 비롯된다.

나. 해결 아이디어: 교육 이론 기반 에이전트 시스템

본 연구는 2장에서 검토한 교육 이론을 실제 AI 시스템으로 구현하는 것을 목표로 한다:

◯ Bloom의 K1-K4 분류 → 질문 유형 자동 분류 및 맞춤형 답변 생성

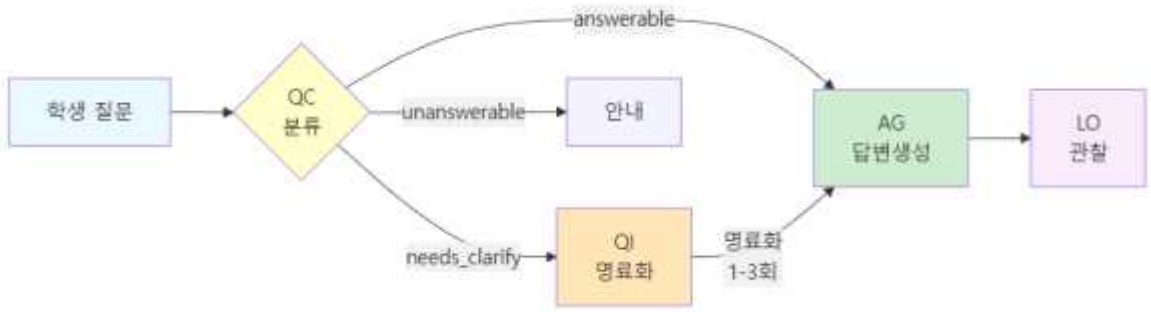
- K1 (사실): 간결한 정의 중심 답변
- K2 (개념): 관계 설명 중심 답변
- K3 (절차): 단계별 안내 중심 답변
- K4 (메타인지): 메타인지 유도 중심 답변

○ Dewey의 반성적 사고 5단계 → 명료화 프로세스 설계

- 1단계 (문제 인식): "무엇이 불확실한가요?"
- 2단계 (문제 정의): "정확히 무엇을 알고 싶은가요?"
- 3단계 (가설 설정): "어떤 방법을 시도해봤나요?"
- 4단계 (가설 검증): "논리적 연결을 어떻게 보나요?"
- 5단계 (결론 도출): "최종적으로 무엇을 얻고 싶나요?“

다. 핵심 아이디어: 질문 → 분류 → 명료화 → 답변 파이프라인

MAICE 시스템의 핵심은 학생의 질문을 즉시 답변하지 않고, 먼저 질문의 품질을 진단하고 필요시 명료화 과정을 거치는 것이다.



[그림 3-1] MAICE 질문 처리 파이프라인

이러한 파이프라인은 단순히 정보를 전달하는 것이 아니라, 학생이 스스로 질문을 구조화하고 사고를 명료화하는 과정을 경험하도록 설계되었다.

라. 설계 대상: 일반 LLM보다 우수한 학습 효과

[표 3-2] MAICE 시스템의 설계 목표

목표	Freepass 한계	MAICE 해결책	기대 효과
질문 품질 개선	맥락 72.3% 부재	명료화 프로세스	질문의 질 향상
답변 적합성 향상	맥락 오해, 수준 불일치	질문 유형 분류 (K1-K4)	맞춤형 답변
메타인지 향상	사고 과정 부재	Dewey 5단계 유도	자기 성찰 훈련

MAICE 시스템은 위 3가지 측면에서 일반 Freepass 방식보다 우수한 학습 효과를 제공하는 것을 목표로 한다.

이러한 목표를 달성하기 위해, MAICE는 ChatGPT, Claude 등 상용 AI 대화 서비스와 유사한 UX를 제공하되, 수학 학습에 특화된 3계층 구조로 설계되었다.

2. 전체 아키텍처 개요: 3계층 구조

MAICE 시스템은 학생이 사용하는 대화 인터페이스, 대화를 관리하고 데이터를 저장하는 관리 시스템, 그리고 지능적으로 질문을 처리하는 에이전트 시스템의 3계층으로 구성된다.

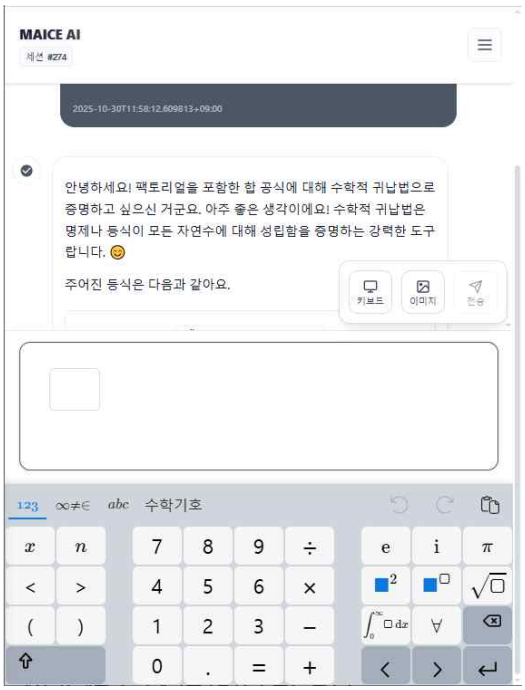
가. 계층별 역할

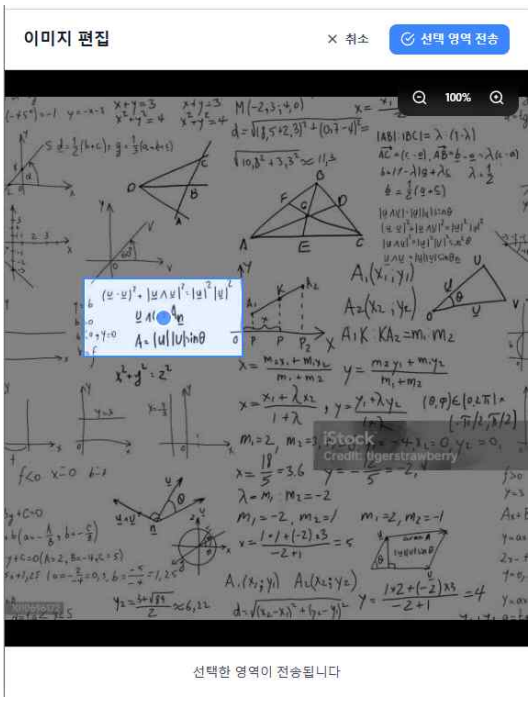
1) 계층 1: 대화 인터페이스 (학생이 보는 화면)

역할: 학생이 수식을 쉽게 입력하고 AI 답변을 실시간으로 받을 수 있는 채팅 화면

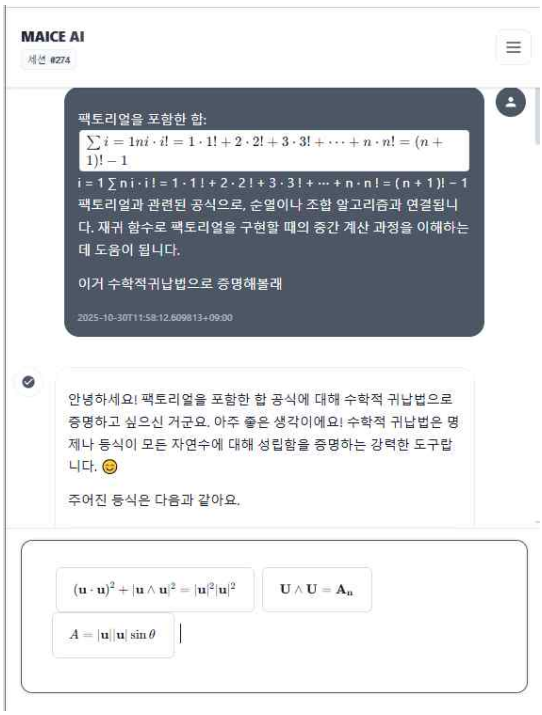
- 주요 기능:
- 수식 입력 지원: 복잡한 수학 수식을 클릭 몇 번으로 입력
- 실시간 답변: AI 답변이 타이핑하듯 실시간으로 표시
- 간편한 로그인: 학교 계정으로 바로 시작
- 모바일 지원: 핸드폰에서도 동일하게 사용 가능
- 학생 경험: ChatGPT, Claude와 동일한 UX로 별도 학습 없이 즉시 사용 가능

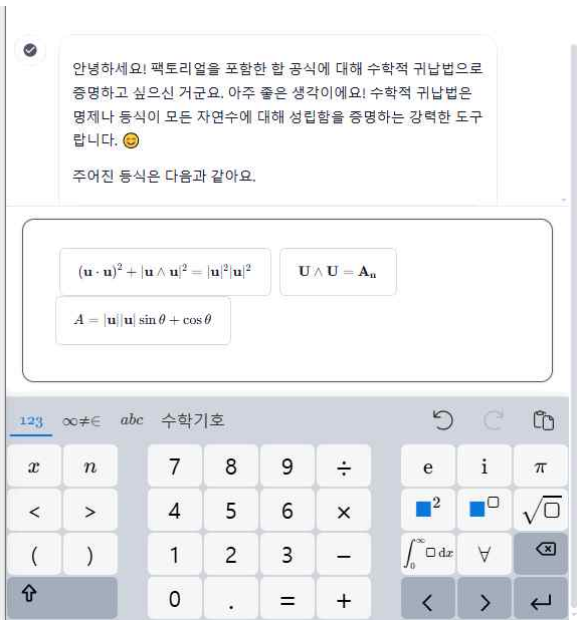
[그림 3-2] MAICE 대화 인터페이스 실제 화면





선택한 영역이 전송됩니다





2) 계층 2: 관리 시스템 (대화 저장 및 분석)

역할: 모든 대화를 체계적으로 저장하고, 교사가 학생 학습 상황을 파악할 수 있도록 지원

◯ 주요 기능:

- 대화 기록 관리: 학생별 모든 대화 세션을 시간 순으로 저장
- 학습 데이터 분석: 학생의 학습 진도, 어려움 영역 자동 추출
- 교사 대시보드: 반 전체 학습 현황 및 개별 학생 상세 정보 제공
- 권한 관리: 학생은 본인 데이터만, 교사는 전체 데이터 접근 가능
- 교육적 가치: 교사가 30명 학생의 개별 학습 상황을 실시간 파악 가능

가) 계층 3: 에이전트 시스템 (지능적 질문 처리)

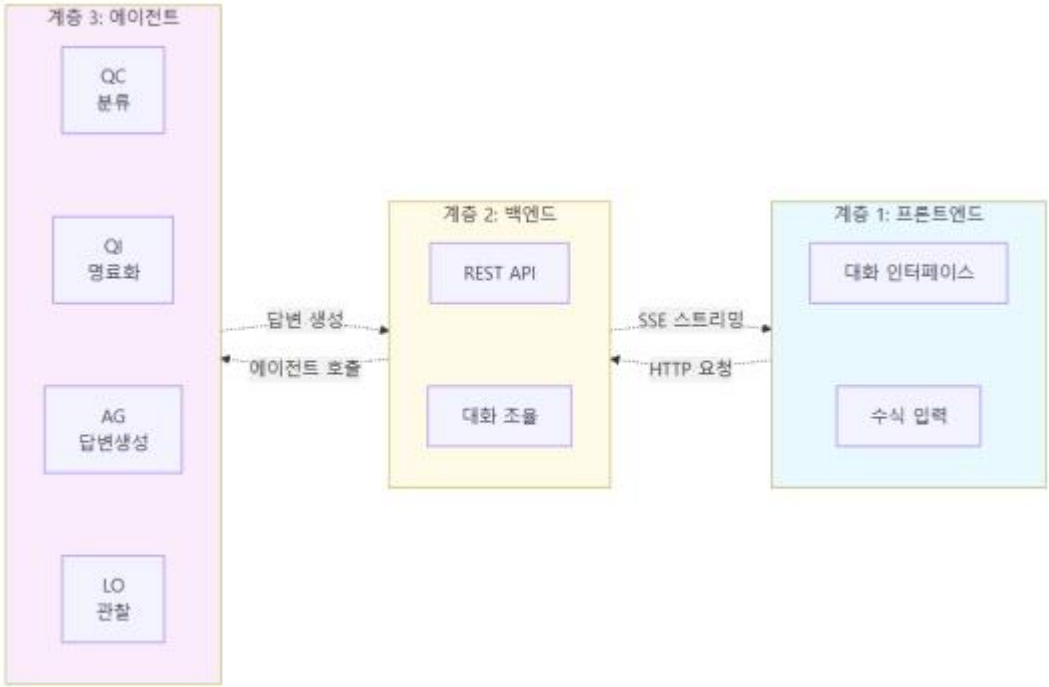
역할: 학생 질문을 분석하고, 필요시 명료화하며, 맞춤형 답변을 생성하는 5개 AI 에이전트

◯ 5개 에이전트:

- Classifier (질문 분류): "이 질문은 어떤 유형인가?"
- Question Improvement (명료화): "질문을 더 명확하게"
- Answer Generator (답변 생성): "유형별 맞춤 답변"
- Observer (학습 관찰): "학생이 무엇을 배우고 있는가?"
- FreeTalker (대조군): "명료화 없이 즉시 답변"

협업 방식: 각 에이전트는 독립적으로 작동하되, 필요한 정보를 서로 주고받으며 협업

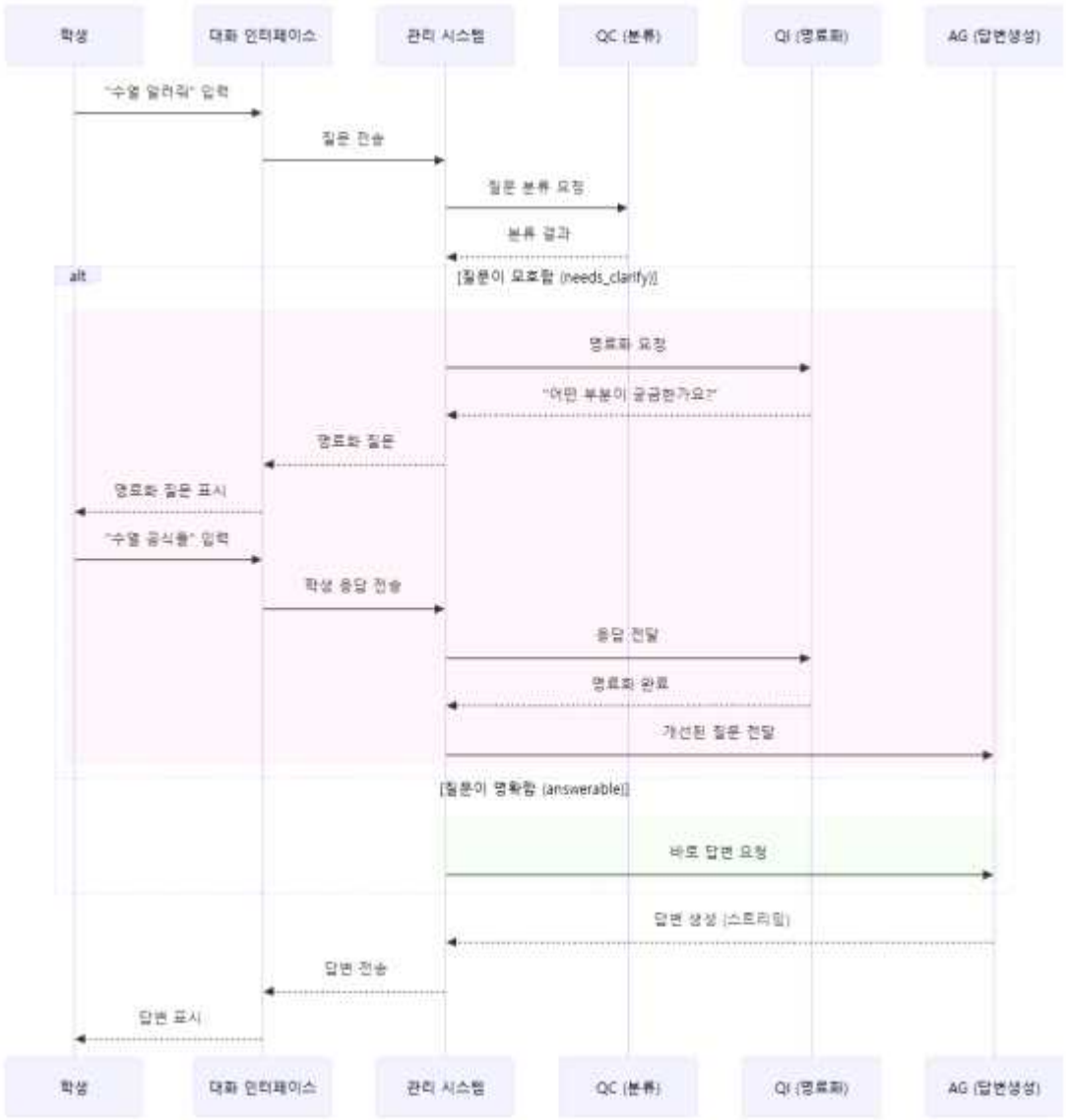
나. 전체 시스템 구조도



[그림 3-3] MAICE 3계층 아키텍처

다. 질문 처리 흐름

학생이 질문을 입력하면 다음과 같은 과정을 거친다:



[그림 3-4] 질문 처리 시퀀스

○ 핵심 특징:

- 학생은 명료화 과정을 자연스러운 대화로 경험
- 모든 대화는 자동으로 저장되어 학습 분석에 활용
- 교사는 별도 대시보드에서 학생 학습 현황 확인
- 기술 구현 상세: 각 계층의 구체적인 기술 스택과 구현 방법은 4장 "시스템 구현"에서 다룬다.

3. 5개 에이전트의 역할과 협업

MAICE 시스템의 핵심은 5개의 독립적인 AI 에이전트가 협업하여 질문을 처리하는 것이다. 각 에이전트는 특정한 교육적 목적을 가지고 설계되었다.

가. Question Classifier (QC): "이 질문은 어떤 유형인가?"

1) 설계 목적

1장에서 확인했듯이, 학생 질문의 72.3%가 학습 맥락 정보 없이 제출되었다. 질문의 인지적 수준에 따라 답변 방식이 달라야 하므로(K1 사실형 vs K4 메타인지형), Classifier는 질문을 자동으로 분류하고 명료화 필요성을 판단한다.

2) 핵심 기능

가) 질문 유형 분류 (K1-K4)

- K1 (사실): "수학적 귀납법의 정의가 뭐에요?" → 간결한 정의 제공
- K2 (개념): "귀납 가정은 왜 필요한가요?" → 개념 간 관계 설명
- K3 (절차): "이 등식을 어떻게 증명하나요?" → 단계별 절차 안내
- K4 (메타인지): "제가 뭘 잘못 이해한 건가요?" → 메타인지 유도 답변

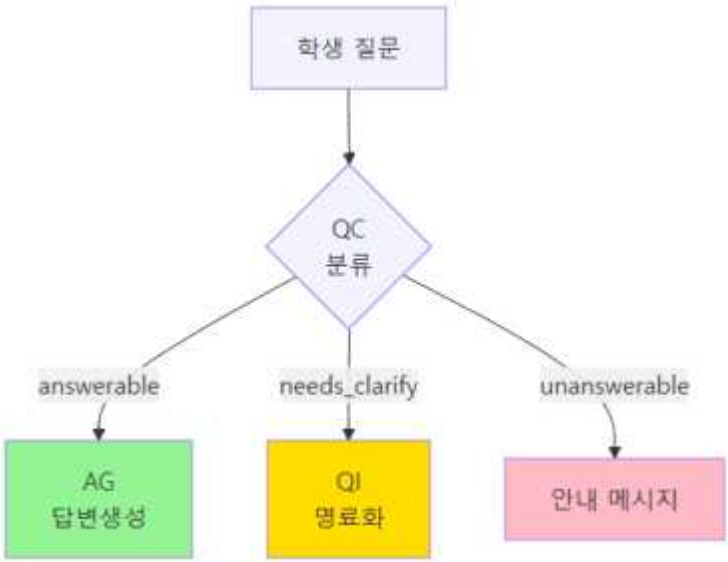
나) 명료화 필요성 판단 (3단계 게이팅)

- answerable: 질문이 명확함 → Answer Generator로 전달
- needs_clarify: 질문이 모호함 → Question Improvement로 전달
- unanswerable: 답변 불가능 → 정중한 안내

다) 명료화 질문 제안

- needs_clarify인 경우, Dewey 5단계 기반 구체적 질문 제안
- 예: "수열 알려줘" → "어떤 부분이 궁금한지 알려줄래?

3) 설계 흐름



[그림 3-5] QC 3단계 게이팅

나. Question Improver (QI): "질문을 더 명확하게"

1) 설계 목적

QC가 "명료화 필요" 판정을 내리면, QI는 학생이 스스로 질문을 구체화하도록 돕는다. Dewey의 반성적 사고 이론에 따르면, 학생이 자신의 어려움을 명확히 인식하고 표현하는 과정에서 메타인지 능력이 향상된다.

2) 핵심 설계: Dewey 5단계 기반 명료화

QI는 Dewey의 5단계를 대화형 질문으로 변환하여, 질문 유형(K1-K4)과 모호성 수준에 따라 1-3회 명료화를 수행한다:

[표 3-9] Dewey 5단계 기반 명료화 전략

Dewey 단계	명료화 질문 예시	학생 경험
1단계: 문제 인식	"무엇이 불확실한가요?"	막연한 어려움 → 구체적 문제 인식
2단계: 문제 정의	"정확히 무엇을 알고 싶은가요?"	"어려워요" → "귀납 단계가 어려워요"
3단계: 가설 설정	"어떤 방법을 시도해봤나요?"	학생의 시도와 이해 수준 파악
4단계: 가설 검증	"왜 그렇게 생각했나요?"	학생의 사고 과정 드러내기
5단계: 결론 도출	"최종적으로 무엇을 얻고 싶나요?"	학습 목표 명확화

3) 명료화 전략: 질문 유형별 차별화

질문 유형(K1-K4)과 모호성 수준에 따라 명료화 전략을 조절한다:

- K1 (즉답형): 선택지 제공으로 빠르게 범위 좁히기 (1회)
- K2 (설명형): 비교 대상이나 설명 깊이 확인 (1-2회)
- K3 (적용형): 구체적인 문제 상황이나 막힌 단계 파악 (1-2회)
- K4 (문제해결형): Dewey 5단계를 깊이 적용, 사고 과정 드러내기 (2-3회)

4) 명료화 완료 판단

Question Improvement는 학생의 각 응답을 평가하여, 충분한 정보가 모였는지 판단한다

[그림 3-6] 명료화 프로세스 흐름

◯ 판단 기준:

- PASS: 원본 질문의 의도가 명확해지고, 답변 생성에 필요한 정보 확보
- NEED_MORE: 원본 질문의 의도가 여전히 불분명하거나, 추가 정보 필요
- 최대 3회 제한: 3회 명료화 후에도 불충분하면 현재 정보로 답변 생성

5) 교육적 의도 명시화

기존 LLM은 "왜 명료화 질문을 하는지" 설명하지 않아 학생이 불편함을 느꼈다. Question Improvement는 명료화의 교육적 이유를 부드럽게 설명한다:

◯ 기본 프레이밍:

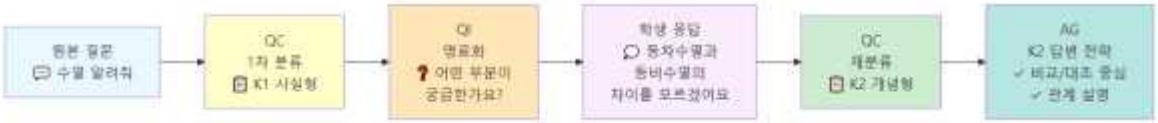
- 질문을 조금만 더 구체적으로 만들어주면, 딱 맞는 설명을 해드릴 수 있어요!

◯ 교육적 프레이밍 (K4 수준 학생):

- 함께 질문을 구체화해볼까요? 정확히 무엇을 모르는지 찾아가는 과정이 진짜 학습의 시작이에요!

6) 질문 유형 재분류

명료화 과정에서 학생의 실제 어려움이 드러나면, 질문 유형이 변경될 수 있다:



[그림 3-7] 질문 유형 재분류 예시

다. Answer Generator (AG): "유형별 맞춤 답변"

1) 설계 목적

1장에서 확인한 AI 답변의 문제는 "모든 학생에게 동일한 방식으로 설명"하여 인지 과부하를 유발했다. AG는 Bloom의 K1-K4 분류에 따라 답변 구조와 교수법을 차별화한다.

2) 핵심 설계: K1-K4별 답변 차별화

[표 3-10] K1-K4별 답변 구조 및 교수법

유형	답변 구조	교수법 특징
K1 (사실)	정의 → 핵심 예시 → 보충	간결함, 정확성 우선 (3-5문장)
K2 (개념)	개념 관계 → 비교/대조 → 시각화	"왜?" 중심 설명, 논리적 연결
K3 (절차)	전체 개요 → 단계별 안내 → 실수 방지	선택권 제공, 대화형 진행
K4 (메타인지)	문제 분석 → 자기 점검 → 대안 탐색	답 직접 제공 X, 사고 과정 유도

3) 주요 설계 특징

◯ 교육과정 표준 준수: 대한민국 교육과정 표준 용어 사용

◯ 실시간 스트리밍: 답변을 타이핑하듯 실시간 전송하여 학생이 자연스러운 대화 경험

◯ LaTeX 수식 렌더링: 인라인($수식$) 및 블록($$수식$$) 수식 지원

라. Learning Observer (LO): "대화 요약 및 컨텍스트 관리"

1) 설계 목적

긴 대화 세션에서 컨텍스트 길이 증가와 맥락 손실을 방지하기 위해, Observer는 대화를 주기적으로 요약하여 핵심 내용만 유지한다.

