Journal of KIIT. Vol. 24, No. 2, pp. 169-178, Feb. 28, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.2.169

Multi-LLM 파이프라인 기반 적응형 튜터링 시스템 콘텐츠 자동 생성: 통계교육 사례를 중심으로

장필식*, 이주양**

Automated Content Generation for Adaptive Tutoring Systems using a Multi-LLM Pipeline: A Case Study in Statistics Education

Phil-Sik Jang*, Ju-Yang Lee**

이 논문은 2026년도 세한대학교 교내 연구비 지원에 의하여 씌여진 것임

요 약

적응형 튜터링 시스템(ITS)은 개인화된 학습 경로와 피드백을 제공하여 학습 효과를 향상시키지만, 고품질 콘텐츠 제작에 막대한 인적 자원이 소요된다. 본 연구는 Multi-LLM 파이프라인을 활용하여 적응형 튜터링 시스템용 교육 콘텐츠를 자동 생성하는 방법론을 제안하고, 인간 전문가 평가를 통해 그 효과를 검증하였다. 연구는 Phase 1(스킬 추출)과 Phase 2(문제 생성)로 구성되었으며, Phase 1에서는 교재로부터 학습 스킬을 추출하기 위해 다양한 LLM 조합을 비교하였다. 분석 결과, GPT가 생성자 역할에서 다른 모델들에 비해 우수한 성능을 보였으며, 'GPT→Claude' 파이프라인이 최고 성능을 기록하였다. Phase 2에서는 'Claude→GPT' 파이프라인이 가장 높은 평균 점수와 함께 품질 안정성 측면에서도 우위를 보였다. 본 연구는 Multi-LLM 기반 적응형 학습 콘텐츠 생성의 실용적 설계 기준을 제시한다는 점에서 의의가 있다.

Abstract

Intelligent Tutoring Systems (ITS) improve learning outcomes through personalized learning paths and adaptive feedback, but developing high-quality instructional content requires substantial human resources. This study proposes an automated content generation methodology for ITS using a Multi-LLM pipeline and evaluates its effectiveness through expert human assessment. The approach consists of two phases: skill extraction and problem generation. Experimental results show that GPT performed best as a generator in Phase 1, with the 'GPT→Claude' pipeline achieving the highest performance, while in Phase 2 the 'Claude→GPT' pipeline demonstrated both the highest mean score and superior quality stability. These findings provide practical design guidelines for Multi-LLM-based automated generation of adaptive learning content.

Keywords

adaptive tutoring systems, automated content generation, multi-LLM pipeline, skill extraction, problem generation

* 세한대학교 항공물류학과 교수
- ORCID: https://orcid.org/0009-0004-3782-131X
** 순천향대학교 글로벌경영대학 SCH 경제경영연구소 연구부교수(교신저자)
- ORCID: https://orcid.org/0009-0002-9677-0415

· Received: Jan. 26, 2026, Revised: Feb. 20, 2026, Accepted: Feb. 23, 2026

· Corresponding Author: Ju-Yang Lee
Institute of SCH Economics and Management, College of Global Business, Soon Chun Hyang University, Korea
Tel.: +82-41-359-6112, Email: sky9357821@naver.com

Ⅰ. 서 론

적응형 튜터링 시스템(ITS, Intelligent Tutoring System)은 학습자의 지식 상태를 진단하고 개인화된 학습 경로와 적응적 피드백을 제공함으로써 학습 효과를 향상시키는 교육 기술이다. 메타분석 연구에 따르면 ITS는 전통적 교실 수업 대비 유의미한 학습 성취 향상을 보이는 것으로 보고되고 있다[1]. 이러한 학습 효과의 핵심에는 학습자 모델링과 지능적 피드백 제공이 있으며, 특히 BKT(Bayesian Knowledge Tracing)과 같은 알고리즘을 통해 학습자의 지식 단위(Knowledge component) 또는 스킬(Skill) 숙달도를 확률적으로 추정하는 방식이 널리 활용되고 있다[2]. UC Berkeley에서 개발된 대표적인 오픈소스 적응형 학습 시스템인 “OATutor”는 학습자 모델링과 단계별 힌트 제공, 오류 기반 피드백을 통한 튜터링 개입을 핵심으로 하는 적응형 튜터링 시스템(ITS)에 해당한다[3]. 이러한 구조는 OATutor가 단순한 개인화 학습 플랫폼을 넘어, 학습 과정 중 개입(Intervention)을 수행하는 지능적 튜터링 시스템으로 기능함을 의미한다.

그러나 ITS의 교육적 효과가 반복적으로 입증되었음에도, 고품질 학습 콘텐츠 제작에 드는 막대한 인적 자원은 시스템 확산의 주요 병목으로 작용해 왔다. T. Murray[4]는 1시간 분량의 