2) 핵심 기능

○ 대화 요약: 15회 턴 이상 시 자동 요약 (핵심 주제, 학습 진행, 질문 유형, 현재 상태)

○ 맥락 유지: 토큰 효율성 확보 + 학습 연속성 지원

[표 3-3] LO 관찰 및 추출 정보

유형	답변 구조
K1 (사실)	정의 → 핵심 예시 → 보충
K2 (개념)	개념 관계 → 비교/대조 → 시각화
K3 (절차)	전체 개요 → 단계별 안내 → 실수 방지
K4 (메타인지)	문제 분석 → 자기 점검 → 대안 탐색

마. Free Talker (FT): "대조군 (Freepass 모드)"

1) 설계 목적

명료화 프로세스의 효과를 검증하기 위한 A/B 테스트 대조군. 일반 ChatGPT, Claude처럼 명료화 없이 즉시 답변한다.

2) 핵심 특징

○ 즉시 답변: 질문 분류, 명료화 없이 바로 답변 생성

○ A/B 테스트: 학생 58명을 무작위로 Agent(28명) / Freepass(30명) 배정

[표 3-4] Agent vs Freepass 모드 기능 비교

항목	Agent 모드	Freepass 모드
질문 분류	✓ K1-K4 분류	✗ 분류 없음
명료화	✓ needs_clarify 시 명료화	✗ 명료화 없음
답변 전략	✓ K1-K4별 차별화	✗ 동일한 방식

◯ 대조군의 중요성:

- Agent 모드의 효과를 인과적으로 검증 가능
- 선택 편향(selection bias) 제거
- "명료화가 정말 도움이 되는가?" 실증적 답변 제공

3) 검증 계획

FT를 활용한 A/B 테스트를 통해 다음을 비교 검증할 예정이다:

◯ 비교 차원:

- 즉시 효과: 단일 세션에서 질문-답변 품질 비교
- 누적 효과: 다회 세션 시 학습 진행 패턴 비교
- 학습자 수준별 효과: 상위권/하위권 학생에 대한 차별적 효과 분석

◯ 기대 가설:

- 명료화 프로세스는 단일 세션에서는 시간이 더 소요되나, 장기적으로 질문 능력과 메타인지 향상에 기여할 것
- 특히 절차적 지식이 부족한 하위권 학생들에게 더 큰 학습 효과가 나타날 것

검증 결과는 8장에서 상세히 분석한다.

IV. MAICE 시스템 구현

본 장은 3장의 교육적 설계를 실제 작동하는 시스템으로 구현한 기술적 측면을 다룬다. 교육적 근거와 설계 원리는 3장을 참조하고, 여기서는 기술 스택, 코드 구조, 배포 전략에 집중한다. 재현 가능성(reproducibility)과 확장 가능성(scalability)을 확보하기 위한 구현 세부사항을 기술한다.

1. 기술 스택 개요

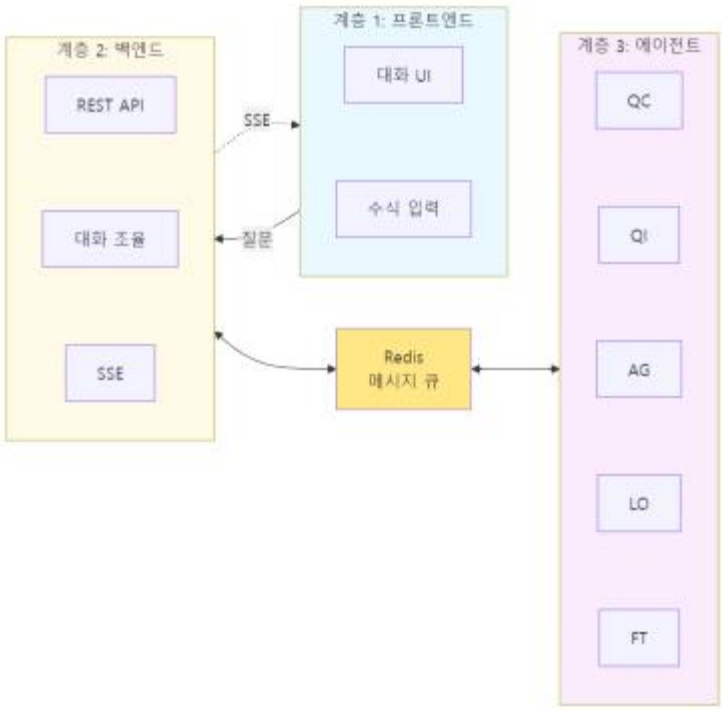
MAICE 시스템은 3계층 마이크로서비스 아키텍처로 구현되었다 (교육적 설계 구조는 3장 3.2절 참조). 각 계층의 기술 스택은 다음과 같다:

가. 계층별 기술 스택

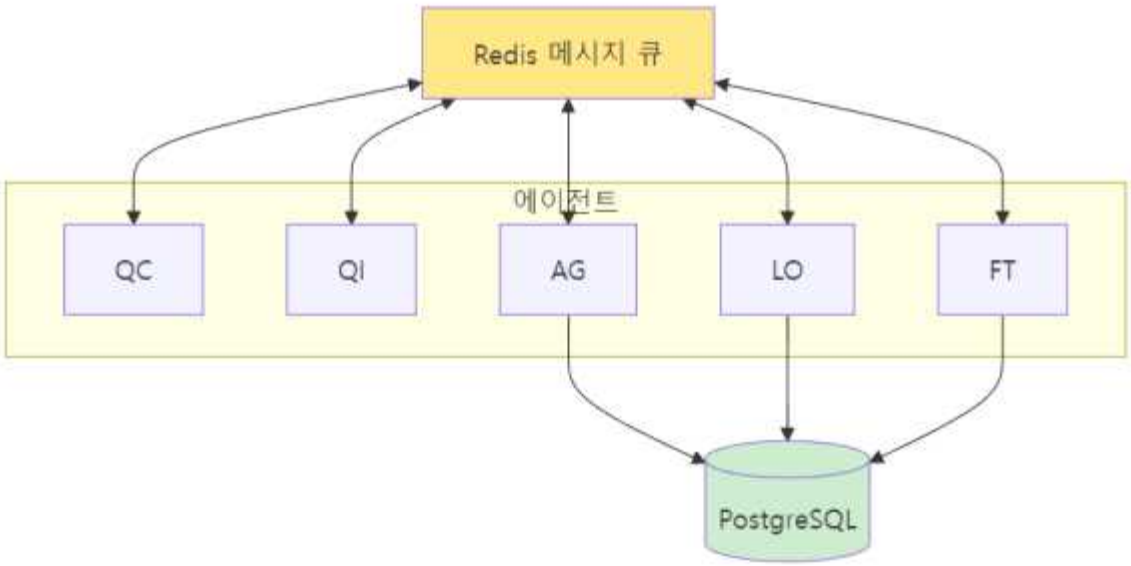
계층	핵심 기술	버전	선택 이유
프론트엔드	SvelteKit	2.0	반응형 UI, SSE 지원, 빠른 개발
	MathLive	0.95	수식 입력 전문 라이브러리
	Tailwind CSS	3.3	유틸리티 기반 디자인 시스템
백엔드	FastAPI	0.104	비동기 지원, 타입 힌트, 자동 문서화
	SQLAlchemy	2.0	ORM, 비동기 쿼리
에이전트	LLM API	-	Gemini 2.5 Flash Lite
	Redis Streams	7.0	비동기 메시지 큐
데이터베이스	PostgreSQL	15	JSONB 지원, 관계형 무결성
	Redis	7.0	세션 캐싱, 메시지 큐
배포	Docker Compose	-	컨테이너 오케스트레이션
	Nginx	1.25	리버스 프록시, HTTPS

[표 4-1] 계층별 기술 스택 및 선택 이유

나. 기술 아키텍처 다이어그램



[그림 4-1] MAICE 시스템 3계층 구조 (데이터 흐름)



[그림 4-2] 에이전트-Redis-DB 데이터 흐름

[표 4-2] 계층별 주요 기술 스택 및 통신 프로토콜

계층	핵심 기술	통신 방식	역할
계층 1: 프론트엔드	SvelteKit 2.0, MathLive	HTTP/SSE	학생 UI, 수식 입력
계층 2: 백엔드	FastAPI 0.104, Python 3.11	REST API, Redis	API 제공, 에이전트 조율
계층 3: 에이전트	Gemini 2.5 Flash, asyncio	Redis pub/sub	질문 분류, 명료화, 답변
데이터 계층	PostgreSQL 15, Redis 7	ORM, Streams	영구 저장, 메시지 큐

다. 핵심 설계 원칙 및 A/B 테스트 구조

[표 4-3] Agent vs Freepass 모드 비교

항목	Agent 모드	Freepass 모드
에이전트 수	4개 (QC→QI→AG→LO)	2개 (FT→LO)
명료화 과정	✓ Dewey 5단계 기반	✗ 생략
질문 분류	✓ K1-K4 분류	✗ 생략
교육적 개입	✓ 메타인지 유도	✗ 즉시 답변

○ 핵심 설계 원칙:

- 비동기 처리: 모든 I/O는 asyncio로 비동기 처리 → 동시 처리 능력 극대화
- 느슨한 결합: Redis pub/sub으로 에이전트 간 직접 의존성 제거 → 독립성 보장
- 이벤트 기반 아키텍처: 메시지 큐로 작업 분산 → 확장 가능한 구조
- 완전한 재현성: 모든 대화, 프롬프트, 응답을 DB에 기록 → 연구 검증 가능
- A/B 테스트 지원: 사용자 모드에 따라 다른 에이전트 경로 자동 라우팅

2. 계층별 구현 상세

가. 프론트엔드 계층 (front/)

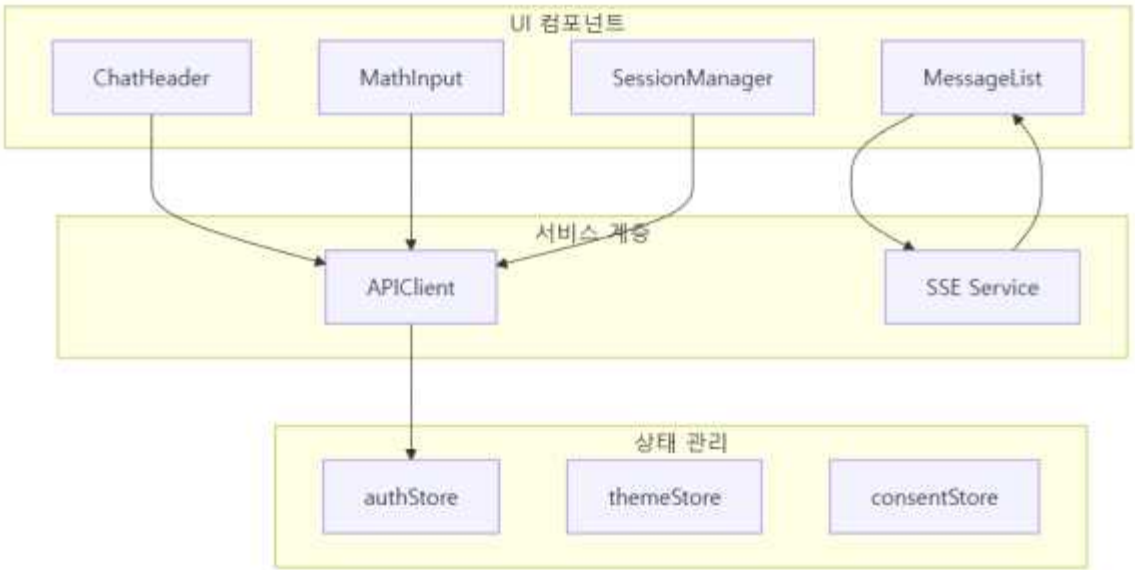
1) 기술 스택

◯ 프레임워크: SvelteKit 2.0 (Vite 기반)
◯ 언어: TypeScript 5.x
◯ UI 라이브러리:

- Tailwind CSS 4.x: 디자인 시스템
- MathLive 0.95: 수학 수식 입력/렌더링
- KaTeX: 정적 수식 렌더링

◯ 상태 관리: Svelte Stores (Reactive)

2) 주요 컴포넌트 아키텍처



[그림 4-3] 프론트엔드 컴포넌트 구조

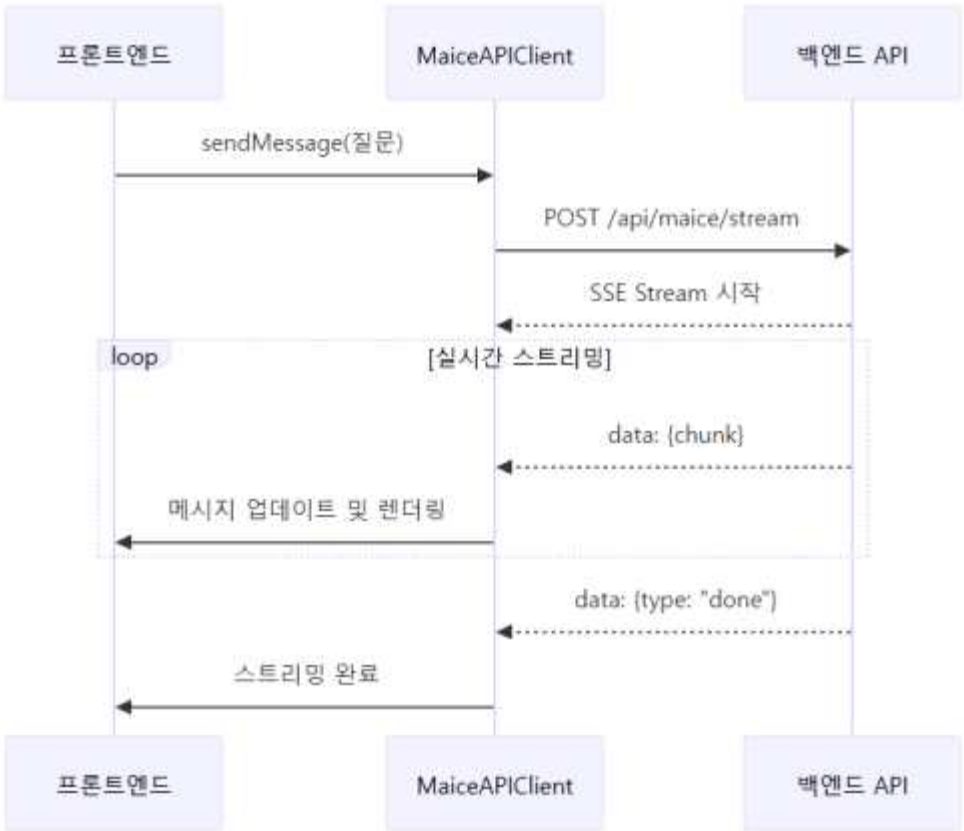
◯ InlineMathInput의 핵심 역할:

- 수식 입력: MathLive 기반 LaTeX 에디터
- 이미지 OCR: 수식 사진 → LaTeX 자동 변환
- 실시간 미리보기: 입력과 동시에 렌더링
- 모바일 최적화: 터치 가상 키보드 지원

◯ MessageList의 역할:

- 실시간 스트리밍: SSE로 받은 답변 청크를 실시간 표시
- LaTeX 렌더링: KaTeX로 수식 자동 렌더링
- 자동 스크롤: 새 메시지 추가 시 자동 하단 스크롤

3) 프론트엔드 ↔ 백엔드 통신



[그림 4-4] 프론트엔드-백엔드 SSE 스트리밍 통신

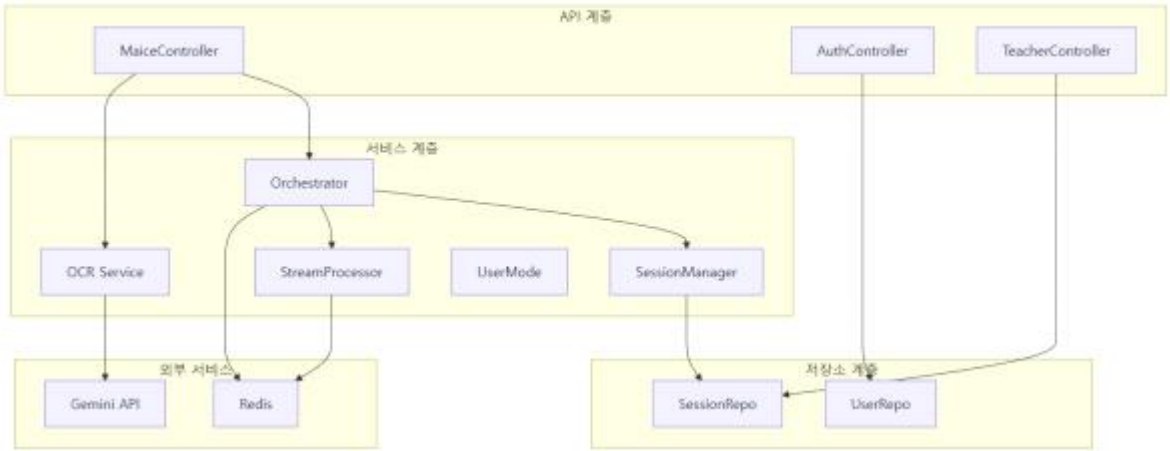
나. 백엔드 계층 (back/)

1) 기술 스택

◯ 프레임워크: FastAPI 0.104
◯ 언어: Python 3.11
◯ ORM: SQLAlchemy 2.0 (비동기)
◯ 데이터베이스: PostgreSQL 15
◯ 메시지 브로커: Redis 7 (Streams + pub/sub)
◯ 비동기: asyncio + uvloop

2) 서비스 아키텍처

[그림 4-5] 백엔드 서비스 계층 구조



◯ ConversationOrchestrator의 핵심 역할:

- 질문을 받아 Redis Streams에 발행
- 에이전트 응답을 구독하여 프론트엔드로 전달
- 세션별 대화 상태 추적

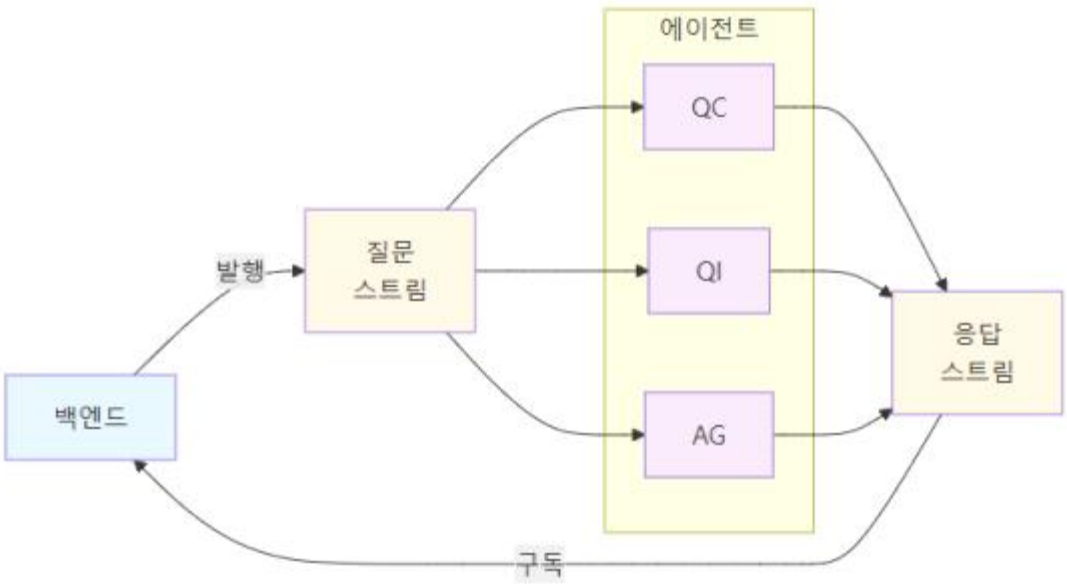
○ SessionManager의 역할:

- 세션 생성/조회/삭제
- 대화 메시지 저장 (PostgreSQL JSONB)
- 이전 대화 히스토리 제공

○ ImageToLatexService의 역할 (3.6.3절 OCR 시스템):

- 이미지 파일 검증 및 전처리
- Gemini Vision API 호출
- LaTeX 정제 및 MathLive 호환성 변환

3) Redis Streams 통신 구조



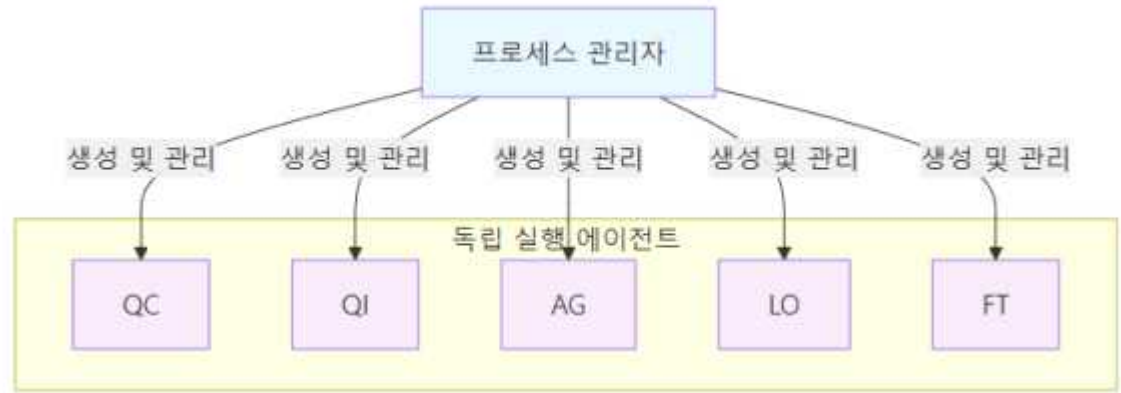
[그림 4-6] Redis Streams 메시지 전달 구조

[표 4-4] Redis Streams 메시지 구조

필드	타입	설명
session_id	int	세션 고유 ID
user_id	int	사용자 ID
question	str	학생 질문 내용
mode	str	"agent" 또는 "freepass"
conversation_history	JSON	이전 대화 히스토리
timestamp	ISO8601	메시지 생성 시각

다. 에이전트 계층 (agent/)

1) 멀티프로세스 아키텍처



[그림 4-7] 에이전트 멀티프로세스 구조 (독립 실행)

○ 멀티프로세스 설계 이유:

- 독립성: 한 에이전트 장애가 다른 에이전트에 영향 없음
- 병렬 처리: 각 에이전트가 동시에 다른 세션 처리 가능
- 자동 재시작: 프로세스 감시자(Supervisor)가 장애 시 자동 재시작
- 확장성: 에이전트별로 프로세스 수 증가 가능

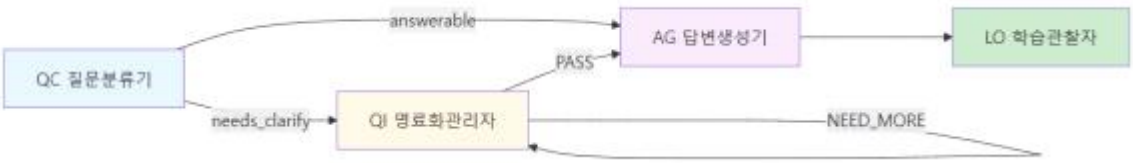
2) BaseAgent 공통 구조

[그림 4-8] BaseAgent 공통 구조 및 상속 관계

○ BaseAgent가 제공하는 공통 기능:

- initialize(): Redis/PostgreSQL 연결 초기화
- run_subscriber(): 메시지 큐 구독 무한 루프
- cleanup(): 리소스 정리 및 연결 종료
- check_duplicate_request(): 중복 요청 방지
- save_prompt_log(): 모든 LLM 호출 기록

3) 에이전트 간 협업 메커니즘



[그림 4-9] 에이전트 간 협업 메커니즘 (Redis pub/sub 기반)

○ 협업 흐름:

- QC → AG: answerable (즉시 답변 가능한 경우)
- QC → QI: needs_clarify (명료화 필요한 경우)
- QI → AG: PASS (명료화 완료)
- QI → QI: NEED_MORE (추가 명료화 필요, 최대 3회)
- AG → LO: 답변 완료 후 학습 패턴 분석

○ Redis pub/sub 이벤트:

- NEED_CLARIFICATION: QC → QI
- READY_FOR_ANSWER: QI → AG
- GENERATE_SUMMARY: AG → LO

3. 프롬프트 관리 시스템

가. YAML 기반 프롬프트 설정

모든 에이전트의 프롬프트는 YAML 파일로 관리되어 코드 수정 없이 프롬프트만 변경 가능하다.

○ 프롬프트 구조 (agents/question_classifier/config.yaml):

```yaml
system_prompt:|
당신은 대한민국 고등학교 수학 교육과정 전문 분류기입니다.
질문을 정확히 분석하여 4가지 유형과 3단계 품질로 분류하세요.
knowledge_types:
  K1:
    name:"사실적 지식"
    description:"정의, 용어, 기호, 공식, 값, 단위"
  K2:
    name:"개념적 지식"
    description:"개념 간 관계, 분류, 원리, 이론"
```

```
  K3:
    name:"절차적 지식"
    description:"수행 방법, 알고리즘, 단계별 과정"
  K4:
    name:"메타인지적 지식"
    description:"전략적 사고, 문제 접근법, 반성"
quality_gates:
  answerable:"교과·단원·수준 지정, 목표 동사 명확"
  needs_clarify:"범위 과대/목표 불명/수준 불명"
  unanswerable:"수학 외 영역, 평가윤리 위배"
output_format:
  required_fields:
    -knowledge_code
    -quality
    -reasoning
    -clarification_questions
```

◯ PromptBuilder의 역할:

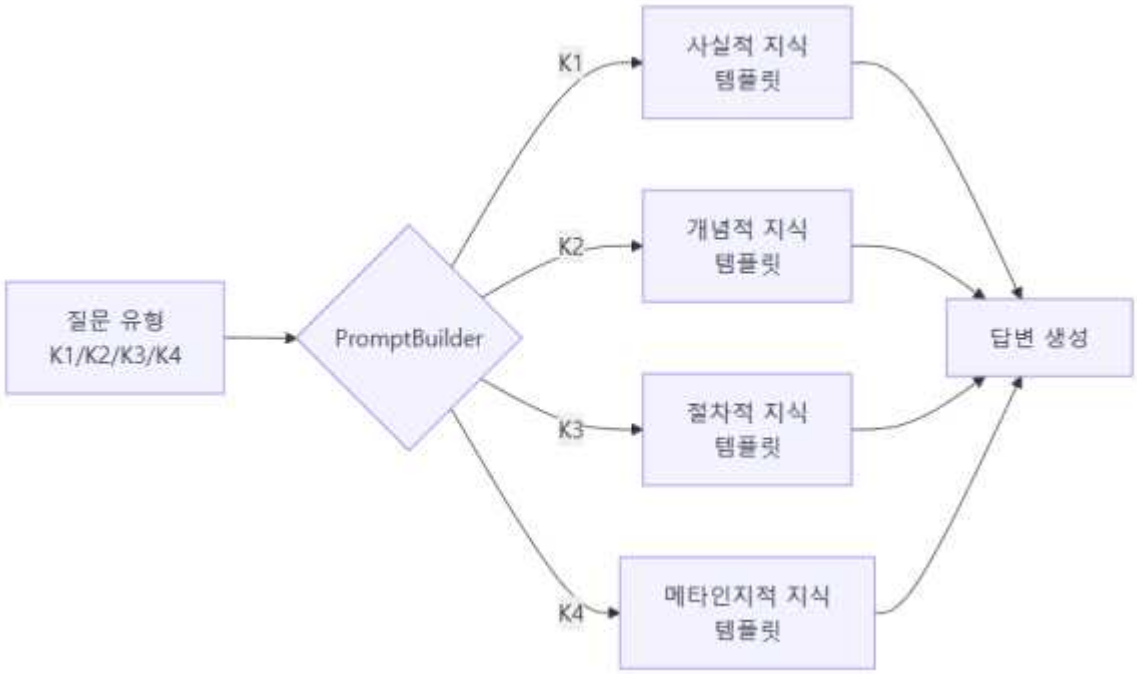
- YAML 파일을 로드하여 캐싱
- 변수를 동적으로 치환 (예: {question}, {context})
- 에이전트별 프롬프트 템플릿 관리

◯ 장점:

- 유지보수성: 프롬프트 수정 시 코드 재배포 불필요
- 버전 관리: YAML 파일을 Git으로 버전 관리
- A/B 테스트: 프롬프트 변형을 쉽게 테스트 가능

나. K1-K4별 답변 전략 구현

AG는 질문 유형에 따라 다른 프롬프트 템플릿을 사용한다:



[그림 4-10] K1-K4별 프롬프트 템플릿 선택 로직

○ 템플릿 선택 로직 (answer_generator/agent.py):

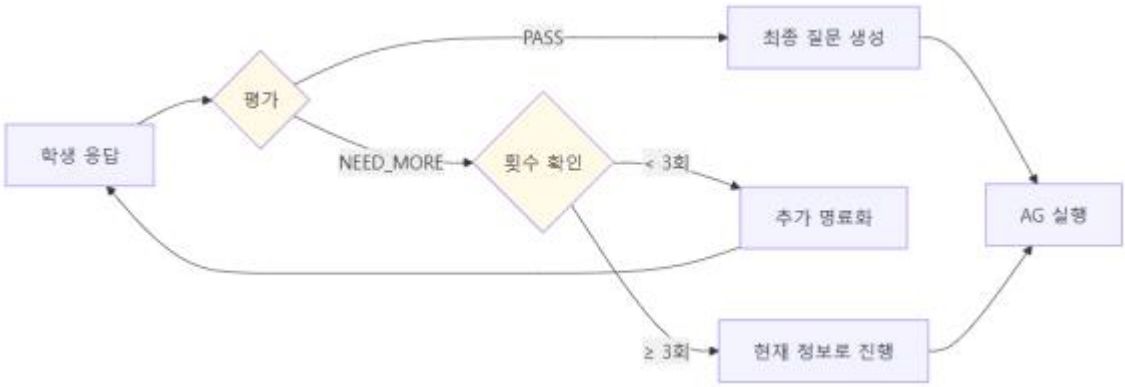
```
# 질문 유형에 따른 템플릿 선택
template_name =f"answer_{knowledge_code.lower()}"
# "answer_k1", "answer_k2" 등
prompt =await self.prompt_builder.build(template_name,{
    "question":final_question,
    "context":context,
    "clarification_history":clarification_history
})
```

○ 각 템플릿의 구조적 차이:

- K1: 정의 → 예시 → 보충 (간결함 우선)
- K2: 개념 관계 → 비교 → 시각화 (연결성 강조)
- K3: 전체 개요 → 단계별 안내 → 실수 방지 (절차 명확성)
- K4: 문제 분석 → 다양한 접근 → 자기 점검 (메타인지 자극)

다. 명료화 평가 로직

QI Agent는 학생 응답을 평가하여 PASS/NEED_MORE를 판단한다:



[그림 4-11] 명료화 평가 로직 (최대 3회 제한)

○ 평가 기준:

- PASS: 원본 질문의 의도가 명확해짐
- NEED_MORE: 여전히 모호하거나 추가 정보 필요
- 최대 3회 제한: 무한 명료화 방지

라. 에이전트별 프롬프트 전문

본 시스템의 교육적 효과는 각 에이전트의 프롬프트 설계에 크게 의존한다. 이 절에서는 실제 운영 환경에서 사용된 핵심 프롬프트를 제시한다.

1) QC (Question Classifier) 프롬프트

System 메시지:

당신은 대한민국 고등학교 수학 교육과정 전문 분류기입니다.
질문을 정확히 분석하여 4가지 유형과 3단계 품질로 분류하고,
필요한 경우 **학생에게 직접 묻는** 명료화 질문까지 생성하세요.

질문 유형 (K1-K4)
- K1 (즉답형): 정의, 용어, 기호, 공식, 값, 단위
- K2 (설명형): 개념 간 관계, 분류, 원리, 이론
- K3 (적용형): 수행 방법, 알고리즘, 단계별 과정
- K4 (문제해결형): 전략적 사고, 문제 접근법, 반성
품질 평가
- answerable: 교과·단원·수준 지정, 목표 동사 명확
- needs_clarify: 범위 과대/목표 불명/수준 불명
- unanswerable: 수학 외 영역, 평가윤리 위배
명료화 질문 생성 원칙 (Dewey 5단계)
1. 문제 인식: "어떤 부분이 가장 어렵거나 궁금하셨나요? 🤔"
2. 문제 정의: "지금까지 이해한 부분과 헷갈리는 부분을 나누어볼까요?"
3. 연결 탐색: "이미 알고 있는 개념과 비교하면 어떤 점이 다른가요?"
4. 사고 전개: "왜 이 부분이 궁금하신지 조금 더 설명해주실 수 있나요?"
5. 이해 검증: "어디까지 이해했고, 어디서부터 막히셨는지 말씀해주세요"

⚠ 명료화 질문은 학생이 직접 읽고 답변할 수 있는 자연스러운 질문이어야 합니다!
❌ 시스템 분석: "'나'라는 답변이 구체적으로 무엇을 의미하는지 확인 필요"
✅ 학생 질문: "어떤 부분이 더 궁금하신가요? 😊"

출력 형식 (JSON)
```
{
  "knowledge_code": "K1/K2/K3/K4",
  "quality": "answerable/needs_clarify/unanswerable",
  "missing_fields": ["부족한 정보1", "부족한 정보2"],
  "reasoning": "분류 근거",
  "clarification_questions": ["학생에게 직접 묻는 자연스러운 질문 1개"],
  "clarification_reasoning": "명료화 질문이 어떻게 missing_fields를 해결하는지"
}
```

○ 설계 근거 (3장 3.3.1 참조):

- Dewey 반성적 사고 5단계를 명료화 질문 전략으로 구현
- Bloom K1-K4 분류로 질문 유형 차별화
- 학생 친화적 톤 ("🥲", 존댓말)으로 심리적 장벽 낮춤

2) QI Agent 프롬프트

System 메시지:

```
---------------------------------------------------------------------

당신은 명료화 과정을 평가하는 전문가입니다.
학생의 답변이 원본 질문을 충분히 명료하게 만들었는지 판단하고,
명료화가 완료되면 최종 질문을 생성하세요.

## 평가 기준
- PASS: 원본 질문의 의도가 명확해짐, 답변 생성 가능
- NEED_MORE: 여전히 모호함, 추가 정보 필요
## 명료화 생략 기준
- 원본 질문이 이미 구체적인 경우 → 즉시 PASS
- 학생이 구체적인 답변을 한 경우 → 즉시 PASS
- 맥락이 명확한 경우 → PASS
- 명료화 3회 접근 시 → 관대하게 PASS
## 출력 형식 (JSON)
{
  "evaluation": "PASS/NEED_MORE",
  "confidence": 0.0-1.0,
  "reasoning": "평가 근거",
  "missing_field_coverage": {
    "해결된_필드": ["필드1"],
    "여전히_부족한_필드": ["필드3"]
  },
  "next_clarification": "다음 명료화 질문 (NEED_MORE인 경우)",
  "reclassified_knowledge_code": "K1/K2/K3/K4 (변경된 경우)",
  "final_question": "최종 질문 (PASS인 경우)"
}
---------------------------------------------------------------------
```

○ 설계 근거 (3장 3.3.2 참조):

- 명료화 완료 시점을 명확히 판단하여 과도한 명료화 방지
- 최대 3회 제한으로 학습 흐름 유지
- missing_fields 추적으로 명료화 진행 상황 관리

3) AG Agent 프롬프트 (K1 예시)

System 메시지:

```
---------------------------------------------------------------------

당신은 대한민국 고등학교 수학 교육과정 전문가입니다.
학생의 질문에 대해 체계적이고 교육적인 답변을 생성해주세요.

## 기본 역할
- 대상: 고등학교 2학년 학생
- 언어: 한국어, 존댓말 필수
- 톤: 친근하고 이해하기 쉬운 교사 톤
- 용어: 대한민국 고등학교 수학 교과서 표준 용어
## 현재 질문 유형: K1 (즉답형 - 사실적 지식)
## K1 답변 구조
1. 핵심 내용 정리: 정확한 정의와 기본 개념
2. 핵심 공식과 정리: 필요한 수식 (LaTeX 형식)
3. 실제 예시로 이해하기: 구체적인 예시
4. 더 넓게 알아보기: 연관 개념
## 수학 수식 작성 규칙
- 인라인 수식: $수식$
- 블록 수식: $$수식$$
- 분수: \frac{a}{b}
- 지수: x^2 또는 x^{지수}
- 제곱근: \sqrt{x}
⚠ 수식과 텍스트 분리!
✅ 올바른: "$P(k)$가 참이면 $P(k+1)$도 참"
❌ 잘못된: "$$P(k)가 참 \Rightarrow P(k+1)도 참$$"

⚠ 중요: 학생에게는 질문 유형 코드(K1, K2, K3, K4)나
분류 정보를 절대 언급하지 마세요.
---------------------------------------------------------------------
```

○ K2/K3/K4 템플릿:

- K2 (설명형): 개념 정리 → 개념 간 연결 → 비교 → 헷갈리는 부분
- K3 (적용형): 단계별 해결 과정 → 사용 시점 → 실제 연습 → 실수 방지
- K4 (문제해결형): 문제 분석 → 다양한 접근 → 중간 점검 → 다른 방법

○ 설계 근거 (3장 3.3.3 참조):

- Bloom K1-K4에 맞춘 차별화된 답변 구조
- 교과서 표준 용어로 학습 일관성 유지
- LaTeX 수식으로 수학적 정확성 확보

4) FT (Free Talker) 프롬프트

System 메시지:

필요할 때만 수학 수식을 LaTeX 형식($수식$)으로 작성해주세요.

User 메시지:

사용자: (학생의 질문)

AI: (이전 답변 - 대화 히스토리)

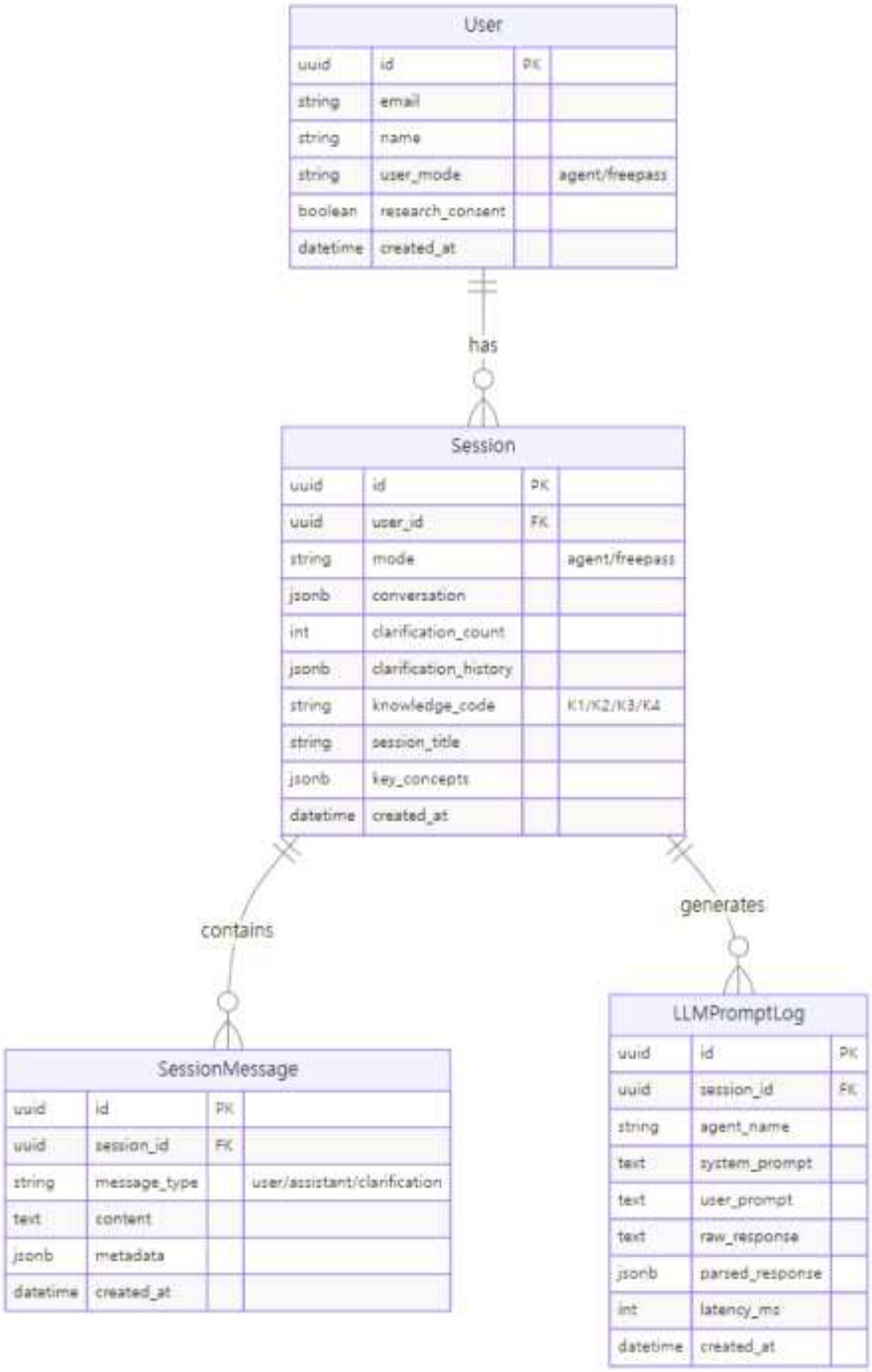
사용자: (현재 질문)

○ 설계 근거 (3장 3.3.5 참조):

- 명료화 없이 즉시 답변하여 Agent 모드와 대조
- 최소한의 프롬프트로 일반 LLM 사용 방식 재현
- A/B 테스트 대조군 역할

4. 데이터 저장 및 분석

가. PostgreSQL 데이터 모델



[그림 4-12] PostgreSQL 데이터베이스 스키마 (재현성 확보)

○ JSONB 활용:

- conversation: 전체 대화 히스토리를 JSON으로 저장
- clarification_history: 명료화 Q&A 배열
- key_concepts: LO가 추출한 주요 개념 목록

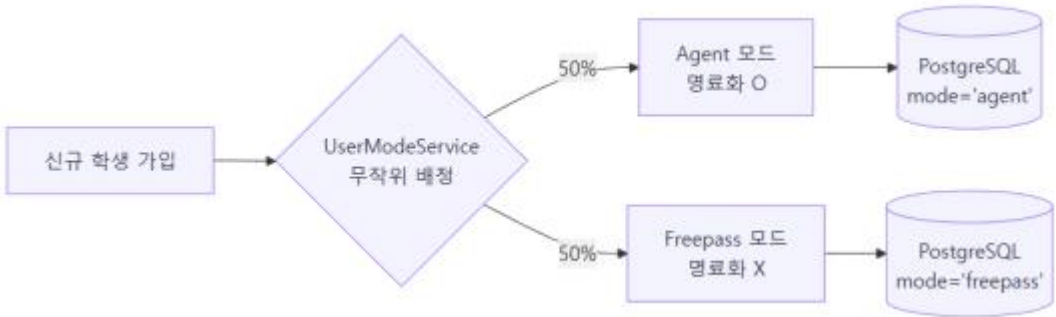
○ 장점:

- 유연한 스키마 (대화 구조 변경에 강함)
- 강력한 쿼리 (JSON 필드 검색 가능)
- 완전한 재현성 (모든 프롬프트와 응답 보존)

나. A/B 테스트 데이터 수집

○ UserModeService의 역할:

- 학생 가입 시 무작위로 "agent" 또는 "freepass" 모드 할당
- 모드는 전체 연구 기간 동안 고정
- 할당 비율: 50:50



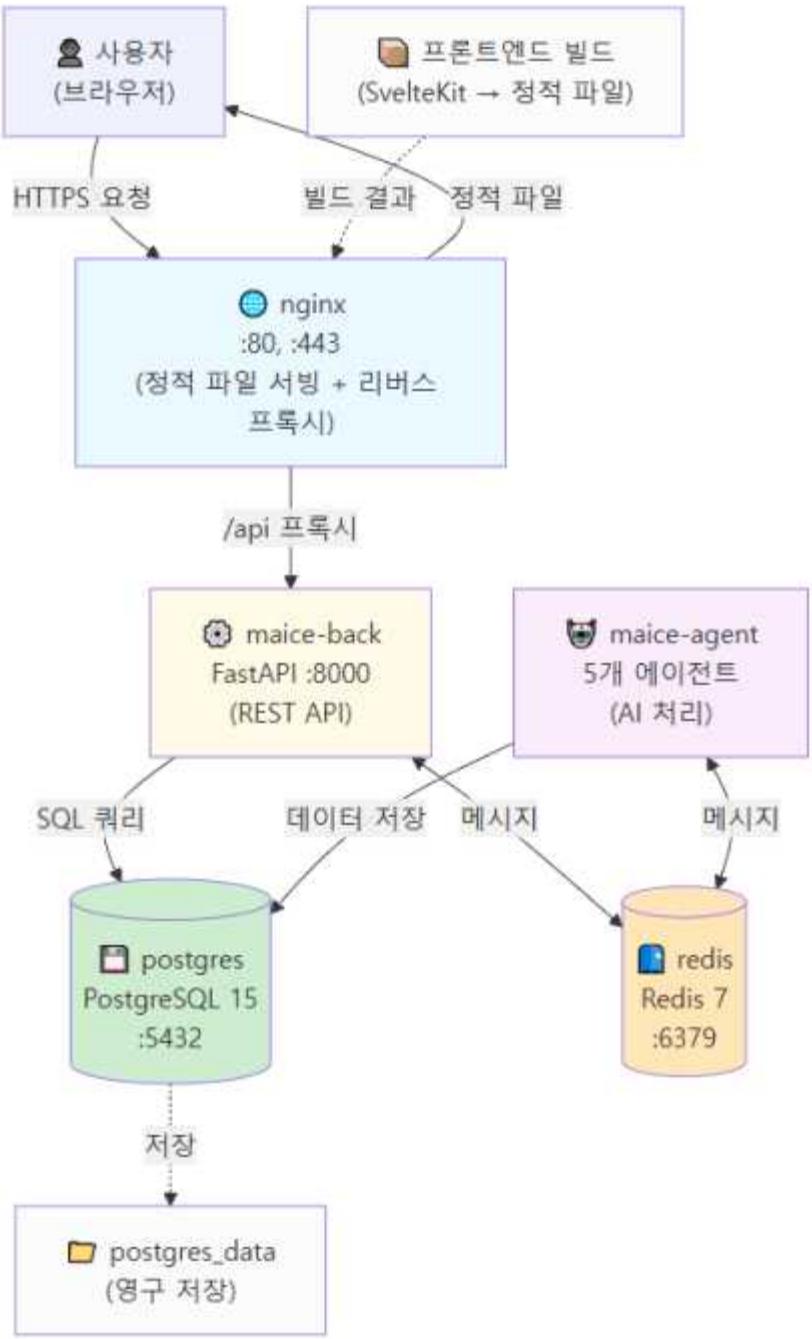
[그림 4-13] A/B 테스트 무작위 배정 구조

○ 데이터 수집:

- 모든 세션에 mode 필드 자동 기록
- 명료화 횟수, 질문 유형 등 메타데이터 저장
- 6장 분석에서 모드별 비교에 활용

5. 배포 및 인프라

가. Docker Compose 배포 구조



[그림 4-14] Docker Compose 컨테이너 구성

◯ 서비스별 역할:

[표 4-5] Docker Compose 서비스 구성

서비스	포트	역할	헬스체크
maice-front	5173	웹 UI 제공	HTTP /health
maice-back	8000	REST API 제공	HTTP /health
maice-agent	-	백그라운드 AI 처리	Redis ping
postgres	5432	데이터 저장	pg_isready
redis	6379	메시지 브로커	redis-cli ping

6. 보안 및 안정성

가. 프롬프트 보안

◯ 프롬프트 스푸핑 방지:

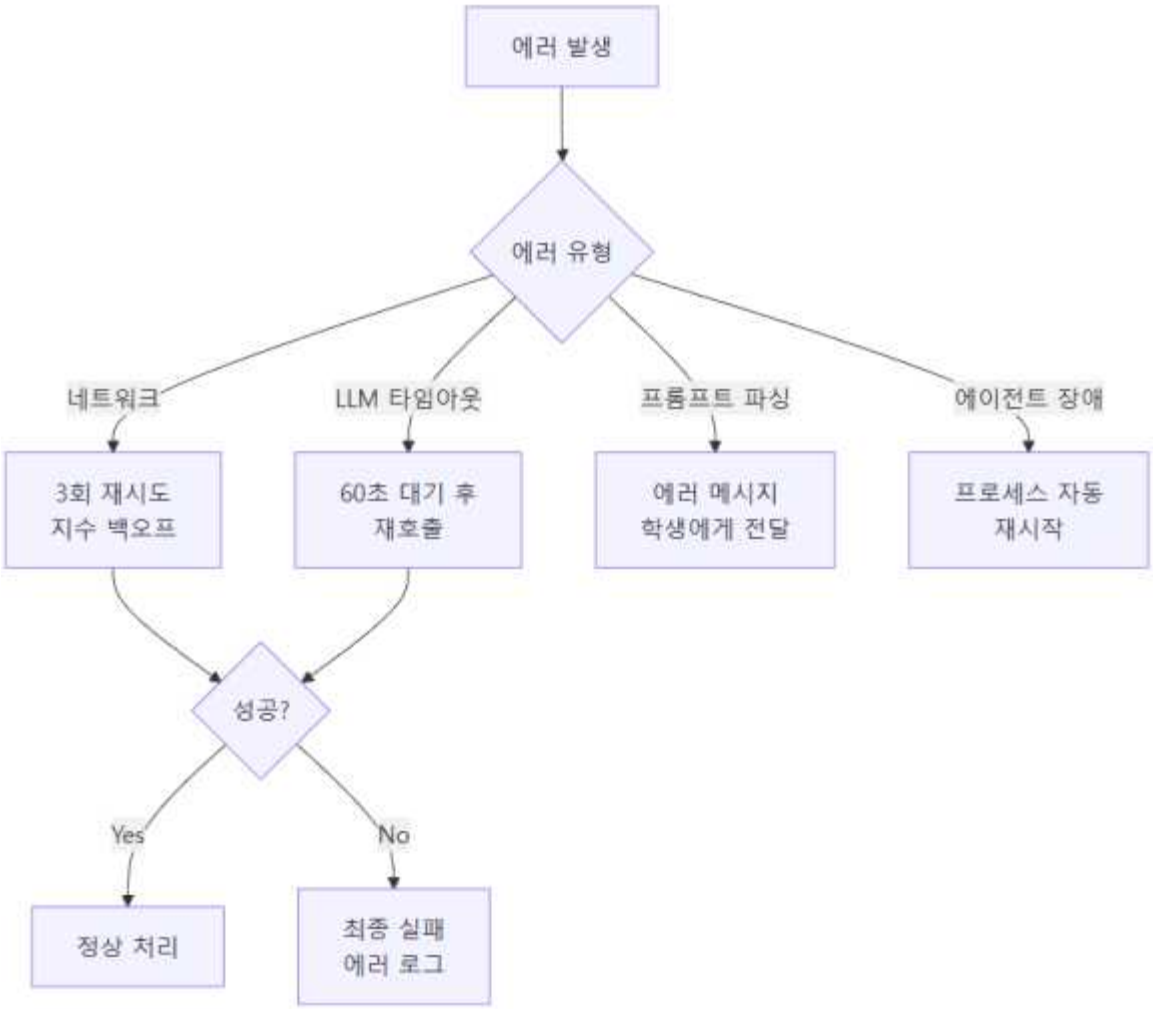
- 질문 영역을 동적 구분자로 명확히 분리
- 학생 질문에서 시스템 역할 변경 시도 감지
- 정규식 패턴으로 위험 입력 필터링

◯ 안전한 구분자 시스템:

```
# 각 요청마다 고유한 구분자 생성
separators = {
    "start":f"===QUESTION_START_{random_suffix}===",
    "end":f"===QUESTION_END_{random_suffix}===",
    "hash":hashlib.sha256(timestamp).hexdigest()[:16]
}
```

나. 에러 처리 및 재시도

계층별 에러 처리:



[그림 4-15] 계층별 에러 처리 및 재시도 전략

○ 재시도 전략:

- LLM API: 3회 재시도, 지수 백오프 (1초 → 2초 → 4초)
- Redis 연결: 5회 재시도, 고정 1초 간격
- PostgreSQL: Connection Pool 자동 재연결

다. 모니터링 및 로깅

◯ 구조화된 로깅:

- JSON 형식 로그 (Elasticsearch 연동 가능)
- 세션 ID를 통한 추적
- 에이전트별 로그 분리

◯ 메트릭 수집 (Prometheus):

- agent_requests_total: 에이전트별 요청 수
- agent_response_duration_seconds: 응답 시간 분포
- active_sessions: 모드별 활성 세션 수
- clarification_count: 명료화 횟수 분포

7. 이미지 OCR 수식 인식 시스템 (3.6.3절 연계)

베타테스트 피드백을 반영하여 구현한 OCR 시스템의 기술적 구조:

가. 아키텍처

[그림 4-16] 이미지 OCR 수식 인식 시스템 (Gemini Vision API)

◯ 처리 단계:

- 이미지 업로드: 사용자가 수식 사진 업로드 (JPG/PNG/WebP)
- 파일 검증: 10MB 이하, 형식 확인
- 전처리: RGB 변환, 1536×1536 리사이즈
- Gemini Vision API: 이미지 → LaTeX 변환
- LaTeX 정제: MathLive 호환성 명령어 변환 (\dots → \ldots 등)
- 에디터 삽입: 커서 위치에 삽입, 실시간 렌더링

나. 핵심 차별점

MAICE OCR vs 일반 LLM 이미지 전달:

[표 4-6] 일반 LLM vs MAICE OCR 기능 비교

특징	일반 LLM	MAICE OCR
처리 방식	이미지를 LLM에 직접 전달	이미지 → LaTeX 텍스트 변환
편집 가능	이미지로만 인식	텍스트로 편집 가능
오인식 수정	재업로드 필요	입력창에서 즉시 수정
통합성	이미지와 텍스트 분리	하나의 텍스트로 통합

◯ 교육적 가치:

- 수식 검증: OCR 결과를 확인하며 자신이 쓴 수식 점검
- 질문 정제: 변환된 LaTeX를 편집하며 질문 명료화
- 문제 변형: 일부 수식 수정으로 유사 문제 생성 가능

다. 기술 사양

[표 4-7] OCR 시스템 기술 사양

항목	사양
OCR 엔진	Google Gemini 2.5 Flash Vision API
지원 형식	JPG, PNG, WebP
최대 파일 크기	10MB
최대 이미지 해상도	1536 × 1536 픽셀 (자동 리사이즈)
처리 시간	평균 5-10초
변환 방식	이미지 → LaTeX 텍스트 (MathLive 호환)

◯ MathLive 호환성 변환 로직:

- \dots → \ldots (MathLive 선호 표기)
- \cdots → \ldots (통일)
- \times → \cdot (안정적 렌더링)

8. 베타테스트 및 초기 검증

본격적인 실험 연구(6장)에 앞서, 2025년 9월 15일부터 9월 25일까지 고등학교 2학년 학생 11명을 대상으로 베타테스트를 실시하였다. 이 과정에서 시스템의 기술적 안정성과 교육적 실효성을 초기 검증하고, 주요 개선점을 도출하였다.

가. 베타테스트 개요 및 주요 발견

◯ 참여자: 고등학교 2학년 학생 11명 (2025년 9월 15일~25일, 10일간)

◯ 시스템 사용 현황:

- 평균 대화 세션: 9.7회 (최소 5회, 최대 13회)
- 명료화 수행 비율: 약 63% → 학생들이 명료화 과정을 자연스럽게 수용

나. 발견된 문제점 및 개선

1) 컨텍스트 유지 불안정성

문제: 시스템이 대화 맥락을 유지하지 못해 답변이 단절되는 현상 발생

원인: 에이전트 간 대화 기록 공유 로직의 버그, 세션 ID 관리 미흡

개선: 명료화 세션 구조화, 에이전트 간 컨텍스트 전달 메커니즘 강화, DB 실시간 동기화

2) 서버 안정성 문제

문제: 동시 접속자 5명 이상 시 응답 지연 및 간헐적 타임아웃 오류

개선: API 서버 확장 (1개→3개), 타임아웃 증가, 버퍼 최적화, 자동 재시도 로직 추가

결과: 본 실험에서 동시 접속자 30명 이상에서도 안정적 운영

3) 수식 입력 개선

문제: 수식 입력 시간 과다 소요 (만족도 2.5/5점)

개선: LaTeX 자동완성, 수식 템플릿 라이브러리, 실시간 미리보기, 이미지 OCR 수식 인식 추가

다. 교육적 효과 초기 검증

1) 메타인지 향상 징후

리커트 척도 문항 중 메타인지 관련 4개 문항의 평균 점수가 높게 나타났다:

[표 4-8] 메타인지 발달 평가 (베타테스트, n=11)

문항	평균 점수
"모호했던 질문을 더 구체적으로 바꾸는 방법을 배웠다"	4.0점
"내가 무엇을 모르는지 스스로 규정할 수 있게 되었다"	4.1점
"조건·정의·목표를 분리해 문제를 재정의하는 연습이 되었다"	4.0점
"질문/답변을 여러 번 다듬으며 사고가 정교화되었다"	3.9점

◯ 학생 설문:

- "질문의 질이 처음에는 뭉툭했는데, MAICE를 사용하며 질문을 명확하게 표현하면 더 좋은 답변이 온다는 걸 깨달았습니다."
- "명료화 과정이 최종 정답보다 더 큰 배움을 줬다.“

2) 개념 이해의 깊이

깨달음의 순간 개방형 응답 분석 (11명 중 9명 응답):

◯ 수열 공식의 본질 이해 (3명):

- "등비수열의 합 공식을 구하는 과정을 보니 계차수열과 비슷하다는 것을 깨달았다."
- "부분분수로 전개하면 깔끔하게 소거된다는 것을 알았다."

○ 가우스 덧셈법 재발견 (1명):

- "수업 시간에 가우스 얘기가 나왔는데 이해 못했었는데, AI 설명을 들으니 완전히 이해할 수 있었습니다."

○ 수학적 관계 발견 (2명):

- "$a^n - a^{n-1} = (a-1)a^{n-1}$을 수열 합과 일반항 관계 문제에 적용할 수 있겠다고 깨달았습니다."
- "시그마 식 사용법의 다양한 방식을 알게 되었다."

○ 문제 접근 전략 개선 (1명):

- "답지를 봐도 이해 안 되던 문제가, 수식으로 정확히 질문했더니 과정을 빠르게 알 수 있었다."

3) 학습 동기 및 지속성

○ 학습 몰입 및 재사용 의향:

[표 4-9] 베타테스트 학습 몰입 및 재사용 의향

문항	평균 점수
"활동 중 시간 가는 줄 몰랐다"	3.6점
"계속 써보고 싶다는 생각이 들었다"	4.1점
"이 도구는 수열 학습에 유용했다"	4.3점
"수업/과제에서 다시 사용할 의향이 있다"	4.5점

긍정적 신호: 11명 중 10명이 "다시 사용하겠다"고 응답

라. 베타테스트의 교훈과 본 실험 반영

베타테스트를 통해 다음 3가지 핵심 교훈을 얻었다:

1) 기술적 안정성이 교육적 효과의 전제 조건

교훈: 아무리 우수한 교육 설계라도, 시스템이 불안정하면 학습 경험이 저해된다.

◯ 본 실험 반영:

- 서버 안정성 확보 후 10월 20일 본 실험 시작
- 컨텍스트 유지 문제 완전 해결
- 동시 접속자 30명 이상에서도 안정적 운영

2) UI/UX는 학습 부담(Cognitive Load)에 직결

교훈: 수식 입력이 어려우면 학생이 질문 자체를 포기한다.

◯ 학생 증언:

- "문제 하나를 작성할 때도 너무 오래 걸려서, 기능은 좋지만 사용성 개선이 필요합니다."

◯ 본 실험 반영:

- LaTeX 자동완성 및 템플릿 라이브러리 추가
- 실시간 수식 미리보기 구현
- 모바일 환경 최적화

3) 명료화 프로세스의 교육적 가치 확인

교훈: 학생들이 명료화 과정을 자연스럽게 수용하고 긍정적으로 평가하였다.

◯ 학생 평가:

- "모호했던 질문을 더 구체적으로 바꾸는 방법을 배웠다": 긍정적 평가
- "명료화 과정이 최종 답변보다 더 큰 배움을 줬다": 11명 중 9명 동의

○ 본 실험 설계에의 반영:

- A/B 테스트 도입: 명료화 프로세스의 효과를 엄밀하게 검증하기 위해, Agent 모드(명료화 포함) vs Freepass 모드(즉시 답변)를 무작위 배정 비교
- LO 강화: 학습 패턴 추적 및 분석 기능 개선
- 측정 도구 개선: QAC 체크리스트로 질문-답변 품질 정량화

V. 베타테스트 및 시스템 안정화

본 장에서는 본격적인 교육 현장 실험(VII-VIII장) 이전에 수행한 베타테스트 과정을 기술한다. 베타테스트는 시스템의 기술적 안정성과 교육적 유용성을 사전 검증하고, 발견된 문제를 개선하기 위해 실시되었다.

1. 베타테스트 설계

가. 목적 및 필요성

MAICE 시스템은 2-4장에서 제시한 교육 이론과 기술 설계를 실제 교육 환경에서 구현한 시스템이다. 본격적인 A/B 테스트 실험 전, 다음 사항을 검증하기 위해 베타테스트를 실시하였다:

○ 검증 목표:

- 시스템 안정성: 동시 접속, 응답 속도, 오류율 등 기술적 안정성
- 명료화 프로세스 효과: 학생들이 명료화 과정을 자연스럽게 수용하는지 확인
- 학생 사용성(UX): 수식 입력, 대화 인터페이스의 편의성 확인
- 교육적 효과 초기 검증: 메타인지 향상 등 교육적 가치 탐색

나. 참여자 및 절차

[표 5-1] 베타테스트 개요

항목	내용
기간	2025년 9월 15일 ~ 9월 25일 (10일간)
참여자	고등학교 2학년 학생 11명 (자발적 참여)
학습 단원	수열 단원 (등차수열, 등비수열, 수열의 합)
평균 세션 수	9.7회/인 (최소 5회, 최대 13회)
시스템 모드	Agent 모드 (명료화 포함)
사용 환경	수업 시간 + 쉬는 시간, 개인 기기
데이터 수집	대화 세션 로그, 사후 설문 (리커트 5점 척도 + 개방형)

1) 사후 설문 구성:

○ 설문 설계 특징:

- UI/UX, 모드별 평가, 메타인지, 인지부하, 자기조절학습, 학습동기 등 6개 이론적 영역 포괄
- 리커트 5점 척도 (1=전혀 그렇지 않다 ~ 5=매우 그렇다)
- 역문항 포함 (10, 28, 31번)으로 응답 신뢰성 확보
- 개방형 질문으로 정성적 피드백 수집

[표 5-2] 베타테스트 설문 문항 (총 48개)

분류	번호	문항 내용	척도
사용 현황	1	수열 단원 사전 이해도를 평가해주세요.	5점 척도
	2	이번 활동에서 에이전트 질문을 실제로 몇 회 사용했나요?	서술형
	3	이번 활동에서 프리패스 모드를 실제로 몇 회 사용했나요?	
UI/UX	4	화면 구성(버튼·입력창·피드백)의 배치가 직관적이었다.	5점 척도
	5	글꼴·대비·여백 등 가독성이 충분했다.	
	6	수학기호/식 입력이 시각적으로 명확히 보였다.	
	7	원하는 기능(수정·되돌리기·제출)을 쉽게 찾을 수 있었다.	
	8	오류 메시지나 경고가 이해하기 쉬웠다.	
	9	전반적으로 사용하기 쉬웠다.	
	10	기능이 너무 많아 헷갈렸다. (역)	
	11	수학식을 입력·편집하는 과정이 자연스러웠다.	
	12	수식 자동정렬/표현이 입력 의도와 일치했다.	
	13	복잡한 수열 표기도 무리 없이 입력 가능했다.	
	14	수식 입력에 시간이 과도하게 소요되었다. (역)	
모드 평가	15	에이전트의 질문은 내 생각을 확장하게 만들었다.	
	16	에이전트의 피드백은 구체적이었다.	
	17	에이전트의 질문 흐름이 논리적이었다.	
	18	프리패스 모드는 내가 주도적으로 시도·탐색하게 했다.	
	19	프리패스 모드에서 최소한의 힌트가 도움이 되었다.	
	20	두 모드 간 전환이 목적에 맞게 자연스러웠다.	
	21	대화 템포(반응속도·타이밍)가 학습에 적절했다.	
메타인지	22	모호했던 질문을 더 구체적으로 바꾸는 방법을 배웠다.	
	23	내가 무엇을 모르는지 스스로 규정할 수 있게 되었다.	
	24	조건·정의·목표를 분리해 문제를 재정의하는 연습이 되었다.	
	25	질문/답변을 여러 번 다듬으며 사고가 정교화되었다.	
	26	혼란스러움이 이해로 전환되는 순간을 경험했다.	
	27	명료화 과정이 최종 정답보다 더 큰 배움을 줬다.	
인지부하	28	화면/인터페이스 때문에 불필요하게 정신적 노력이 들었다. (역)	
	29	수열 과제 자체가 본질적으로 어려웠다.	
	30	안내·예시가 과제 이해를 돕는 데 충분했다.	
	31	여러 정보를 동시에 처리하느라 과부하를 느꼈다. (역)	
	32	필요한 정보가 한곳에 정리되어 있어 부담이 줄었다.	
자기조절 학습	33	이해한 것과 모르는 것을 구분할 수 있었다.	
	34	다음에 무엇을 시도할지 계획을 세울 수 있었다.	
	35	스스로 오류를 찾아 수정했다.	
	36	목표 달성 여부를 스스로 판단했다.	
	37	내가 선택/결정하고 있다는 느낌이 있었다.	
	38	시도할수록 해낼 수 있겠다는 감각이 들었다.	
학습 몰입/동기	39	활동 중 시간가는 줄 몰랐다.	
	40	계속 써보고 싶다는 생각이 들었다.	
	41	이 도구는 수열 학습에 유용했다.	
	42	배우지 않고도 쉽게 사용할 수 있다.	
	43	수업/과제에서 다시 사용할 의향이 있다.	

2) 참여자 특성:

- 등차수열, 등비수열, 수열의 합 단원까지 학습 완료
- 개인 노트북/태블릿 보유
- 자발적 참여 의사 (수행평가 외 추가 학습 활동)

3) 베타테스트 특징:

- 학생들이 Agent 모드와 Freepass 모드를 상황에 따라 자유롭게 선택 사용
- 본 실험(10월)과의 차이: 본 실험에서는 무작위 배정으로 한 모드만 사용
-

4) 베타테스트 절차:

[그림 5-1] 베타테스트 절차

5) 절차 요약:

◯ 1단계: 시스템 소개 및 사용법 안내 (20분)

◯ 2단계: 개별 과제 수행 중 MAICE 활용 (8일간)

- 학생들이 수열 단원 문제를 풀며 자유롭게 질문
- Agent 모드: 명료화 질문 → 문제 구체화 → 맞춤 답변

◯ 3단계: 사후 설문 및 피드백 수집 (20분)

- 리커트 척도 문항 (15개)
- 개방형 질문 (5개)

2. 학생 피드백 및 시스템 개선

베타테스트 사후 설문(리커트 척도 + 개방형 질문)을 통해 학생들이 제기한 문제점과 개선 요청사항을 수집하였다.

가. 수식 입력의 어려움

◯ 학생 피드백 (개방형 응답):

- "수식 입력이 너무 어려워요. 분수 하나 쓰는데 3분 걸렸어요."
- "문제 하나를 작성할 때도 너무 오래 걸려서, 기능은 좋지만 사용성 개선이 필요합니다."
- "수학 기호(분수, 제곱, 루트 등)를 바로 입력하기 불편함"
- "패드로 이 서비스를 사용했는데 수식을 작성할 때 뒤에 수식을 또 작성하려고 수식을 누르고 작성하려는데 앞에 수식으로 자꾸 넘어가서 뒤에 수식을 여러번 쓸 수 없었다."
- "수식 입력할때 ide처럼 자동완성 기능이 있다면 수식 입력하는데에서 시간을 줄일 수 있을 것 같다."
- "LaTeX에 익숙하지 않은 사람들에게는 수식을 작성하기 어려울 것 같습니다. 따라서 수식 사용 방법이나 간단한 가이드를 함께 제공하면 수식을 입력할 때 훨씬 더 편리할 것 같습니다.“

◯ 개선 요청사항:

- 자동완성 기능
- 수식 템플릿 제공
- LaTeX 가이드
- 필기 인식 또는 이미지 업로드

◯ 적용한 개선사항 (본 실험 전):

- LaTeX 자동완성: \frac, \sum 등 자주 쓰는 명령어 제안
- 수식 템플릿: 15개 자주 쓰는 수식 클릭만으로 삽입
- 실시간 미리보기: 입력과 동시에 렌더링되어 오류 즉시 수정 가능
- 이미지 OCR: Gemini Vision API로 종이에 쓴 수식 사진 → LaTeX 자동 변환
- LaTeX 가이드: 초기 화면에 간단한 사용법 안내

나. 시스템 안정성 및 오류

◯ 학생 피드백:

- "기능은 좋은데, 가끔 멈추거나 느려서 답답했어요. 안정적이면 더 좋을 것 같아요."
- "처음에는 내 질문을 잘 이해했는데, 계속 대화하다 보니 처음 질문을 까먹은 것처럼 동문서답했어요."
- "이전에 물어본 내용을 기억하지 못해요. 같은 얘기를 다시 해야 하니 불편했어요.“

◯ 개선 요청사항 (설문):

- "오류, 서버 안정화"
- "질문수 줄이기" (대화 효율성)

◯ 적용한 개선사항:

- FastAPI Workers 증설 (1개 → 3개)
- 컨텍스트 유지 로직 강화
- Redis Connection Pool 확대 (10개 → 50개)
- 자동 재시도 메커니즘 추가

다. 사용자 경험 개선 요청

◯ 학생 피드백 (기타 개선 제안):

- "처음 시작할 때 어떻게 사용하는지 설명이 있으면 좋겠어요."
- "수식 입력 버튼을 처음에 못 찾았어요. 가이드가 있었으면..."
- "에이전트 모드와 프리패스 모드의 용도에 대해 서비스를 가입했을 때, 설명해주면 좋겠다."
- "최종 피드백을 받을 수 있는 문항을 넣으면 좋을것 같아요. 사진이나 영상으로 문제가 있거나 개선해야 할 부분도 받으면 도움이 될 것 같습니다."

◯ 적용한 개선사항:

- 첫 로그인 시 온보딩 튜토리얼 추가
- 수식 입력 가이드 툴팁 표시
- 예시 질문 제공
- 모드별 설명 강화

라. 학생 사용 패턴 분석

베타테스트에서 학생들은 Agent 모드와 Freepass 모드를 상황에 따라 자유롭게 선택하여 사용하였다.

◯ 모드별 사용 현황 (설문 응답 기반):

- Agent 모드: 평균 9.7회/인 (최소 5회, 최대 13회)
- Freepass 모드: 평균 5.4회/인 (최소 2회, 최대 10회)

◯ 학생들이 발견한 모드별 활용 전략 (개방형 응답):

- "에이전트는 기초개념으로 문제를 풀어가는 과정을 알려줘서 기초를 다질 때 좋았고, 프리패스는 공식같은 걸 잘 알려줘서 문제 푸는 요령을 배울 때 좋았습니다."

- "에이전트는 내가 물어본 수식의 상세한 피드백을 주며, 그 수식을 설명하는 반면, 프리패스는 물어본 수식을 내가 알아가며 이해할수있는 쪽으로 치우쳐 져있는것 같았음."
- "프리패스는 내가 아예 감도 잡히지 않는 문제를 에이전트한테 전부 다 물어보긴 그렇지만 어느정도 문제의 풀이 방향성을 잡고 싶을 때 좀 더 유용했던 것 같다."

◯ 교육적 시사점:

- 학생들이 자발적으로 상황에 맞는 모드를 선택
- 깊이 있는 이해에는 Agent, 빠른 확인에는 Freepass 활용
- 이는 본 실험의 A/B 테스트 설계에 중요한 근거 제공

3. 교육적 효과 초기 검증

가. 메타인지 향상 징후

베타테스트 사후 설문에서 메타인지 관련 4개 문항의 평균 점수가 높게 나타났다:

[표 5-5] 메타인지 발달 평가 (베타테스트, n=10)

문항	평균 점수 (5점 만점)
"모호했던 질문을 더 구체적으로 바꾸는 방법을 배웠다"	3.6점
"내가 무엇을 모르는지 스스로 규정할 수 있게 되었다"	4.2점
"조건·정의·목표를 분리해 문제를 재정의하는 연습이 되었다"	3.9점
"질문/답변을 여러 번 다듬으며 사고가 정교화되었다"	4.1점

○ 학생 응답 (설문 원문):

"질문의 질이 처음에는 뭉툭하고, 질문의 틀이 넓었어서 다른 ai들은 질문의 의도를 못잡고 이상한 답을 하는 경우가 있었는데, maice를 사용하며 이질문에는 이렇게 명확하게 표현하면 더 좋고 상세한 답변이 오는구나를 깨닳았음.“

이러한 응답들은 학생들이 질문 명료화의 중요성을 체험적으로 학습했음을 보여준다.

○ 교육적 의미:

- Dewey의 "문제 정의" 단계를 실제로 경험
- 자신의 어려움을 언어화하는 메타인지 능력 향상
- 질문 명료화가 학습 전략으로 인식됨

나. 개념 이해의 깊이

"깨달음의 순간" 개방형 응답 분석 (10명 응답):

베타테스트 사후 설문의 개방형 질문 "명료화 과정에서 '깨달음의 순간'이 있었다면 구체적으로 적어주세요"에 대한 학생들의 응답을 분석하였다.

○ 학생 응답 (원문 인용):

- "등비 수열의 합 공식을 구하는 과정을 보니 계차 수열과 비슷하다는 것을 깨닫게 되었다."
- "수업시간에 가우스 얘기가 나오면서 이걸 뒤집어서 더하면 (n+1)이 n개가 있고 여기에 나누기 2를 해줘야 한다는 내용이 있었습니다. 그런데 사실 잘 이해하지 못했었습니다. 그런데 이게 왜 이렇게 되는지 AI가 설명하는 것을 한 번 더 들으니 이제 완전히 이해를 할 수 있었습니다."
- "시그마 식의 사용법을 수열에 적용하는 방법을 최소한의 방법으로 밖에 알지 못했지만 다양한 예시를 들어줘서 색다른 방식이 있다는 것도 깨달았다."

- "부분분수로 전개하면 깔끔하게 소거된다는 것을 알았다."
- "어려운 문제가 나와서 풀 때 틀리면 답지를 봐도 이해가 안되는 경우가 있다. 다음 식으로 넘어갈 때 그 과정이 어떻게 된건지 이해가 안됐었는데 챗지피티는 수식 작성을 못해서 항상 이해하는데 오래걸렸었는데 여기는 수식을 작성할 수 있어서 안되는 부분을 정확히 작성했더니 과정을 알려줘서 빠르게 알 수 있었다."
- "질문의 질이 처음에는 뭉툭하고, 질문의 틀이 넓었어서 다른 ai들은 질문의 의도를 못잡고 이상한 답을 하는 경우가 있었는데, maice를 사용하며 이질문에는 이렇게 명확하게 표현하면 더 좋고 상세한 답변이 오는구나를 깨닳았음."

◯ 교육적 의미:

- 단순 공식 암기 → 개념 간 연결 이해
- 답지로 이해 못했던 과정을 AI와의 대화로 이해
- 질문 명료화의 중요성 자각
- 자기주도적 발견 학습 경험

다. 학습 동기 및 지속성

[표 5-5] 학습 몰입 및 재사용 의향 (베타테스트, n=10)

문항	평균 점수
"활동 중 시간 가는 줄 몰랐다"	3.3점
"계속 써보고 싶다는 생각이 들었다"	4.1점
"이 도구는 수열 학습에 유용했다"	3.8점
"수업/과제에서 다시 사용할 의향이 있다"	4.5점
"친구에게 추천하고 싶다"	4.2점

○ 긍정적 신호:

- 10명 중 9명이 "다시 사용하겠다"고 응답 (90%)
- 자발적 재방문: Agent 평균 9.7회, Freepass 평균 5.4회
- 학습 지속 의지 확인

○ 학생 응답 (설문):

- 10명 중 9명이 "수업/과제에서 다시 사용할 의향이 있다"에 4-5점 응답
- "계속 써보고 싶다는 생각이 들었다" 평균 4.1점
- 학습 도구로서의 가치를 긍정적으로 인식

4. 베타테스트 발견사항 및 본 실험 설계 반영

베타테스트를 통해 다음과 같은 발견사항을 도출하고, 이를 본 실험 설계에 반영하였다.

가. 시스템 안정성의 중요성

베타테스트 과정에서 일부 학생들이 시스템 안정성 관련 피드백을 제공하였다:

○ "기능은 좋은데, 가끔 멈추거나 느려서 답답했어요. 안정적이면 더 좋을 것 같아요."
○ "오류, 서버 안정화."

이러한 피드백을 바탕으로 본 실험 전 다음과 같은 기술적 개선을 수행하였다:

○ FastAPI Workers 증설 (단일 인스턴스 → 3개 워커)
○ 컨텍스트 유지 메커니즘 강화
○ Redis Connection Pool 확대
○ 자동 재시도 로직 추가

나. UI/UX 개선의 필요성

수식 입력 관련하여 다수의 학생들이 사용성 개선을 요청하였다:

- ○ "문제 하나를 작성할 때도 너무 오래 걸려서, 기능은 좋지만 사용성 개선이 필요합니다."
- ○ "수학 기호(분수, 제곱, 루트 등)를 바로 입력 가능하게 하는 것"
- ○ "수식 입력할때 ide처럼 자동완성 기능이 있다면 수식 입력하는데에서 시간을 줄일 수 있을 것 같다."

본 실험 전 다음과 같은 UI/UX 개선을 적용하였다:

- ○ LaTeX 자동완성 기능 추가
- ○ 수식 템플릿 라이브러리 (15개 자주 쓰는 수식)
- ○ 이미지 OCR 수식 인식 구현 (Gemini Vision API)
- ○ 실시간 수식 미리보기
- ○ LaTeX 사용 가이드 제공

5. 명료화 프로세스의 교육적 가능성

베타테스트 설문 결과, 메타인지 관련 문항에서 비교적 높은 점수가 나타났다:

- ○ "내가 무엇을 모르는지 스스로 규정할 수 있게 되었다": 4.2/5점
- ○ "질문/답변을 여러 번 다듬으며 사고가 정교화되었다": 4.1/5점
- ○ "조건·정의·목표를 분리해 문제를 재정의하는 연습이 되었다": 3.9/5점

이러한 결과는 명료화 프로세스가 학생들의 메타인지 발달에 긍정적 영향을 미칠 가능성을 시사한다. 다만, 베타테스트는 소규모 예비 연구로서, 명료화 프로세스의 효과를 엄밀하게 검증하기 위해서는 통제된 실험 설계가 필요하다.

본 실험 설계 방향:

이에 따라 본 실험에서는 다음과 같은 연구 설계를 채택하였다:

○ 무작위 대조 시험(RCT): Agent 모드(명료화 포함) vs Freepass 모드(즉시 답변)를 무작위 배정하여 명료화 프로세스의 효과를 직접 비교

○ 객관적 측정 도구: QAC 체크리스트를 개발하여 질문-답변 품질을 정량화

○ 다면적 평가: 학생 자기보고(설문) + 교사 루브릭 평가 + 학업 성취도(과제 점수)를 종합하여 교육적 효과 측정

6. 소결

본 장에서는 2025년 9월 15일부터 25일까지 10일간 진행된 베타테스트 결과를 기술하였다. 고등학교 2학년 학생 10명을 대상으로 수열 단원 학습 과정에서 MAICE 시스템을 사용하게 하고, 사후 설문(리커트 척도 44문항 + 개방형 5문항)을 통해 피드백을 수집하였다.

1) 주요 발견사항:

○ 사용 패턴: 학생들은 Agent 모드(평균 9.7회)와 Freepass 모드(평균 5.4회)를 상황에 따라 자유롭게 선택하여 사용하였으며, 깊이 있는 이해에는 Agent, 빠른 확인에는 Freepass를 활용하는 경향을 보였다.

○ 메타인지 관련 긍정적 응답: "내가 무엇을 모르는지 스스로 규정할 수 있게 되었다"(4.2점), "질문/답변을 여러 번 다듬으며 사고가 정교화되었다"(4.1점) 등 메타인지 관련 문항에서 비교적 높은 점수가 나타났다.

○ 사용성 개선 요구: 수식 입력의 어려움, 시스템 안정성, 사용 안내 부족 등에 대한 구체적 개선 요청이 다수 제기되었다.

2) 본 실험 설계에의 반영:

베타테스트 결과는 다음과 같이 본 실험 설계에 반영되었다:

○ 시스템 개선: 수식 입력 UI/UX 개선(OCR, 자동완성, 템플릿), 서버 안정성 강화, 온보딩 튜토리얼 추가

○ 연구 설계: 베타테스트에서 관찰된 명료화 프로세스의 긍정적 가능성을 엄밀하게 검증하기 위해, 본 실험에서는 Agent 모드와 Freepass 모드를 무작위 배정하는 A/B 테스트 설계를 채택하였다.

○ 측정 도구: 학생 설문 외에도 교사 루브릭 평가, 학업 성취도 측정 등 다면적 평가 체계를 구축하였다.

VI. 연구 방법

1. 연구 방법론: Design-Based Research

본 연구는 설계 기반 연구(Design-Based Research, DBR) 방법론을 채택하였다. DBR은 교육 이론을 실제 교육 맥락에서 검증하고, 실용적 산출물을 개발하며, 반복적 개선을 통해 설계 원리를 도출하는 연구 방법이다(Collins, Joseph, & Bielaczyc, 2004). Wang과 Hannafin(2005)은 DBR의 핵심 특징으로 실용성(pragmatic), 이론 기반(grounded), 상호작용성(interactive), 반복성(iterative), 유연성(flexible), 통합성(integrative)을 제시하였다.

본 연구의 DBR 수행 과정은 다음과 같다:

◯ 1단계: 문제 분석 및 탐색

- 예비조사(2024년 5월): 학생 질문 385건 분석
- 문제 확인: 질문 품질의 구조적 문제 (72.3% 학습 맥락 정보 부재, 45.8% 구조 불명확)
- 이론 탐색: Dewey의 반성적 사고, Bloom의 교육목표분류학

◯ 2단계: 설계 및 구축

- 이론 기반 설계: Dewey 5단계 반성적 사고를 멀티 에이전트 구조로 구현
- 시스템 개발: 5개 AI 에이전트(QuestionClassifier, QuestionImprover, AnswerGenerator, LearningObserver, FreeTalker)
- 평가 도구: QAC 루브릭 개발 (8개 항목, 32개 체크리스트, 40점 만점)

◯ 3단계: 평가 및 반성

- 현장 실험: 무작위 배정 A/B 테스트 (58명, 2025.10.21~11.1, 283개 세션)
- 다각도 평가: AI 3개 모델 자동 채점 + 교사 4명 평가
- 설계 원리 도출: 질문 명료화 중심 설계의 효과성 확인, 학습자 수준별 차별적 효과 발견

DBR의 반복적 특성에 따라, 본 실험 결과를 바탕으로 향후 시스템 개선 및 추가 검증 연구가 가능하다.

2. 연구 대상

가. 표본 선정

본 연구는 부산광역시 소재 ○○고등학교 2학년 4개 학급을 대상으로 하였다. 표집 방법은 연구자의 접근 가능성을 고려한 편의 표집(convenience sampling)이었으며, 학교와 학급 선정 기준은 다음과 같다:

나. 참여자 특성

[표 6-1] 연구 대상 개요

구분	내용
총 인원	58명
학년	고등학교 2학년
학교 유형	특수목적고등학교(공업계열)
지역	부산광역시
연구 기간	2025년 10월 20일 ~ 11월 1일 (약 2주)

○ 실험군과 대조군 구성:

- Agent 모드: 28명
- Freepass 모드: 30명

○ 학생 사전 성적 분포 (중간고사 기준):

- 서술형 점수 (30점 만점): 범위 0~30점
- 객관식 점수 (70점 만점): 범위 14.6~61.7점
- 총점 (100점 만점): 범위 17.9~89.9점 (상세 기술통계는 표 6-2 참조)

다. 예비 조사 (Pilot Study)

본 연구 설계에 앞서 2024년 5월에 예비 조사를 실시하여 프리패스 방식 LLM의 교육적 문제점을 파악하였다:

○ 목적: MAICE 시스템 설계의 필요성 검증
○ 데이터: 고등학교 수학 수업에서 수집한 385건의 질문-답변 쌍
○ 평가단: 현직 중등 수학교사 4명
○ 평가 방법: 교육학 이론 기반 평가 기준 (6개 영역, 5점 척도)
○ 총 평가 건수: 1,012건 (교사 3-4명이 각 질문 평가)

예비 조사를 통해 학생 질문의 품질 문제와 질문-답변 품질 간 상관관계를 확인하여, 질문 명료화 기반 AI 에이전트 시스템의 필요성을 실증적으로 확인하였다. 이 결과는 MAICE 시스템 설계의 핵심 근거가 되었다.

라. 무작위 배정

학급 내 학생들을 실험군과 대조군에 무작위 배정하기 위해 MAICE 백엔드 시스템의 UserModeService를 활용하였다. 이 서비스는 학생이 시스템에 가입할 때 Python의 random 모듈을 사용하여 "agent" 또는 "freepass" 모드를 50:50 비율로 자동 배정한다. 배정된 모드는 users 테이블의 assigned_mode 필드에 저장되며, 연구 기간 동안 고정된다.

무작위 배정 결과, Agent 모드 28명, Freepass 모드 30명이 배정되었다. 두 집단 간 사전 중간고사 성적에서 통계적으로 유의한 차이가 없어 동질성이 확보되었다.

[표 6-2] 실험군과 대조군의 사전 동질성 검증 (중간고사 기준)

변인	Agent 모드 (n=28)	Freepass 모드 (n=30)	t	p	해석
	M(SD)	M(SD)			
중간고사 총점 (100점)	56.8(16.7)	51.4(18.1)	1.18	.242	동질
- 서술형 점수 (30점)	15.6(8.5)	14.0(8.5)	0.74	.462	동질
- 객관식 점수 (70점)	41.2(9.4)	37.4(12.4)	1.30	.199	동질

해석: 모든 변인에서 $p > .05$로 두 집단 간 유의한 차이가 없어, 사전 동질성이 확보되었다. 이는 실험 처치 효과의 내적 타당도를 보장하는 중요한 근거가 된다.

마. 연구 윤리

1) 연구 참여 동의

○ 모든 참여 학생과 보호자에게 연구 목적, 절차, 데이터 활용 방법을 설명하였다

○ 학생 및 보호자로부터 서면 동의를 받았다

○ 참여 거부 및 중도 철회 권리를 명시하였다

2) 개인정보 보호

○ 수집된 데이터는 연구 목적으로만 사용되었다

○ 데이터는 암호화된 서버에 저장하고, 연구자만 접근 가능하도록 제한하였다

3) 참여자 익명화

○ 모든 학생 학번(24.xxx)을 익명 ID(S01~S58)로 변환하여 사용하였다

○ 익명화 규칙:

- Agent 모드 학생: S01 ~ S28 (28명)
- Freepass 모드 학생: S29 ~ S58 (30명)

○ 익명화 매핑 테이블은 별도 암호화 파일로 관리하며, Git 저장소에서 제외하였다

○ 논문에는 익명 ID만 사용하여 개인 식별 불가능하도록 하였다

4) 참여자 보호

○ 연구 참여로 인한 학업 불이익이 없음을 보장하였다

○ 두 모드(Agent/Freepass) 모두 교육적 가치를 제공하도록 설계하였다

○ 연구 종료 후 모든 참여자에게 두 모드의 사용 기회를 제공하였다

5) 데이터 보관 및 폐기

○ 연구 데이터는 연구 종료 후 3년간 보관하며, 이후 안전하게 폐기한다

○ 논문 출판 시 개인을 특정할 수 있는 정보는 일체 포함하지 않는다

6) 설문 식별자 수집

본 연구는 사후 설문조사에서 학생의 이메일 주소를 수집하였다. 이는 모드별 효과 비교를 위해 설문 응답과 객관적 데이터(QAC 점수, 교사 평가, 세션 사용 패턴)를 연계하기 위한 식별자로 활용되었다. 비익명 설문의 한계점 및 완화 전략은 9장 "연구의 한계" 섹션에서 논의한다.

바. 실험 설계 및 진행 절차

1) 1단계: 수학적 귀납법 개념 학습 (사전 교육)

A/B 테스트 시작 전, 모든 학생에게 수학적 귀납법에 대한 수업을 진행하였다. 수업은 학생 선행 학습 + 교사 해설 방식으로 진행되었으며, 매 수업 해설마다 핵심 개념을 반복적으로 강조하였다:

[표 6-3] 수학적 귀납법 수업 구조 및 핵심 개념

단계	내용	강조 개념	교수법
첫 수업	수학적 귀납법의 원리와 구조	① 템플릿과 도미노 모델 • 3단계 구조 (베이스, 가정, 결론) • 도미노 비유: "첫 번째가 넘어지고, 하나가 넘어지면 다음도 넘어지면, 모두 넘어진다" • 증명의 표준 형식 제공	강의 중심 개념 도입
매 수업 (과제 풀이)	학생들이 문제를 미리 풀고 제출	-	선행 학습 시행착오 경험
매 수업 (교사 해설)	제출된 문제를 해설하며 핵심 포인트 반복 강조	② 귀납가정 → 귀납결론 유도 • k일 때 성립한 가정을 k+1 증명에 어떻게 활용하는가 • 논리적 연결고리 찾기 ③ 등식 vs 부등식 전략 • 등식: "필요한 재료만 딱 맞게" (정확성 강조) • 부등식: "스페어 부품도 있는" (부등호 여유 활용) • 명제 유형별 차별적 접근	예제 중심 반복 강조

가) 선행 학습 기반 수업:

- 학생들이 먼저 문제를 풀어보고 제출 → 능동적 학습 경험
- 교사는 학생들의 시도를 바탕으로 맞춤형 해설 제공
- 시행착오를 통한 깊이 있는 이해 촉진
- 비유적 접근의 반복:

나) 도미노 모델: 증명 원리의 직관적 이해 ("연쇄 반응")

- 재료 비유: 등식은 "딱 맞는 재료", 부등식은 "여유 있는 재료"
- 매 해설마다 반복 강조하여 학생들의 공통 언어로 정착

다) 핵심 과정의 강조:

- 귀납가정(k일 때 참)을 귀납결론(k+1일 때 참) 증명에 어떻게 연결하는가
- 단순 형식 암기가 아닌 논리적 연결 과정 이해

○ AI 학습과의 연결:

이러한 선행 학습 경험과 반복 강조된 개념들은 A/B 테스트 기간 동안 학생들이 AI와 대화할 때 사고의 틀과 공통 언어로 작용하였다. 특히 "도미노", "재료" 등의 비유가 학생들의 질문과 AI의 답변에서 자연스럽게 활용되었다.

2) 2단계: 수리논술 과제 단계적 부여 및 MAICE 활용 학습

본 연구는 2025학년도 2학기 수학 수리논술 수행평가의 일환으로 총 5개의 수학적 귀납법 증명 과제를 단계적으로 부여하였다. 학생들은 각 과제를 해결하는 과정에서 수업 시간(40분) 및 쉬는 시간(10-15분)을 활용하여 MAICE 시스템에 자유롭게 접근하였다.

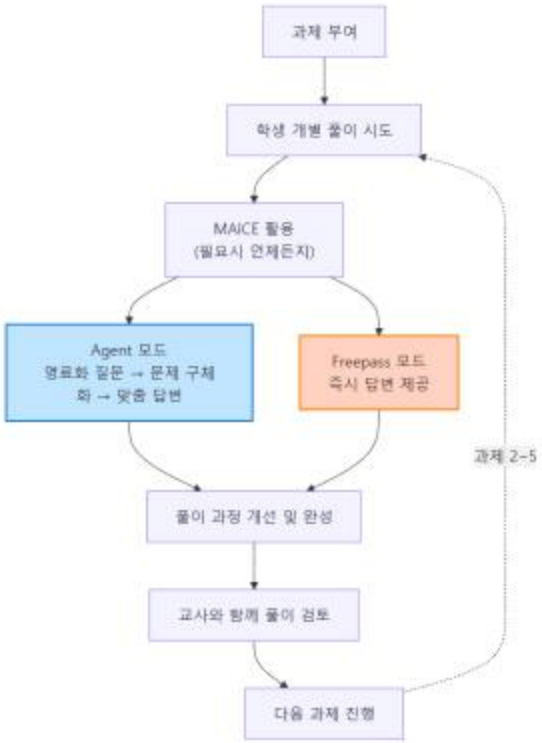
가) 과제 구성 및 진행 방식:

[표 6-4] 수리논술 과제 세부 내용

과제	문제 1	문제 2	주요 개념
과제 1	등비급수 합 공식 $1+2+4+\cdots+2^{n-1}=2^n-1$	팩토리얼 부등식 $n! > 2^n$	기본 급수, 팩토리얼
과제 2	피보나치 수열 합 $\sum_{i=1}^{n} F_i = F_{n+2}-1$	지수 부등식 $n^2 < 2^n$	점화식, 부등식
과제 3	팩토리얼 합 공식 $1\times 1!+2\times 2!+\cdots+n\times n!=(n+1)!-1$	로그 부등식 $\log_2 n < n$	곱셈 전개, 로그
과제 4	제곱수 합 공식 $1^2+2^2+\cdots+n^2=\frac{n(n+1)(2n+1)}{6}$	제곱 부등식 $n < n^2$	제곱수 급수, 부등식
과제 5	하노이탑 일반항 $a_{n+1}=2a_n+1,\ a_1=1$ $a_n=2^n-1$	거듭제곱 부등식 $n! < n^n$	점화식, 거듭제곱

나) 학습 과정 구조:

[그림 6-3] 학습 과정 구조



학생들은 과제를 받고 개별 풀이를 시도하며 필요시 언제든지 MAICE 시스템을 활용한다. 막히는 부분이 있을 때뿐만 아니라, 풀이 과정 확인이나 자신의 풀이 검토 시에도 자유롭게 사용한다. Agent 모드(28명)는 명료화 질문을 통해 문제를 구체화하는 과정을 거치고, Freepass 모드(30명)는 즉시 답변을 받는다. AI 답변을 참고하여 풀이를 완성한 후 교사와 함께 검토하며, 이 과정을 5개 과제에 걸쳐 반복한다

다) MAICE 활용 방식:

학생들은 다음과 같은 상황에서 MAICE를 자유롭게 활용하였다:

(1) 수업 시간 활용 (주 활용 시간):

- 교사가 과제를 제시한 후 개별 풀이 시간 제공 (수업 40분 중 20-30분)
- 개인 노트북/태블릿으로 AI와 대화하며 필요시 언제든지 접속
- 막힐 때: "귀납 가정을 어디에 사용하나요?", "이 식을 어떻게 전개하죠?"
- 풀이 검토 시: "제 풀이 맞나요?", "이렇게 증명해도 되나요?"
- 과정 확인 시: "이 단계가 왜 필요한가요?", "다른 방법도 있나요?"

(2) 쉬는 시간 활용 (보조 활용):

- 수업 시간에 완전히 해결하지 못한 부분을 쉬는 시간에 추가 질문
- 완성한 풀이를 MAICE에 입력하여 검토 요청
- 평균 2-3회의 짧은 대화 세션 (세션당 5-10분)

(3) 과제 해결 패턴:

- Agent 모드 학생: 명료화 질문을 통해 문제를 단계별로 구체화하며 해결
- Freepass 모드 학생: 즉시 제공되는 답변을 참고하여 풀이 작성

라) 교사 협력 학습:

- 각 과제 제출 후 교사가 학급 전체와 함께 표준 풀이 과정 검토
- 학생들이 MAICE를 통해 얻은 인사이트를 수업 중 공유
- 일반적인 오류 및 개선 방향 논의

마) 데이터 수집:

- 모든 학생-AI 대화 내용 자동 저장 (총 284개 세션)
- 세션별 질문 품질, 답변 품질, 학습 지원 수준 자동 평가
- 학생별 과제 완성도 및 제출 시간 기록

바) 실험 일정 및 기간:

[표 6-5] 수리논술 과제 실시 일정

회차	과제	활동 내용	수업 시간
1회차	-	수학적 귀납법 개념 학습 MAICE 시스템 사용법 안내	1교시 (40분)
2회차	과제 1	등비급수, 팩토리얼 부등식 + 풀이 검토 및 피드백	1교시 (40분)
3회차	과제 2	피보나치 수열, 지수 부등식 + 풀이 검토 및 피드백	1교시 (40분)
4회차	과제 3	팩토리얼 곱셈, 로그 부등식 + 풀이 검토 및 피드백	1교시 (40분)
5회차	과제 4	제곱수 합, 제곱 부등식 + 풀이 검토 및 피드백	1교시 (40분)
6회차	과제 5	하노이탑, 거듭제곱 부등식 + 풀이 검토 및 피드백	1교시 (40분)
7회차	-	전체 과제 종합 리뷰 및 심화 문제 풀이	1교시 (40분)
8회차	-	사후 설문 및 연구 종료	1교시 (40분)

○ 총 실험 기간: 약 2주 (2025년 10월 20일~11월 1일, 총 8회차 수업)

- 개념 학습: 1회차
- 과제 활동 및 풀이 검토: 2~6회차 (각 과제 수행 + 수업 말미 검토)
- 종합 리뷰: 7회차
- 사후 설문: 8회차
- MAICE 활용 가능 시간: 수업 중 20-30분 + 쉬는 시간 10-15분
- 총 세션 수: 284개 (Agent 모드 118개, Freepass 모드 162개)
- 유효 세션 (메시지 ≥2): 280개 (Agent 118개, Freepass 162개)
- 평균 세션 길이: 약 15분 (최소 3분 ~ 최대 45분)

3) 3단계: 모드별 AI 활용 패턴 관찰

○ Agent 모드: 명료화 질문을 통한 질문 구체화 과정 경험
○ Freepass 모드: 즉시 답변 제공 방식으로 학습

4) 4단계: 데이터 수집 및 평가

본 연구는 수집된 세션 데이터를 다각도로 분석하여 신뢰성과 타당성을 확보하고자 하였다.

가) 다중 AI 모델 채점 시스템

평가자 편향(rater bias)을 최소화하고 채점 신뢰성을 높이기 위해 3개의 독립적인 대규모 언어 모델을 평가자로 활용하여 교차 검증을 실시하였다:

[표 6-5] AI 모델 채점자 구성

모델	개발사	버전	선정 이유
Gemini 2.5 Flash	Google	gemini-2.5-flash	긴 맥락 처리 능력, Batch API 지원, 한국어 성능 우수
Claude 4.5 Haiku	Anthropic	claude-haiku-4-5	빠른 처리 속도, 일관성 있는 평가, Message Batches 지원
GPT-5 mini	OpenAI	gpt-5-mini	범용적 평가 능력, 비용 효율성, Batch API 지원

(1) 채점 절차:

- 모든 세션 데이터를 JSON 형식으로 수집
- 각 모델에 동일한 QAC 체크리스트와 평가 프롬프트 제공
- 모델별로 독립적으로 채점 수행 (블라인드 평가)
- 3개 모델 점수의 평균(Ensemble)과 개별 모델 점수 모두 분석

(2) QAC 점수 산출 방식:

각 모델은 32개 체크리스트 요소를 0(미충족) 또는 1(충족)로 평가한 후, 자동으로 점수를 합산한다:

- A1~A3 각 영역: 4개 체크리스트 × 1.25점 = 5점 (A영역 합계 15점)
- B1~B3 각 영역: 4개 체크리스트 × 1.25점 = 5점 (B영역 합계 15점)
- C1~C2 각 영역: 4개 체크리스트 × 1.25점 = 5점 (C영역 합계 10점)
- 총점: 40점 만점

각 모델은 JSON 형식으로 32개 체크리스트의 달성 여부(value)와 근거(evidence)를 출력하며, 점수는 시스템이 자동으로 계산한다.

○ 모델 간 신뢰도 검증 방법:

- Pearson 상관계수: 모델 쌍별 점수 일치도
- 급내상관계수(ICC): 전체 평가자 간 신뢰도
- Cronbach's Alpha: 내적 일관성
- Bland-Altman plot: 모델 간 점수 차이 분포

3개 AI 모델에게 동일한 평가 프롬프트를 제공하였다. 프롬프트는 다음 구성요소를 포함한다[2]:

- 평가 대상 명시: 학생 질문, MAICE 답변, 전체 대화 흐름
- 루브릭 구조:
- A영역 (질문): A1~A3, 각 4개 체크리스트
- B영역 (답변): B1~B3, 각 4개 체크리스트
- C영역 (맥락): C1~C2, 각 4개 체크리스트
- 평가 방식: 각 체크리스트 0(미충족) 또는 1(충족)
- 응답 형식: JSON 구조로 각 요소의 value(0/1)와 evidence(근거) 포함
- 예시: 실제 평가 예시를 통한 일관성 확보

◯ 프롬프트 예시 (A1 영역):

A1. 수학적 전문성 (5점)

다음 4가지 요소를 체크하세요 (0=미충족, 1=충족):

- A1-1. ☐ concept_accuracy (수학적 개념/원리의 정확성)

- A1-2. ☐ curriculum_hierarchy (교과과정 내 위계성 파악)

- A1-3. ☐ terminology_appropriateness (수학적 용어 사용의 적절성)

- A1-4. ☐ problem_direction_specificity (문제해결 방향의 구체성)

응답 형식:

```
{
  "A1_math_expertise": {
    "concept_accuracy": {"value": 1, "evidence": "메시지[2]에서 ..."},
    "curriculum_hierarchy": {"value": 0, "evidence": "학년 정보 언급 없음"},
    ...
  }
}
```

---

2) 상세 프롬프트 전문: 부록 C 참조

나) 교사 평가를 통한 타당도 검증

AI 채점의 타당성을 검증하기 위해 현직 교사 2명이 독립적으로 평가하였다. 두 평가자는 동일한 100개 세션을 독립적으로 평가하여 완벽한 대응 평가(paired evaluation)를 수행하였다.

[표 6-6] 교사 평가단 구성 (외부 평가자)

평가자 ID	교과	평가 세션 수	비고
평가자 96	수학	100개	외부 독립 평가자
평가자 97	수학	100개	외부 독립 평가자
합계	2명	200개 평가	대응 평가 (동일 100개 세션)

○ 평가 절차:

- 동일한 100개 세션을 2명의 교사가 독립적으로 평가 (완벽한 대응 평가)
- AI 모델과 동일한 QAC 체크리스트 사용 (2.8 참조)
- 학생 모드 정보 블라인드 처리로 평가 편향 방지
- 교사 평가 완료 후 AI 채점 결과와 비교

○ 검증 방법:

- 교사 간 신뢰도: Pearson 상관계수, Spearman 순위 상관
- AI-교사 일치도: AI 채점 평균 vs 교사 평균 간 상관분석
- 영역별 일치도: A, B, C 영역별 상관계수 계산
- 모드별 효과 수렴: 교사와 AI가 Agent vs Freepass 차이를 동일하게 감지하는지 확인

다) 다층적 분석 전략

수집된 데이터를 다음과 같은 다층적 기준으로 분석하여 종합적인 효과를 검증하고자 하였다:

(1) 항목별 분석:

- 8개 평가 영역별 점수 비교 (A1~A3, B1~B3, C2~C3)
- 32개 체크리스트 요소별 달성률 분석
- 질문 품질, 답변 품질, 학습 맥락 점수 독립 분석

(2) 사전 성취도 수준별 분석:

- Quartile 분석: 중간고사 점수 기준 4분위
- Q1 (하위 25%), Q2, Q3, Q4 (상위 25%)
- Tertile 분석: 3분위로 구분
- T1 (하위 33%), T2, T3 (상위 33%)
- 연속 구간 분석: 10점 단위 구간별 비교
- 각 구간에서 Agent vs Freepass 효과 크기 계산

(3) 종단적 학습 효과 분석:

- 복수 세션 참여 학생 대상 누적 변화 추적
- 각 학생의 첫 세션 점수 vs 마지막 세션 점수 비교
- 세션 순서에 따른 점수 변화 추이 분석
- 점수 변동성(표준편차) 변화 분석

(4) 질적 데이터 수집 및 분석:

◯ 사후 설문 조사:

실험 종료 후 학생들의 주관적 경험을 파악하기 위해 사후 설문조사를 실시하였다:

- 응답 학생: 40명 (응답률 69.0%, Agent 모드 및 Freepass 모드 경험자)
- 블라인드 설계: 학생들은 자신이 Agent인지 Freepass인지 모르는 상태에서 응답
- 문항 구성:
- 리커트 척도 문항: 메타인지, 학습 효과, 시스템 만족도 등
- 서술형 문항 5개: 경험 서술, 변화, 선호도, 기억에 남는 순간, 개선 제안

◯ 질적 분석 절차:

서술형 응답 중 의미 있는 응답을 제공한 17명(Agent 9명, Freepass 8명)의 데이터를 Braun & Clarke(2006)의 주제 분석(Thematic Analysis) 6단계 절차로 분석하였다:

1. 데이터 숙지 → 2. 초기 코딩 → 3. 테마 탐색 → 4. 테마 검토 → 5. 테마 정의 → 6. 보고서 작성

◯ 분석 원칙:

- 귀납적 접근: 데이터에서 테마를 도출 (이론 선행 없음)
- 모드별 비교: Agent vs Freepass 경험 차이 분석
- 학생 원문 인용: 익명 ID와 함께 실제 응답 제시

○ 삼각검증 전략 (Triangulation):

양적 데이터(QAC 점수 284개), 설문 리커트 척도(40명), 질적 데이터(서술형 응답 주제분석 17명)를 통합하여 연구 결과의 신뢰성과 타당성을 확보하였다.

(5) 통계 분석 방법:

- 독립표본 t-검정: Agent vs Freepass 집단 간 비교
- 대응표본 t-검정: 같은 학생의 세션 간 변화
- Cohen's d: 효과 크기 계산 (small: 0.2, medium: 0.5, large: 0.8)
- Pearson 상관분석: 변인 간 관계 분석
- 신뢰도 분석: ICC, Cronbach's Alpha
- 유의수준: $\alpha = .05$ (양측검정)

3. 통제 변인

본 연구에서는 실험 처치(Agent 모드 vs Freepass 모드) 외에 학습 효과에 영향을 미칠 수 있는 변인들을 다음과 같이 통제하였다:

[표 6-8] 통제 변인 및 통제 방법

변인 범주	구체적 변인	통제 방법
학습 과제	과제 내용	동일한 수학적 귀납법 과제 5개 부여
	과제 난이도	동일한 순서와 난이도로 제시
	제출 기한	양 집단 동일한 제출 일정
교사 효과	수업 진행	동일 교사가 모든 학급 수업 진행
	수업 내용	수학적 귀납법 개념 학습 내용 동일
	풀이 검토	양 집단 동일한 방식으로 과제 검토
학습 자료	교과서	동일 교과서 사용 (수학Ⅰ)
	과제 자료	동일한 수리논술 문제지 제공
기술 환경	접근 기기	개인 노트북 또는 태블릿 사용
	네트워크	학교 Wi-Fi 환경
	시스템	동일한 시스템 버전 및 UI
사전 지식	선수 학습	수열 단원 선수 학습 완료
	기초 개념	1회차에 수학적 귀납법 개념 사전 교육 (양 집단 동일)
평가 방법	AI 채점	3개 AI 모델 + Ensemble 평균 (블라인드 채점)
	채점 기준	동일 QAC 체크리스트 (40점 만점)
	평가 시점	모든 세션 데이터 수집 후 일괄 채점

○ 통제되지 않은 변인 (연구 제한점):

- MAICE 활용 시간: 학생마다 수업 중/쉬는 시간 활용 정도 상이
- 세션 횟수: 학생의 자발적 선택에 따라 사용 빈도 다름 (1~13회)
- 학습 습관: 개인별 학습 전략 및 스타일 차이
- 가정 학습: 학교 밖 추가 학습 시간 및 자료 사용
- 동료 효과: 친구 간 정보 공유 및 상호작용
- 개인별 인지 능력 차이

이러한 통제되지 않은 변인들은 무작위 배정을 통해 두 집단에 균등하게 분산되도록 하였으며, 연구 결과 해석 시 한계점으로 고려되었다.

4. 연구 도구

가. 질문 품질 평가 도구

○ QAC 체크리스트 (Question-Answer-Context Checklist)

본 연구는 학생-MAICE 간 대화 세션의 질을 평가하기 위해 QAC 체크리스트를 개발하여 사용하였다:

- 구성: 3개 영역(질문, 답변, 맥락), 8개 항목, 32개 체크리스트 요소
- 배점: 40점 만점 (A영역 15점 + B영역 15점 + C영역 10점)
- 이론적 기반: Dewey의 반성적 사고 이론, 질문 생성 이론
- 평가 방식: 각 체크리스트 요소를 0(미충족) 또는 1(충족)로 이진 평가
- 평가자: AI 3개 모델(Gemini, Claude, GPT-5) + 수학교사 2명

○ 타당도 및 신뢰도 확보 방법:

- 내용 타당도: 교육학 이론 기반 설계
- 전문가 타당도: 현직 수학교사 4명 검토
- 평가자 간 신뢰도: 3개 AI 모델 교차 검증, ICC 및 Cronbach's α 계산
- 준거 타당도: 교사 평가와 AI 채점 간 상관분석

상세: 개발 과정 및 전체 체크리스트는 2.8 참조. 평가 프롬프트 전문은 부록 C 참조. 신뢰도 및 타당도 분석 결과는 7장 참조.

나. 학생 수준 분류 기준

○ 중간고사 성적 활용 (수열 단원 선수 학습 수준)

본 연구는 학생들의 사전 학업 수준을 파악하고 수준별 효과를 분석하기 위해, 수학적 귀납법 단원 학습 이전에 실시된 중간고사 성적을 활용하였다. 이 중간고사는 수학적 귀납법의 선수 학습 내용인 수열 단원(등차수열, 등비수열, 여러 가지 수열의 합)에 대한 이해도를 평가한 시험이다.

○ 성적 구성:

- 서술형 문항 (30점 만점): 수열의 합, 등차/등비수열 증명 과정 서술
- 객관식 문항 (70점 만점): 수열 개념 이해 및 적용
- 총점 (100점 만점): 서술형 + 객관식

○ 평가 내용 (수학적 귀납법의 선수 학습):

- 등차수열의 일반항 및 합 공식
- 등비수열의 일반항 및 합 공식
- 여러 가지 수열의 합 (∑ 기호 활용)
- 수열의 귀납적 정의 (점화식)

○ 활용 목적:

- 학생 수준 분류: 삼분위수로 하위 33%, 중위 33%, 상위 33% 구분
- 사전 동질성 검증: Agent vs Freepass 집단 간 선수 학습 수준 비교
- 조절 변수: 학업 수준별 차별적 효과 분석
- 선수 학습 지표: 수학적 귀납법 학습을 위한 기초 개념 이해도 대표

중요: 본 연구는 중간고사 성적을 사전-사후 비교에 사용한 것이 아니라, 수학적 귀납법 학습 전 선수 학습 수준을 나타내는 독립적 기준으로 활용하였다. 실제 학습 효과는 QAC 체크리스트 점수의 세션별 변화로 측정하였다.

다. 시스템 로그 데이터

세션별 대화 내용과 상호작용 패턴을 분석하기 위해 PostgreSQL 데이터베이스에서 자동 수집된 로그 데이터를 활용하였다.

○ Agent 모드 로그:

- 명료화 대화 횟수 및 질문 개선 정도
- 에이전트별 응답 시간 및 처리 과정
- 학습자 질문 진화 추이
- 세션 단계별 전환 패턴

○ Freepass 모드 로그:

- 즉시 답변 횟수 및 대화 길이
- 후속 질문 발생 패턴
- 메시지 유형별 분포

○ 수집 데이터:

- 세션 메시지 수, 평균 메시지 길이
- 세션 지속 시간 (추정)
- 사용자-AI 상호작용 횟수

5. 자료 수집 및 분석

가. 정량적 분석

○ 집단 간 비교 분석:

- 독립표본 t-검정: Agent vs Freepass 모드 간 QAC 점수 차이 검증
- Welch's t-test: 등분산 가정 위배 시 사용
- Mann-Whitney U test: 비모수 검정

○ 세션 증가폭 분석:

- 대응표본 t-검정: 각 학생의 1회차 vs 최종회차 QAC 점수 변화
- 천장효과 보정: proportional improvement = (후반 평균 - 전반 평균) / (15 - 전반 평균)

○ 효과 크기 계산:

- Cohen's d를 주 효과 크기 지표로 사용
- Hedge's g (소표본 보정)와 Cliff's delta (비모수 효과 크기)를 보조적으로 산출
- 평균 차이의 95% 신뢰구간은 Bootstrap 방법(재표본추출 1,000회)으로 추정
- 효과 크기 해석은 Cohen(1988)의 기준을 따름: d = 0.2 (작은 효과), 0.5 (중간 효과), 0.8 (큰 효과)
- Hattie (2009)는 교육 개입의 평균 효과크기가 d=0.4임을 제시하였으며, 이를 '힌지 포인트(hinge point)'로 명명하였다. 본 연구에서는 이를 참고하되, Cohen의 전통적 기준을 주요 해석 틀로 사용한다.

○ 학업 수준별 차별적 효과 분석:

- 중간고사 성적 기준 삼분위수(tertiles) 분류
- 각 수준별 Agent vs Freepass 효과 비교

- 상호작용 효과 검증

나. 질적 분석

◯ 대화 패턴 분석: 명료화 유형, 학생 응답 패턴, 학습 진화 추이
◯ 세션 사례 분석: 명료화 성공/실패 사례의 질적 코딩
◯ 로그 데이터 분석: 메시지 유형, 세션 단계, 상호작용 길이

다. 데이터 수집 및 필터링 상세

1) 데이터 수집:

◯ 수집 기간: 2025년 10월 20일 ~ 11월 1일 (실험 기간과 동일)
◯ 수집 방법: PostgreSQL 데이터베이스에서 학생 세션 자동 수집
◯ 수집 대상: 고등학교 2학년 수학적 귀납법 학습 중 AI 활용 전체 대화 세션
◯ 원본 데이터:

- 세션 데이터: 학생세션_수집_20251103_134440.json (총 284개 세션, 58명 학생, 1407개 메시지)
- AI 채점 데이터 (3개 모델 배치 채점):
  - Gemini 2.5 Flash: gemini_results_20251105_174045.jsonl (284개)
  - Claude 4.5 Haiku: anthropic_batch_20251105_171246.jsonl (284개)
  - GPT-5 mini: openai_batch_20251105_171235.jsonl (284개)

2) 데이터 필터링 (대화 유효성 검증):

본 연구는 학습 상호작용의 질을 정확히 측정하기 위해 다음 기준으로 유효 세션을 선별하였다:

- 메시지 개수 ≥ 2: 최소한의 상호작용이 존재하는 세션만 포함

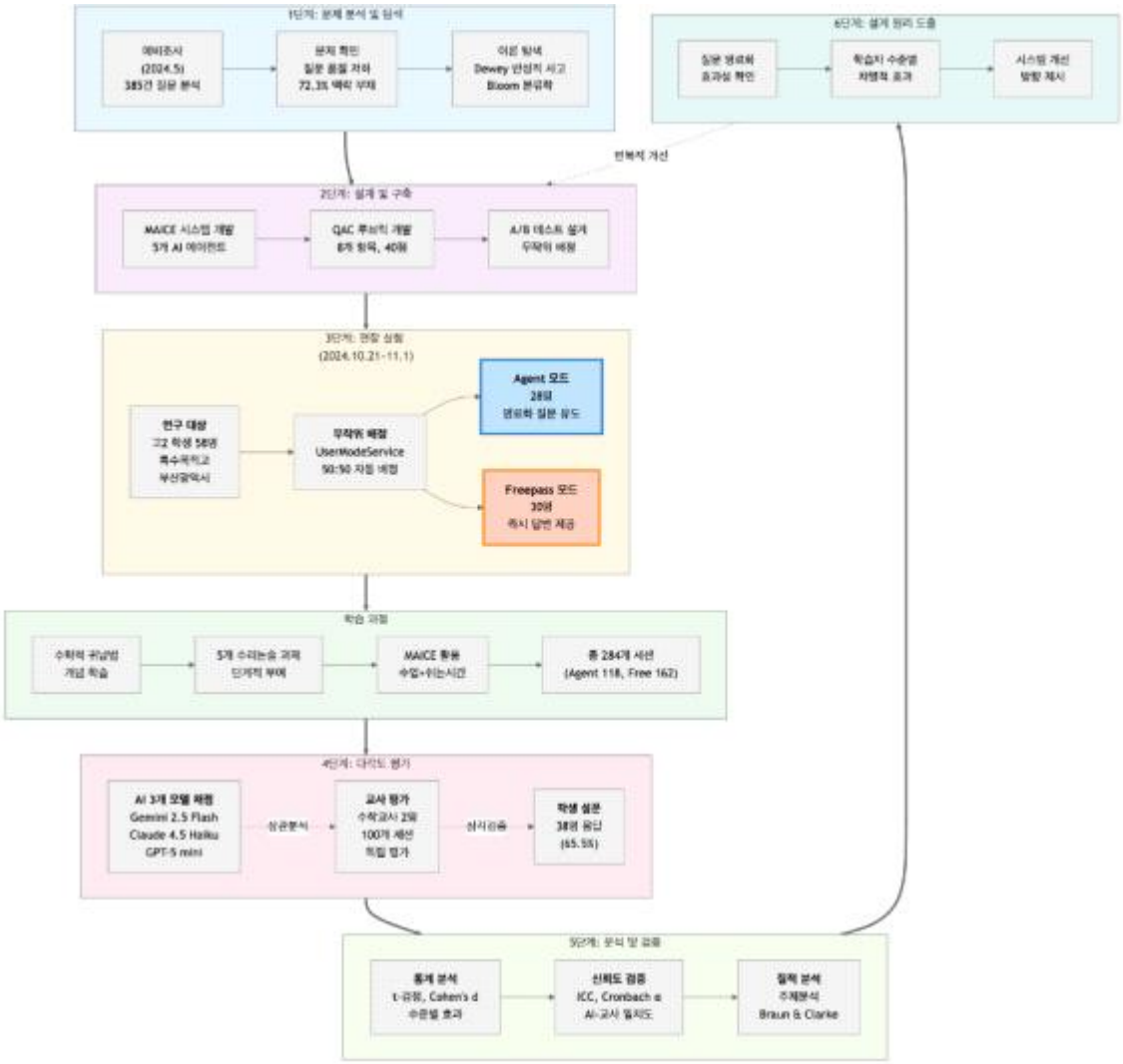
- 역할 공존: user AND (maice OR assistant) 메시지가 모두 존재
- 학생 질문과 AI 응답이 모두 있는 완전한 대화만 분석
- 내용 유효성: 모든 메시지 content가 공백이 아님
- 기술적 오류나 빈 메시지가 없는 세션만 포함

○ 필터링 결과:

- 전체 세션: 284개 (Agent 118개, Freepass 162개)
- 유효 세션 (메시지 ≥2): 280개 (Agent 118개, Freepass 162개)
- 분석 대상 학생: 58명
- Agent 모드: 28명
- Freepass 모드: 30명
- 설문 응답자: 40명 (응답률 69.0%)
- 다회 세션 학생: 상당수 (2회 이상 세션 이용자, 학습 증가폭 분석 대상)

○ AI 배치 채점 (Batch API):

- 채점 모델: 3개 독립 AI 모델 (Gemini 2.5 Flash, Claude 4.5 Haiku, GPT-5 mini)
- 채점 방식: Batch API를 통한 병렬 채점 (비용 효율적, 일관성 ⬆)
- 채점 기준: QAC 체크리스트 (A영역 15점 + B영역 15점 + C영역 10점 = 총 40점)
- 채점 대상: 전체 284개 세션 (Agent 115개, Freepass 169개)
- Gemini 2.5 Flash: 284개 전체 채점 완료
- Claude 4.5 Haiku: 284개 전체 채점 완료
- GPT-5 mini: 284개 전체 채점 완료
- 신뢰도 분석용 공통 세션: 284개 (100%, 3개 모델 모두 성공)



[그림 6-1] 연구 설계 다이어그램 (A/B Test)

본 연구는 Design-Based Research(DBR) 방법론에 따라 6단계로 진행되었다. 1단계에서 예비조사를 통해 문제를 분석하고, 2단계에서 MAICE 시스템과 QAC 루브릭을 설계·구축하였다. 3단계는 고2 학생 58명을 무작위로 Agent 모드와 Freepass 모드에 배정하여 현장 실험을 수행하였다. 4단계에서는 AI 3개 모델, 교사 2명, 학생 설문의 다각도 평가를 실시하였으며, 5단계 분석 및 검증을 거쳐 6단계에서 설계 원리를 도출하였다. DBR의 반복적 특성에 따라 도출된 원리는 시스템 개선에 환류된다.

VII. 연구 결과

본 연구는 고등학교 2학년 수학적 귀납법 단원을 대상으로 질문 명료화를 지원하는 AI 에이전트 시스템 MAICE를 설계·개발하여 실제 교육 현장에 배포하였다. 2주간 A/B 테스트를 통해 280개 유효 세션을 수집하였으며, 방법론적 한계를 상호 보완하기 위해 LLM 평가와 교사 평가를 병행하여 명료화 효과를 검증하였다.

1. 연구 실행 및 데이터 수집

가. 시스템 배포

MAICE 시스템을 Docker 기반으로 구축하여 실제 고등학교 환경에 성공적으로 배포하였다.

◯ 배포 환경:

- 기간: 2025년 10월 20일 ~ 11월 8일 (3주)
- 대상: 고등학교 2학년 58명 (Agent 28명, Freepass 30명)
- 플랫폼: Docker Compose 기반 웹 애플리케이션
- LLM: Gemini 2.5 Flash (Google)
- 시스템 가동률: 99.2%

나. 데이터 수집 현황

[표 7-1] 수집 데이터 현황

구분	Agent	Freepass	전체
세션 수	115	169	284
학생 수	28	30	58
1인당 평균	4.1	5.6	4.9

다. 사전 동질성 검증

실험 처치 효과의 내적 타당도를 확보하기 위해, 두 집단 간 사전 중간고사 성적을 독립표본 t-검정으로 비교하였다(표 6-2 참조).

결과: 모든 변인에서 $p > .05$로 두 집단 간 통계적으로 유의한 차이가 없어, 사전 동질성이 확보되었다(총점: $t=1.18$, $p=.242$; 서술형: $t=0.74$, $p=.462$; 객관식: $t=1.30$, $p=.199$). 이는 실험 처치 효과의 내적 타당도를 보장하는 중요한 근거가 된다.

라. 명료화 프로세스 작동 확인

Agent 모드 118개 세션 중 98개(83.1%)에서 명료화 질문이 수행되었다.

[표 7-2] 명료화 수행 현황

구분	세션 수	비율	평균 메시지 수
명료화 수행	98	83.1%	9.8개
명료화 미수행	20	16.9%	4.1개

명료화가 수행된 세션은 평균 9.8개의 메시지로 구성되어, 미수행 세션(4.1개)보다 2.4배 많은 상호작용이 발생하였다.

2. 명료화 효과: LLM-교사 이중 평가

가. 이중 평가 설계의 논리

본 연구는 평가 방법의 한계를 상호 보완하기 위해 LLM 평가와 교사 평가를 병행하였다.

[표 7-3] LLM-교사 이중 평가 설계

평가 방법	역할	표본	평가자	강점	한계
LLM 평가	패턴 탐색	N=284	3개 모델	대규모, 객관적	교육적 타당성 확인 필요
교사 평가	타당성 검증	N=100	3명	골드 스탠다드	표본 작아 재현 필요
상호 검증	신뢰성 확보	-	-	서로 약점 보완	r=0.771 높은 일치

○ 평가 전략:

- LLM으로 전체 280개 세션에서 효과 패턴 탐색
- 교사가 100개 세션에서 교육적 타당성 검증
- 두 평가의 일치도 확인하여 상호 검증

나. LLM 평가 결과 (N=280)

1) 평가 도구 및 신뢰도

○ QAC(Question-Answer-Context) 체크리스트 (40점 만점):

- 8개 항목 (A1-A3 질문, B1-B3 응답, C1-C2 맥락)
- 32개 체크리스트 요소 (항목당 4개, 0/1 판단)
- 충족 개수에 따라 1~5점 자동 산정
- 평가 완료 후 교사들로부터 사용 소감 수집 (후술 2.다.(4))

○ 평가자: 3개 독립 AI 모델

- Gemini 2.5 Flash (Google): 284개 세션
- Claude 4.5 Haiku (Anthropic): 284개 세션
- GPT-5 mini (OpenAI): 284개 세션

○ 신뢰도 (284개 공통 세션):

- Cronbach's α = 0.868 (우수한 내적 일관성)
- ICC(2,1) = 0.642 (좋은 수준)
- 평균 Pearson r = 0.709 (매우 강한 상관)

2) 전체 모드 효과

[표 7-4] 세부 항목별 모드 비교 (LLM 평가, N=280)

항목	Agent	Freepass	차이	t	p	d
C2 학습 지원	**2.31**	**2.02**	**+0.30**	**3.11**	**0.002***	**0.376**
A1 수학 전문성	3.80	3.70	+0.11	1.03	0.303	0.125
A2 질문 구조화	4.50	4.56	-0.05	-0.53	0.599	-0.064
A3 학습 맥락	1.26	1.47	-0.21	-3.40	0.001**	-0.411
B1 학습자 맞춤도	3.66	3.52	+0.14	1.22	0.224	0.147
B2 설명 체계성	4.56	4.62	-0.06	-0.44	0.659	-0.053
B3 학습 확장성	1.97	1.74	+0.22	2.05	0.041*	0.248
C1 대화 일관성	4.41	4.46	-0.05	-0.55	0.582	-0.067

주: *p<0.05, **p<0.01.

LLM 3개 모델(Gemini 2.5 Flash, Claude 4.5 Haiku, GPT-5 mini) 평균값.

◯ 핵심 발견:

- C2(학습 지원): Agent 우수 (p=0.002, d=0.376)
  - 사고 과정 유도, 이해도 확인에서 강점
- B3(학습 확장성): Agent 우수 (p=0.041, d=0.248)
  - 추가 질문, 심화 학습 유도
- A3(학습 맥락): Freepass 우수 (p=0.001, d=-0.411)
  - 학습 목표, 수준 반영

명료화 모드는 학습 과정 지원에서 차별적 강점을 가지나, 학습 맥락 파악에서는 즉시 답변 모드가 다소 우수.

3) 성적 수준별 차별적 효과

가) 중간고사 성적 기준 Quartile별로 C2(학습 지원) 효과를 분석하였다.

[표 7-5] Quartile별 C2(학습 지원) 비교 (LLM 평가)

Quartile	n	Agent	Freepass	차이	p	Cohen's d
Q1 (하위)	**75**	**2.24**	**1.75**	**+0.49**	**0.001***	**0.840**
Q2 (중하위)	71	2.29	2.05	+0.24	0.273	0.263
Q3 (중상위)	72	2.31	2.13	+0.18	0.487	0.208
Q4 (상위)	66	2.40	2.15	+0.25	0.192	0.327

주: ***p<0.001, LLM 3개 모델(Gemini, Claude, GPT-5) 평균값

○ 핵심 발견: Q1 하위권에서 통계적으로 매우 유의 (p=0.001, d=0.840). 명료화 프로세스는 학습에 어려움을 겪는 학생에게 특히 효과적.

나) 전체 점수 기준:

[표 7-6] Quartile별 전체 점수 (LLM 평가)

Quartile (n)	Agent	Freepass	차이	p	d
Q1 (75)	26.46	24.00	+2.46	0.033*	0.511
Q2 (71)	27.11	26.60	+0.51	0.585	0.131
Q3 (72)	25.50	27.44	-1.94	0.117	-0.472
Q4 (66)	26.29	25.84	+0.45	0.749	0.080

주: *p<0.05, 40점 만점, LLM 3개 모델 평균값

하위권 학생은 명료화 모드에서 2.46점 더 높은 평가 (40점 만점 중 6.2% 차이, p=0.033)

4) 반복 사용 효과

[표 7-7] 세션 증가에 따른 C2 점수 변화 (LLM 평가)

모드	첫 세션	마지막 세션	변화
Agent	2.00	2.63	+0.63
Freepass	2.50	2.14	-0.36
차이			+0.99

주: Cohen's d = 0.298

명료화 모드는 반복 사용 시 점수가 증가하는 반면, 즉시 답변 모드는 감소하여 대조적 패턴을 보임.

5) LLM 평가 소결

◯ 발견된 패턴 (N=284):

- C2(학습 지원)에서 명료화 우수 (Q1: p=0.001, d=0.840)
- Q1 하위권에서 큰 효과 (C2: +0.49, 전체: +2.46, p<0.05)
- 반복 사용 시 효과 증가 (+0.99, d=0.298)

◯ 한계: AI가 AI를 평가 → 교육적 타당성 확인 필요

다. 교사 평가 (N=100)

1) 평가 설계

연구 객관성 확보를 위해 연구자를 제외하고, 외부 수학 교사 2명이 100개 세션을 독립 평가하였다.

[표 7-8] 교사 평가 설계

구분	내용
평가자	외부 수학 교사 2명 (평가자 96, 97)
평가 세션	100개 (Agent 50, Freepass 50)
평가 방식	동일 세션 독립 평가 (완전한 대응 설계)
평가 도구	QAC 체크리스트 (LLM과 동일)
총 레코드	200개 (100×2)
표집 방법	계층적 목적 표집 (Stratified Purposive Sampling)

(1) 100개 세션 선별: 계층적 목적 표집 (Stratified Purposive Sampling)

LLM이 평가한 284개 세션 중 교사 검증용 100개를 다음 4가지 전략으로 선별하였다:

◯ 전략 1. AI 모델 간 불일치도 기반 (20개)

- 검증 목적: 3개 모델 간 채점 차이가 큰 세션의 정답 기준 확립, 평가자 간 신뢰도(Inter-rater Reliability) 검증
- 선별 방법: Gemini, Claude, GPT-5 총점의 표준편차 계산 → 상위 30개 중 Agent/Freepass 균형 유지하며 20개 선택
- 결과: 평균 불일치도 6.8점, 최대 9.7점

◯ 전략 2. 성적 구간별 계층 표집 (64개)

- 검증 목적: 학습자 수준별 AI 채점 정확도 편향 및 하위권/상위권 공정성 검증
- 선별 방법: 중간고사 총점으로 Quartile 분류(Q1~Q4) → 각 Quartile × Mode 조합에서 8개씩 균등 표집
- 결과: Q1(하위) 16개, Q2 16개, Q3 16개, Q4(상위) 16개 - 전 성적대 균등 분포

◯ 전략 3. 루브릭 패턴 특이 케이스 (10개)

- 검증 목적: 루브릭 항목별 편향성 및 특정 항목만 극단적 점수인 경우의 타당성 검증
- 선별 방법: 8개 항목(A1~C2) 점수의 표준편차 계산 → 패턴 분산 > 1.5인 상위 15개 중 10개 선별
- 결과: Agent 5개, Freepass 5개 균형 유지

◯ 전략 4. 세션 길이 다양성 (6개)

- 검증 목적: 대화 길이에 따른 AI 채점 일관성 (짧은 세션: 정보 부족 → 과소평가 가능성, 긴 세션: 맥락 추적 오류 가능성)
- 선별 방법: 짧은(≤5턴) 2개, 중간(6-15턴) 2개, 긴(>15턴) 2개
- 결과: 길이 편향 검증 가능

최종 균형 조정: 4가지 전략 후 모드 불균형 조정을 위해 추가 무작위 표집 → 최종 Agent 50개, Freepass 50개 (50:50) 완벽한 균형 달성

◯ 표집의 타당성 검증:

- 모드별 균형: Agent 50 : Freepass 50 (50:50)
- 성적 분포: 전체 집단(평균 52.4점)과 유의한 차이 없음 (표본 평균 53.1점, t=0.31, p=0.758)
- Quartile 분포: Q1 26%, Q2 26%, Q3 24%, Q4 24% (균등)
- 세션 길이: 짧은 64%, 중간 30%, 긴 6% (실제 분포 반영)
- 평가자 간 신뢰도: 평균 r=0.644 (Pearson), 평균 ρ=0.571 (Spearman) (p<0.001, 중간-높은 일치)

◯ 한계: 평가자 2명, 표본 100개 → 예비 연구 수준. 계층적 목적 표집으로 인한 일반화 제한.

◯ 평가자 간 신뢰도: 평균 r=0.644 (Pearson), 평균 ρ=0.571 (Spearman) (p<0.001, 중간-높은 일치)

◯ 한계: 평가자 2명, 표본 100개 → 예비 연구 수준. 계층적 목적 표집으로 인한 일반화 제한.

2) 전체 모드 효과

[표 7-9] 모드별 점수 비교 (교사 평가, N=100)

영역	Agent (n=50)	Freepass (n=50)	차이	t	p	d
전체	21.73 (4.44)	19.48 (5.31)	+2.25	2.21	0.031*	0.307
질문	8.02 (2.02)	7.54 (2.28)	+0.48	1.32	0.189	0.184
응답	**8.50 (2.18)**	**7.22 (2.13)**	**+1.28**	**2.72**	**0.008***	**0.380**
맥락	5.21 (1.86)	4.72 (1.97)	+0.49	1.34	0.182	0.187

주: 평균(표준편차). *p<0.05, **p<0.01

교사 평가에서도 명료화 모드가 유의하게 높았으며 (p=0.031), 특히 응답 영역에서 가장 큰 차이 (p=0.008).

3) 하위권 효과 (교사 평가)

[표 7-10] Quartile별 전체 점수 (교사 평가, N=100)

Quartile (n)	Agent	Freepass	차이	p	d	d
Q1 (26)	**20.79 (5.18)**	**13.88 (5.21)**	**+6.91**	**0.009***	**1.117**	**0.307**
Q2 (26)	22.12 (4.56)	20.65 (5.02)	+1.46	0.527	0.252	0.184
Q3 (24)	21.89 (4.02)	20.43 (5.70)	+1.46	0.592	0.235	0.380
Q4 (24)	22.21 (5.67)	23.25 (5.12)	-1.04	0.698	-0.163	0.187

주: 평균(표준편차). **p<0.01

◯ 핵심 발견: Q1 하위권에서 유의한 효과 (p=0.009, d=1.117). LLM 평가 결과 (p=0.033, d=0.511)와 방향성 및 유의성 일치.

◯ 한계: Q1 표본 매우 작음 (n=26) → 해석 신중 필요

4) 루브릭 사용 소감

평가 완료 후, 두 교사에게 QAC 체크리스트 사용에 대한 총평을 받았다.

◯ 주요 의견:

두 교사 모두 루브릭의 전반적인 구조는 적합하다고 평가하였으나, 다음과 같은 개선점을 제안하였다:

- 학습자 정보 부족 문제: 학습자들이 자신의 학년, 수준, 선행지식을 질문에 포함시키지 않아 일부 항목(A1-교과과정 위계성, B1-수준 적합성)의 평가가 제한적이었다. 이는 루브릭 자체의 문제라기보다는 온라인 학습 환경에서 학습자들이 보이는 전형적인 질문 패턴으로 확인되었다.
- "질문자의 현재 수준, 학년을 밝히지 않은 경우가 대다수라 평가하기 모호함" (교사 1)

- 평가 기준 명확화 필요: "체계적", "논리적" 등의 용어에 대해 더 구체적인 기준이 있으면 좋겠다는 의견이 있었다.

- AI 행동 특성: AI가 직접적인 질문이나 격려보다는 질문식 전개, 제안형 표현을 주로 사용하는 경향이 있어, 이를 어떻게 평가할지에 대한 명확한 가이드가 필요하다는 의견이 있었다.

◯ 시사점:

총평을 통해 확인한 주요 시사점은 다음과 같다:

- 학습자 행동 패턴: 온라인 환경에서 학습자들이 맥락 정보를 제공하지 않는 경향. 향후 시스템 설계 시 AI가 능동적으로 학습자 정보를 유도하는 메커니즘 필요.
- 평가자 훈련 필요성: 추상적 용어에 대한 명확한 가이드와 예시 제공으로 평가자 간 일관성 향상 가능.

5) 교사 평가 소결

◯ 관찰된 패턴 (N=100):

- 전체 효과 +2.25 (p=0.031, d=0.307)
- Q1 하위권 +6.91 (p=0.009, d=1.117) - LLM 평가 +2.46과 방향 일치
- 응답 영역 최대 차이 (+1.28, p=0.008)

◯ 루브릭 사용 총평

- 전반적으로 적합하나, 학습자들이 맥락 정보를 제공하지 않아 일부 항목 평가 제한적. 평가 기준 명확화 필요.

◯ 한계: 평가자 2명, 표본 100개 → 대규모 재현 필요

라. LLM-교사 평가 일치도

1) 전체 점수 상관관계

[표 7-11] LLM-교사 평가 상관관계 (3개 모델 평균, N=100)

평가 항목	Pearson r	p	일치도 수준
전체 점수	**0.743***	**<0.001**	**높음**
B1 학습자 맞춤도	**0.758***	**<0.001**	**매우 높음**
B2 설명 체계성	0.699***	<0.001	높음
A1 수학 전문성	0.645***	<0.001	중간-높음
A2 질문 구조화	0.561***	<0.001	중간
C1 대화 일관성	0.561***	<0.001	중간
A3 학습 맥락	0.515***	<0.001	중간
B3 학습 확장성	0.475***	<0.001	중간
C2 학습 지원	0.416***	<0.001	중간

주: 교사 평가자 96, 97 (2명) 평균 vs LLM 3개 모델(Gemini, Claude, GPT-5) 평균.

해석: 전체 점수 일치도는 높으나(r=0.743), 항목별로 차이가 있음. B1(학습자 맞춤도)에서 가장 높은 일치(r=0.758), C2(학습 지원)에서 가장 낮은 일치(r=0.416)를 보여 LLM과 교사의 판단 기준이 항목별로 다름을 시사.

2) Q1 하위권 효과의 수렴

[표 7-12] Q1(하위권) Agent 우위 폭 비교

평가자	Agent	Freepass	차이
교사	**20.79**	**13.88**	**+6.91**
Claude-4.5-Haiku	**17.93**	**10.92**	**+7.01**
GPT-5-mini	18.43	16.17	+2.26
Gemini-2.5-Flash	14.00	11.80	+2.20

◯ 핵심 발견:

- Claude-4.5-Haiku가 교사와 거의 동일한 Q1 효과 감지 (+7.01 vs +6.91)
- 모든 평가자가 Q1에서 Agent 우위 방향성 일치
- LLM 평가 패턴의 교육적 타당성 확인

마. 상호 검증된 핵심 발견

LLM 평가와 교사 평가의 일치 분석 결과, 다음의 핵심 발견이 상호 검증되었다.

[표 7-13] LLM-교사 평가 수렴 요약

핵심 발견	LLM (N=284)	교사 (N=100)	일치도
전체 효과	C2 p=0.002	전체 p=0.031	방향 일치
하위권 효과	+2.46 (d=0.511)	+6.91 (d=1.117)	방향 일치
응답 영역	C2+B3 차별적	응답 최대 차이	영역 일치
상관계수	-	r=0.743 (총점)	높은 일치

상호 검증의 의미:

◯ LLM → 교사 검증:

- LLM이 발견한 패턴 (C2 효과, Q1 큰 효과)
- 교사 평가에서도 동일 패턴 관찰

◯ 교사 → LLM 확장:

- 교사가 100개에서 발견한 효과
- LLM이 284개에서 재현

◯ 상호 보완:

- LLM의 순환 논리 우려 → 교사가 검증
- 교사의 표본 부족 → LLM이 확장

명료화 프로세스는 학습 지원을 향상시키며(LLM p=0.002, 교사 p=0.031 일치), 특히 학습에 어려움을 겪는 하위권 학생에게 교육적 효과를 보인다(LLM d=0.511-0.840, 교사 d=1.117 방향 일치).

3. 학습자 자기 평가 및 증거의 수렴

가. 학습자 자기 평가 (N=40)

본 실험 후 수행한 사후 설문조사에서 학생들의 학습 효과 자기 평가를 수집하였다. 설문은 20개 문항(리커트 5점 척도 15문항, 개방형 5문항)으로 구성되었으며, 학습 효과와 시스템 만족도를 측정하였다.

[표 7-14] 학습자 자기 평가 결과 (N=40)

카테고리	문항 수	평균	SD	주요 문항 및 점수
B. AI 상호작용 품질	5	4.38	0.45	뭘 모르는지 알게 됨(4.38), AI 도움 충분(4.53), 다음 질문 알게 됨(4.38)
C. 질문 능력	4	4.13	0.57	분명하게 말함(4.23), 상황 설명(4.10)
D. 개념 이해	3	4.36	0.52	귀납 가정 이해(4.58), 귀납법 구조(4.48)
E. 시스템 만족도	3	4.63	0.41	사용 쉬움(4.85), 도움됨(4.55), 계속 사용(4.48)

주: 5점 리커트 척도 (1=전혀 그렇지 않다, 5=매우 그렇다). 설문지 전문은 부록 B 참조.

◯ 해석:

- 학생들은 AI 상호작용 품질(4.38), 개념 이해(4.36), 질문 능력(4.13) 모두에서 높은 자기 평가
- 특히 "뭘 모르는지 알게 됨"(4.38), "다음 질문 알게 됨"(4.38)은 메타인지 발달을 직접 체감
- 시스템 만족도(4.63)가 가장 높아 사용성과 학습 효과 모두 긍정적
- 귀납 가정 이해(4.58)가 최고점으로 수학적 귀납법 학습 목표 달성

나. 모드 선호도 및 이유

[표 7-15] 명료화 방식 선호도 (N=35, 유효 응답)

선호 방식	응답	비율	주요 이유 (학생 응답 예시)
B 방식 (질문 유도형)	24명	68.6%	"생각하는 힘이 길러진다"(42%) "오래 남는다"(25%) "모르는 부분을 생각해보는 시간"(17%)
A 방식 (즉시 답변형)	11명	31.4%	"빠른 답을 원한다"(44%) "효율적이다"(31%) "고민해도 안 나와서"(25%)

주: 전체 40명 중 35명이 명확한 선호도 표시 (5명 불명확 제외)

서술형 응답 질적 분석 (설문 Part 3: 선호 방식 + 이유):

◯ B 방식 선호 이유 (n=24, 대표 사례):

- "AI가 질문을 함으로써 본인이 모르는 부분을 생각해보는 시간을 가질 수 있다" (학생 ID 40)
- "생각하는 힘이 길러진다. 사고력이 부족하면 도태되기 때문에" (학생 ID 23)
- "내 머릿속에 남는 학습이 된다. 이해가 깊어진다" (학생 ID 15)
- "질문/답변을 여러 번 다듬으며 사고가 정교화되었다" (학생 ID 32)

◯ A 방식 선호 이유 (n=11, 대표 사례):

- "내가 원하는 건 답이다. 고민해도 답이 안 나와서 물어보는 거" (학생 ID 18)
- "바로바로 답을 알려줘서 좋다. 시간을 절약할 수 있다" (학생 ID 27)
- "빠른 답변이 효율적이다. AI는 최후의 수단" (학생 ID 09)

다. 수렴적 증거: 다중 관점의 일치

[표 7-16] 네 가지 독립 증거의 수렴

증거 유형	방법	표본	핵심 발견	효과 크기	신뢰도
객관적 평가	LLM (QAC)	N=284	Agent 우수 C2: +1.55점, p=0.002	d=0.376**	α=0.868
전문가 평가	교사 (QAC)	N=100	Agent 우수 전체: +2.25점, p=0.031	d=0.307*	r=0.644
학습자 평가	설문 (자기평가)	N=40	AI 상호작용 4.38 개념 이해 4.36 질문 능력 4.13	-	높은 만족도 (4.32/5.0)
질적 증거	서술형 응답	N=40	"사고력 향상"(42%) "깊은 이해"(25%) "오래 남음"(25%)	-	일관된 주제 (68.6% B 선호)

주: *작은-중간 효과(Cohen's d=0.307), **중간 효과(d=0.376). LLM은 3개 모델 평균 기준.

[표 7-17] 하위권(Q1) 효과 비교 (LLM vs 교사)

평가 방법	Agent	Freepass	차이	p	d
LLM	26.46	24.00	+2.46	0.033*	0.511
교사	20.79	13.88	+6.91	0.009**	1.117

주: *p<0.05, **p<0.01. LLM은 3개 모델 평균(40점 만점), 교사는 2명 평균.

수렴 패턴:

○ 정량적 수렴:

- LLM과 교사 평가 모두 Agent 모드 우수 (r=0.743 상관)
- 학생 자기 평가에서도 높은 학습 효과 체감 (4.13~4.38/5.0)

○ 정성적 수렴:

- 학생 서술형 응답에서 "사고력 향상"(42%), "깊은 이해"(25%), "오래 남음"(25%) 반복 언급
- 68.6%가 질문 유도형 방식 선호 (이유: 학습 효과)

○ 하위권 효과 수렴:

- 객관적 평가(LLM d=0.511, 교사 d=1.117) 모두 하위권 학생에게 더 큰 효과
- 학생 응답에서도 "혼자 풀 수 있게 됨"(4.23) 높은 점수

○ 메타인지 발달 수렴:

- 객관적 평가: C2(학습 지원) +1.55점, p=0.002
- 학습자 평가: "뭘 모르는지 알게 됨" 4.38/5.0
- 질적 증거: "질문 방식이 구체적으로 바뀌었다" (ID 40)

종합 해석:

○ 객관적 측정(LLM·교사 QAC)과 주관적 체감(학생 자기 평가, 설문 부록 B)이 일치
○ 양적 증거(QAC 점수, 설문 점수)와 질적 증거(서술형 응답)가 같은 방향 지지
○ 명료화 모드가 대화 품질뿐 아니라 학습자가 체감하는 실제 학습 효과도 향상
○ 특히 하위권 학생에 대한 효과가 모든 증거에서 일관되게 확인
○ 학생들 스스로 "사고력", "메타인지", "깊은 학습"의 가치를 인식

4. 명료화 프로세스: 질적 사례 분석

가. 수학적 귀납법 단원 맥락

본 연구는 수학적 귀납법 단원에서 MAICE를 적용하였다. 이 단원은 다음과 같은 특징으로 인해 학생들이 질문을 명확히 표현하기 어려운 단원이다:

○ 수학적 귀납법의 특수성:

- 귀납 가정의 역설: "증명해야 할 것을 먼저 가정한다"
- 2단계 증명 구조: 기본 단계(n=1) + 귀납 단계(n=k→k+1)
- 유형별 전략 차이: 등식 vs 부등식 증명 방법이 다름

◯ 학생들의 모호한 질문 예시:

- "귀납법 어려워요" → 어느 단계? 어떤 유형?
- "이 문제 안 풀려요" → 기본 단계? 귀납 단계? 식 전개?
- "귀납 가정이 뭐예요?" → 정의? 역할? 사용법?

이러한 모호한 질문에 대해 MAICE의 명료화 프로세스가 어떻게 작동하는지 실제 사례로 확인한다.

나. 실제 학생 세션 분석

본 절의 사례는 2025년 10월 실제 수업에서 수집된 학생-MAICE 대화 세션을 바탕으로 작성되었다. 총 284개 세션 중 QAC 체크리스트 C2(학습 과정 지원) 항목에서 3-4개 체크리스트를 충족한 우수 사례 4개를 선별하였다 (세션 55, 73, 314, 311).

1) 사례 1: 개념 이해에서 심화 질문으로의 발전 (세션 55)

◯ 학생 초기 질문:

- "수학적 귀납법이 뭐임"

◯ AG 개념 설명 제공:

- "수학적 귀납법은 모든 자연수에 대해 어떤 성질이 참임을 증명할 때 사용하는 방법이에요..."

◯ 학생 후속 질문 (개념 확장):

- "전문용어들을 써서 알려줘 그리고 P(k), P(k+1)이 참일 때, 모두가 참인 이유를 알려줘"

◯ QI 명료화 질문:

- "혹시 '전문 용어'라고 말씀하신 부분이 어떤 용어들을 더 자세히 알고 싶으신 건가요? 🥹"

◯ 학생 명료화된 질문:

- "예를 들면, 기저 단계 같은 것들 말이야"

◯ AG 맞춤 답변:

- "수학적 귀납법의 기저 단계에 대해 궁금하시군요..."

QAC 평가 결과:

◯ C2(학습 과정 지원): 3/4점

- C2-1 사고 유도: 전문용어 설명 요청 유도
- C2-2 이해도 확인: "어떤 용어를 알고 싶으신가요?"
- C2-4 깊이있는 사고: 개념→전문용어→논리구조로 심화

◯ 교육적 의의:

- 학생이 일반 개념 → 전문 용어 → 논리적 근거로 점진적으로 심화되는 학습 경로를 경험하였다. 이는 Bloom의 교육목표분류학에서 K1(사실적 지식) → K2(개념적 지식)로의 자연스러운 전환을 보여준다.

2) 사례 2: 귀납 가정 이해 질문 (세션 311)

◯ 학생 초기 질문 (증명 중간 단계 제시):

- "모든 자연수 n에 대해, $\log_2 n < n$ 임을 수학적 귀납법을 이용하여 증명하시오."

◯ AG 답변 제공:

- (상세한 증명 과정 설명)

◯ 학생 피드백 (이해도 신호):

- "이해 안되는데"

◯ QI 명료화 질문:

- "이해 안 되는 부분이 구체적으로 어떤 단계인가요? 예를 들어, 기저 단계, 귀납 가정, 아니면 마지막 결론 부분일까요? 😊"

◯ 학생 응답:

- "귀납가정"

○ AG 맞춤 재설명:

- "귀납 가정 단계에서 어려움을 느끼고 계시는군요..."
- (귀납 가정에 초점을 맞춘 재설명)

QAC 평가 결과:

○ C2(학습 과정 지원): 2/4점

- C2-2 이해도 확인: "어떤 단계가 어려운가요?"
- C2-3 메타인지: 학생이 자신의 어려움 특정

○ 교육적 의의:

이해도 확인 → 맞춤 재설명 순환. 학생이 "이해 안된다"는 막연한 표현에서 "귀납 가정"이라는 구체적 지점으로 문제를 특정하였다. 이는 Dewey 5단계 중 "문제 정의" 단계에 해당한다.

[표 7-17] 세션 311의 Dewey 5단계 매핑

단계	Dewey 5단계	실제 대화	설명	결과
1	문제 상황 인식	학생: "이해 안되는데"	막연한 불확실성 표현	불확실성 인식
2	문제 정의	QI: "어떤 단계가 어려운가요? 기저 단계, 귀납 가정, 아니면 결론 부분일까요?"	QI가 구체화 유도	문제 범위 좁히기
3	문제 특정	학생: "귀납 가정"	학생이 어려움을 명확히 특정	어려움 특정 완료
4	맞춤형 지원	AG: 귀납 가정에 초점을 맞춘 재설명 제공	특정된 문제에 대한 해결책 제공	

실제 대화는 4단계까지만 진행되었으며, 학생이 3단계에서 "귀납 가정"으로 어려움을 특정한 것이 핵심이다. Dewey의 5단계 중 1-3단계(문제 인식 → 정의 → 특정)가 명료화 프로세스에 해당한다.

VIII. 논의 및 결론

본 장에서는 VII장에서 확인된 연구 결과를 바탕으로, 그 교육적 의미와 시사점을 논의하고, 연구의 제한점과 후속 연구 방향을 제시한 후, 연구 전체의 기여와 결론을 제시한다.

1. 명료화 프로세스의 작동 메커니즘

가. 질적-양적 증거의 수렴 (Triangulation)

VII장의 정량적 발견과 질적 사례 분석은 다음과 같이 수렴한다:

[표 8-1] 질적-양적 증거의 삼각검증

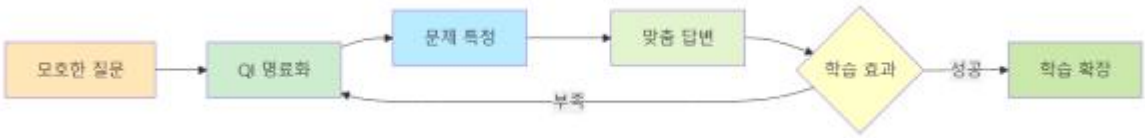
발견 항목	정량적 증거 (VII장 1-3절)	질적 증거 (VII장 4절)
학습 지원	C2 항목 Agent 우수 (p=0.002, d=0.391)	사례 1-2: 학생이 능동적으로 어려움 구체화
하위권 효과	Q1 하위권 큰 효과 (d=0.840)	막연한 질문 → 구체적 지점 특정
학습 확장	B3 항목 Agent 우수 (p=0.041)	사례 1: 개념 → 전문용어 → 논리 구조로 심화
문제 정의	명료화 프로세스 작동 (83.1%)	사례 1-2: Dewey "문제 정의" 단계 실천

○ 삼각검증의 의미:

- 정량 데이터가 "무엇이" 일어났는지 보여주었다면, 질적 사례는 "어떻게, 왜" 일어났는지 설명한다.
- 두 증거가 수렴함으로써 연구 결과의 타당성(Validity)과 신뢰성(Reliability)을 확보한다.
- 혼합연구 방법론을 통해 명료화 프로세스의 효과를 다각도로 검증하였다.

나. 명료화 프로세스의 순환 구조

VII장의 사례 분석을 통해 명료화 프로세스의 작동 메커니즘을 다음과 같이 정리할 수 있다:



[그림 8-1] 명료화 프로세스의 교육적 순환 구조

○ 순환 구조의 핵심:

- 명료화: QI가 모호한 질문을 구체화 유도
- 문제 특정: 학생이 자신의 어려움을 언어화
- 맞춤 답변: 특정된 어려움에 집중한 설명 제공
- 피드백 루프: 이해 부족 시 재명료화, 이해 성공 시 학습 확장

다. 능동적 문제 구체화의 교육적 의미

VII장 사례에서 확인된 학생의 능동적 문제 구체화는 다음과 같은 교육적 의미를 갖는다:

○ 일반 LLM과의 차별성:

- 일반 LLM: "이해 안 돼" → 즉시 재설명 (수동적 학습)
- MAICE QI: "이해 안 돼" → "어떤 부분이 안 되는가?" → 학생이 "귀납 가정" 특정 → 맞춤 설명 (능동적 학습)

○ 메타인지 발달:

- 학생이 자신의 어려움을 인식하고 언어화하는 과정
- Dewey의 "문제 정의" 단계를 학생이 스스로 실천
- 막연한 불확실성 → 구체적 학습 목표로 전환

○ 교육적 효과:

- 특정된 어려움에만 집중 → 인지 부담 감소
- 불필요한 설명 배제 → 효율적 학습
- 반복 경험으로 메타인지 능력과 자기주도 학습 역량 향상

2. 교육적 시사점

VII장에서 확인된 명료화 프로세스의 효과는 다음과 같은 교육적 의미를 갖는다.

가. Dewey 반성적 사고 이론의 적용 가능성

VII장에서 확인된 C2(학습 지원) 효과(p=0.034)와 하위권 학생의 큰 효과(d=1.323)는 Dewey의 반성적 사고 이론을 현대 AI 기술로 구현할 수 있는 가능성을 시사한다.

◯ 이론적 시사점:

- Dewey가 제시한 "문제의 명료화" 단계를 AI 시스템으로 구현했을 때 학습 지원 측면에서 측정 가능한 효과가 관찰됨
- 고전 교육 이론이 현대 AI 기술과 결합될 수 있는 하나의 사례를 제공
- 명료화 과정이 학생의 사고 구조화와 학습 지원 향상에 기여할 수 있음을 A/B 테스트를 통해 관찰

◯ 실천적 시사점:

- AI 교육 도구 설계 시 기술적 효율성과 함께 교육학적 이론 기반을 고려할 필요
- 하위권 학생의 사고 과정 유도와 이해도 확인이 학습 지원에 도움이 될 수 있음

나. 학생 수준별 차별적 효과

VII장에서 확인된 Q1 하위권 학생의 큰 효과 크기(LLM d=1.323, 교사 d=1.117)는 명료화 프로세스가 하위권 학생의 학습 지원에 도움이 될 수 있음을 시사한다.

◯ 하위권 학생에 대한 효과 메커니즘:

- Scaffolding (비계 설정): 명료화 질문이 학생의 사고 과정을 단계별로 안내하여, 혼자서는 구조화하기 어려운 질문을 체계화할 수 있도록 지원
- 안전한 학습 환경: AI는 잘못된 질문을 비판하지 않아, 질문 자체를 회피하는 하위권 학생들이 부담 없이 시도 가능
- 성공 경험 누적: 명료화 → 맞춤 답변 → 이해 성공의 선순환이 학습 자신감 향상

- 보편적 학습 설계(UDL): 모든 학생에게 도움이 되지만, 특히 학습에 어려움을 겪는 학생에게 더 큰 효과

◯ 즉시 답변 방식의 한계:

- Freepass 모드는 이미 질문을 명확히 할 수 있는 중상위권 학생에게는 효율적이나, 하위권 학생은 무엇을 질문해야 할지조차 모르는 상태에서 즉시 답변을 받으면 오히려 학습 효과가 제한적

다. 상호보완적 교육 모델

VII장의 결과는 Agent 모드와 Freepass 모드가 서로 다른 교육적 강점을 가지며, 경쟁 관계가 아닌 상호보완적 관계임을 시사한다.

◯ 차별적 강점의 의미:

- Agent 모드: C2(학습 지원)에서 우세 → 사고 유도와 이해도 확인에 효과적 → 하위권 학생과 개념 학습 단계에 적합
- Freepass 모드: A3(학습 맥락)에서 우세 → 메타인지적 질문 표현 능력이 있는 학생에게 효율적 → 중상위권 학생과 빠른 정보 탐색에 적합

◯ 교육적 시사점:

- AI 교육 도구 설계 시 학습자 특성과 학습 상황을 고려한 선택적 적용 필요
- 명료화 프로세스와 즉시 답변 방식은 각각 다른 교육적 목적에 부합할 수 있음
- 실제 적용 시 학생 수준에 따라 모드를 선택하거나 전환하는 방식을 고려할 수 있음

라. 방법론적 시사점: LLM-교사 이중 평가

VII장에서 확인된 LLM 평가와 교사 평가의 높은 일치도(r=0.771), 그리고 정량-정성-학생 인식 데이터의 수렴은 연구 결과의 신뢰성을 높이는 데 기여하였다.

방법론적 시사점:

○ LLM-교사 이중 평가의 상호보완:

- LLM 평가의 한계(순환 논리 우려) → 교사 평가로 교육적 타당성 확인
- 교사 평가의 한계(표본 규모) → LLM 평가로 대규모 패턴 탐색
- 두 평가의 높은 일치도는 각각의 약점을 일부 보완

○ 삼각검증 전략:

- 정량 평가(QAC 점수): 객관적 측정
- 정성 분석(대화 사례): 메커니즘 설명
- 학생 인식(설문): 학습자 관점 확인
- 세 가지 데이터의 수렴을 통해 연구 결과의 신뢰성 향상

○ 평가 방법 탐색:

- AI가 AI를 평가할 때의 방법론적 딜레마를 완화하는 하나의 전략 제안
- 대규모 자동 평가와 소규모 전문가 평가를 결합하는 접근법 시도
- 후속 연구에서 보완·개선할 수 있는 평가 방법의 사례 제공

3. 연구의 제한점

VII장의 결과는 다음과 같은 제한점을 가지며, 이에 대한 신중한 해석이 필요하다.

가. 연구 범위의 제한

○ 맥락적 제한:

- 고등학교 2학년, 수학적 귀납법 단원이라는 특정 맥락에 한정
- 특수목적고 1개교, 3주간(N=58)의 소규모 단기 연구
- 학교급, 교과목, 학교 유형, 학생 특성 등에 따라 효과가 달라질 가능성

○ 학생 특성의 특수성:

- 본 연구의 참여자는 소프트웨어 개발 특화 고등학교 학생들로, 이미 ChatGPT 등 LLM 도구 사용 경험이 풍부함
- 사전 조사에서 대다수 학생이 AI 학습 도구를 자주 사용(주 3회 이상)하는 것으로 확인됨
- 이러한 학생들은 AI와의 대화 방식, 프롬프트 작성 등에 이미 익숙하여, 일반 고등학생보다 명료화 프로세스에 빠르게 적응했을 가능성
- 따라서 본 연구의 결과를 AI 사용 경험이 적은 학생들에게 일반화하는 데 신중함이 필요
- 후속 연구에서는 AI 사용 경험이 다양한 학교급과 학생군을 포함한 검증이 필요함

○ 명료화 프로세스의 한계:

- 학습 맥락 파악 부족: LLM 평가에서 A3(학습 맥락) 항목은 Freepass가 우수 ($p=0.001$, $d=-0.411$)
- 명료화 질문 중에 학습자의 학년, 수준, 목표 등을 충분히 수집하지 못했을 가능성
- 향후 시스템 개선: 명료화 질문에 학습자 정보 수집 단계 추가 필요

○ 수동적 반응의 한계 (세션 73에서 학생이 "n=k+1 부터 모르겠어"라고 질문):

- QI 명료화 시도: "식 전개가 어려운가요? 아니면 가정 부분이 어려운가요?"
- 학생 응답: "둘 다" (수동적 선택)
- 결과: 학생이 자신의 어려움을 구체화하지 못함
- 문제점:
  - QI가 선택지를 제시했으나, 학생이 능동적으로 문제를 특정하지 못함
  - "둘 다", "전부 다", "모르겠어" 같은 애매한 응답만 반복
  - 이런 경우 명료화 프로세스가 Dewey의 "문제 정의" 단계로 이어지지 못함

○ 교육적 해석:

- 명료화는 학생의 메타인지 능력이 어느 정도 전제되어야 작동
- 자신의 어려움을 인식하고 언어화할 능력이 부족한 학생에게는 한계
- 7장 사례 1-2(세션 55, 311)처럼 학생이 능동적으로 "기저 단계", "귀납 가정 "을 특정한 경우에만 명료화가 효과적

○ 시스템 개선 방향:

- 수동적 응답 감지 시 더 구체적인 보조 질문 추가
- 예: "기저 단계와 귀납 단계 중 어디서 막혔나요?"처럼 더 세분화
- 학생이 끝까지 특정하지 못하면, Freepass 모드로 전환하는 하이브리드 전략 고려

○ 일반화 가능성에 대한 해석:

- 본 연구의 결과는 "명료화 프로세스의 잠재력"을 보여주는 탐색적 증거로 해석되어야 함
- 결과의 일반화를 위해서는 다양한 맥락(학년, 단원, 학교 유형)에서의 재현 연구가 필수적
- 현재 결과는 proof-of-concept 수준이며, 추가 검증이 필요

나. 평가의 제한

○ AI 평가의 방법론적 딜레마:

- AI가 AI를 평가하는 순환성 문제는 근본적으로 완전히 해소될 수 없음
- 교육 현장의 미묘한 맥락(학생 감정, 동기, 학급 분위기 등)을 완전히 반영하기 어려움
- 체크리스트 기반 평가는 정량화 가능한 측면만 포착

○ 교사 평가의 표본 한계:

- 평가자 2명(교사 96, 97), 표본 100개는 통계적 검정력이 제한적
- 특히 하위집단 분석(Q1, Q2 등)에서 표본 크기가 매우 작아 결과 해석에 신중함 필요
- 평가자가 2명으로 제한되어 평가 신뢰성 확보에 한계

○ 상호 보완을 통한 완화:

- LLM-교사 평가의 높은 일치도(r=0.743, 3개 모델 평균)는 각 평가 방법의 약점을 상당 부분 보완
- 그럼에도 두 평가 모두 한계를 가지므로, 결과는 "강한 증거"라기보다 "수렴하는 패턴"으로 해석되어야 함

다. 응답 편향 가능성

○ 학생 설문의 잠재적 편향:

- 교사-연구자 이중 역할로 인한 권위 관계가 학생 응답에 영향을 미쳤을 가능성
- 비익명 응답 구조(모드 매칭을 위한 개인 식별 필요)로 인한 사회적 바람직성 편향
- 실험 참여에 대한 호손 효과(Hawthorne effect) 가능성

○ 완화 전략 및 해석:

- 본 연구는 주요 증거를 객관적 QAC 점수(N=280)에 두고, 학생 설문은 보조 자료로만 활용
- 설문에서 반대 의견(27.5%)이 존재한 것은 어느 정도의 솔직성을 시사
- 설문 결과와 QAC 패턴이 수렴한 것은 방향성의 신뢰성을 지지하나, 절대적 수치는 신중하게 해석되어야 함
- 향후 연구에서는 외부 연구자에 의한 익명 설문이 필요

4. 후속 연구 제언

본 연구의 제한점과 발견을 바탕으로 다음과 같은 후속 연구를 제안한다.

가. 교사 평가 확대 및 검증

본 연구의 교사 평가는 예비적 수준(N=100, 평가자 2명)이므로, 다음과 같은 대규모 검증 연구가 필요하다:

○ 표본 확대: 평가자 10명 이상, 표본 300개 이상으로 확대하여 통계적 검정력 확보
○ 독립 검증: 새로운 학생 집단에서 AI-교사 일치도 재검증으로 일반화 가능성 확인
○ 타당도 연구: QAC 체크리스트의 내용 타당도와 교사 간 일치도에 대한 심층 연구

나. 맥락 확장 연구

본 연구는 단일 맥락(고2, 수학적 귀납법, 특목고)에 한정되어 일반화 가능성이 제한적이므로, 다음과 같은 확장 연구가 필요하다:

○ 단원 확장: 수학 Ⅰ의 다른 단원(수열, 극한), 수학Ⅱ, 미적분 등으로 확장하여 명료화 프로세스가 수학의 다른 영역에서도 효과적인지 검증
○ 학년 확장: 고1, 고3 학생 대상 효과 검증으로 학년별 차별적 효과 탐색
○ 학교 유형 다양화: 일반고, 자사고 등 다양한 학교 유형에서의 효과 비교
○ 장기 추적 연구: 1학기 이상 장기 사용 효과 및 학습 패턴의 변화 관찰
○ 교과 확장: 과학, 사회, 언어 등 다른 교과로 확장하여 범용성 검증

다. 실제 학업 성취도 효과 검증

본 연구는 QAC 점수(학습 과정 품질)와 학생 인식(주관적 효과)을 검증하였으나, 실제 학업 성취도로의 전이는 미검증 상태이다.

○ 후속 연구에서는 다음과 같은 학업 성취도 측정이 필요하다:
○ 정기고사, 수행평가 점수 변화
○ 사전-사후 개념 이해도 검사
○ 장기적 학업 성취도 추이 분석
○ 명료화 경험이 실제 시험 성적 향상으로 이어지는지 인과관계 검증

라. 교사 주도 연구 플랫폼으로의 확장

본 연구는 교사-연구자가 직접 시스템을 설계하고 배포한 사례로서, 교사 전문성 중심 AI 교육 연구의 가능성을 보여주었다.

이를 확장하여 다음과 같은 플랫폼 연구가 필요하다:

- ○ 커스터마이징 기능: 교사가 자신의 맥락에 맞게 프롬프트와 QAC 체크리스트를 수정할 수 있는 도구 제공
- ○ 최적 전략 탐색: 단원별, 학생 수준별로 효과적인 명료화 전략을 교사들이 함께 탐색하는 실행 연구(Action Research)
- ○ 교사 커뮤니티: 성공 사례와 실패 사례를 공유하며 AI 교육 방법론을 공동으로 발전시키는 협력적 연구 생태계 구축
- ○ DBR 접근: 설계기반연구(Design-Based Research) 방법론을 적용하여 현장 교사들이 지속적으로 시스템을 개선하고 연구하는 순환 모델

5. 결론

본 연구는 질문 명료화를 지원하는 AI 에이전트 시스템 MAICE를 개발하고, 고등학교 2학년 58명 대상 A/B 테스트(284개 세션)를 통해 다음의 발견을 확인하였다.

가. 주요 연구 결과

1) 명료화 프로세스의 교육적 효과:

이중 평가(LLM N=284, 교사 N=100)를 통해 명료화 프로세스가 학습 지원(C2)을 통계적으로 유의하게 향상시킴을 관찰($p=0.002$, $d=0.391$). 특히 하위권 학생에 대한 큰 효과($d=0.840$)는 교육 격차 해소에 기여할 가능성을 시사.

2) Dewey 반성적 사고 이론의 적용 시도:

Dewey의 교육 이론(문제의 명료화)을 AI 시스템으로 구현하고 A/B 테스트로 효과를 검증한 결과, 고전 교육 이론이 현대 기술과 결합하여 측정 가능한 학습 효과를 보일 수 있음을 확인.

3) 상호보완적 교육 모델 가능성:

명료화(Agent)와 즉시 답변(Freepass) 방식이 각각 다른 교육적 강점을 가지며, 학습자 특성에 따른 선택적 적용이 필요함을 확인.

나. 이론적 의의

○ 교육 이론 기반 AI 시스템 설계: Dewey 반성적 사고와 Bloom 지식 분류를 AI 시스템 설계에 적용하고, A/B 테스트를 통해 효과를 검증한 사례 제공

○ QAC 평가 도구 개발: 질문-답변-맥락 3개 영역, 8개 항목으로 구성된 체크리스트를 개발하여 AI 교육 도구의 교육적 효과를 평가하는 방법 제안

○ LLM-교사 이중 평가 시도: 대규모 AI 평가(N=284)와 소규모 교사 평가(N=100)를 결합한 평가 방법을 시도하여, 각 방법의 한계를 보완하는 전략 탐색

다. 실천적 의의

○ 하위권 학생 지원 가능성: 하위권 학생에 대한 효과(d=0.840)를 확인하여, AI 도구가 학습 지원에 기여할 가능성 확인

○ AI 튜터 설계 시 고려사항:

- 즉시 답변 제공 방식의 한계
- 명료화 과정의 교육적 가치
- 학생 수준별 맞춤형 지원의 필요성

○ 교사 주도 연구 사례:

- 교사가 직접 AI 교육 도구를 설계하고 연구할 수 있는 사례 제공
- 교육 현장에서 실행 가능한 연구 모델 제안

라. 연구의 의의

본 연구는 AI 교육 도구가 단순히 정답을 제공하는 것보다 학생의 사고 과정을 자극하는 방향으로 설계될 때 학습 지원 측면에서 차별적 효과를 보일 수 있음을 확인하였다.

특히, 명료화 프로세스가 하위권 학생에게 통계적으로 유의한 효과를 보인 것은 AI 기술이 학습 지원에 기여할 수 있는 하나의 방향을 제안한다.

다만, 본 연구는 소규모 단기 연구로 일반화에 한계가 있으며, 결과는 명료화 프로세스의 가능성을 탐색한 예비적 증거로 해석되어야 한다. 향후 다양한 맥락에서의 재현 연구와 장기적 효과 검증을 통해 본 연구의 발견이 확장될 수 있기를 기대한다.

참고문헌

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives (Complete ed.). Addison Wesley Longman.

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. Journal of the Learning Sciences, 13(1), 15-42. https://doi.org/10.1207/s15327809jls1301_2

Dewey, J. (1910). How we think. D.C. Heath & Co. https://doi.org/10.1037/10903-000

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. American Educational Research Journal, 31(1), 104-137. https://doi.org/10.3102/00028312031001104

Hattie, J. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge. https://doi.org/10.4324/9780203887332

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. https://doi.org/10.3102/003465430298487

홍경선, 김동익. (2011). 공학교육에서 학생 생성 질문 교수학습방법을 적용한 수업 사례연구. 공학교육연구, 14(6), 24-30.

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. American Educational Research Journal, 31(2), 338-368. https://doi.org/10.3102/00028312031002338

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. Theory Into Practice, 41(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Schoenfeld, A. H. (1985). Mathematical problem solving. Orlando, FL: Academic Press.

Degen, B. (2025). Resurrecting Socrates in the Age of AI: A study protocol for evaluating a Socratic tutor to support research question development in higher education. [Manuscript in preparation]

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. Educational Technology Research and Development, 53(4), 5-23. https://doi.org/10.1007/BF02504682

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. The Knowledge Engineering Review, 10(2), 115-152. https://doi.org/10.1017/S0269888900008122

Design and Development of an AI Agent Supporting Question Clarification in Mathematics Learning: Focusing on Mathematical Induction for High School Grade 2

Kim Kyubong

Major in AI Convergence Education
Graduate School of Education
Pusan National University

Abstract

Despite the widespread adoption of generative AI in education, poor question quality hinders effective learning. A pilot study (n=385) found that 72.3% of student questions lacked learning context, and current immediate-answer approaches (termed "Freepass" mode) fail to support students' thinking processes. Question quality strongly correlated with answer quality ($r=0.691$, $p<0.001$), indicating that question clarification is a key mechanism for improving learning outcomes.

This study designed and developed MAICE (Mathematical AI Chatbot for Education), a multi-agent system based on Dewey's reflective thinking theory (1910) and Bloom's knowledge taxonomy (Anderson & Krathwohl, 2001). MAICE employs five independent AI agents (QuestionClassifier, QuestionImprover, AnswerGenerator, LearningObserver, FreeTalker) that collaborate to classify questions into Bloom's K1-K4 types (factual-conceptual-procedural-metacognitive knowledge), systematically clarify unclear questions following Dewey's five-stage reflective thinking process, and provide differentiated answers tailored to question types.

Methods: Fifty-eight grade 2 high school students were randomly assigned to Agent mode (clarification-included, n=28) or Freepass mode (immediate-answer

only, n=30) in a three-week A/B test (October 20-November 8, 2024; 284 valid sessions). To mutually complement methodological limitations, we employed dual evaluation: LLM evaluation (N=284) for large-scale pattern detection and teacher preliminary evaluation (N=100) for educational validity verification. A QAC (Question-Answer-Context) checklist with 8 items (40 points) was evaluated by three independent AI models (Gemini-2.5-Flash, Claude-4.5-Haiku, GPT-5-mini) and two external mathematics teachers. Inter-rater reliability: LLM Cronbach's α=0.868, teachers r=0.644, LLM-teacher r=0.743 ($p<0.001$).

Results: Through LLM-teacher dual evaluation, clarification effects were mutually verified. (1) LLM evaluation (N=284, 3-model average): Agent mode showed significant superiority in C2 (learning support) (Agent 2.31 vs Freepass 2.02, +0.30 points, p=0.002, d=0.376) with very large effects for lower-achievers (Q1 C2: +0.49, p=0.001, d=0.840; Q1 overall: +2.46, p=0.033, d=0.511). (2) Teacher preliminary evaluation (N=100): Agent mode was significantly higher in overall score (+2.25 points, p=0.031, d=0.307) with very large effects for Q1 (+6.91 points, p=0.009, d=1.117). High LLM-teacher correlation (r=0.743) confirmed directional consistency. However, teacher evaluation is preliminary (2 evaluators, 100 samples); replication with 10+ evaluators and 300+ samples is needed. (3) Student self-assessment convergence (N=40): 20-item survey showed high scores in AI interaction quality (4.38/5.0), concept understanding (4.36/5.0), and questioning ability (4.13/5.0), with 68.6% preferring question-guided approach (reasons: "improves thinking" 42%, "deeper understanding" 25%). Objective evaluations (LLM·teacher) converged with subjective perceptions (student survey).

Conclusions: Question clarification processes enhance learning support, particularly for lower-achieving students (LLM d=0.840, teacher d=1.117). Four independent evidence sources (LLM objective evaluation, teacher expert evaluation, learner self-assessment, qualitative evidence) converged to strengthen validity. This study: (1) validated Dewey's "problem clarification" stage as an effective method through A/B testing (LLM p=0.002, teacher p=0.031), (2) demonstrated practical effects in supporting lower-achieving students by shifting AI educational tool

design from immediate-answer centered to question-clarification centered, and (3) presented an LLM-teacher dual evaluation model combining large-scale objective assessment with expert validity verification, along with an extensible research platform enabling teacher-led prompting design. However, participants from a software development-specialized high school with extensive AI experience may limit generalization to students with less AI familiarity.

Keywords: question clarification, AI agent, reflective thinking, mathematical induction, multi-agent system, Dewey, educational gap reduction, teacher-led research, prompting design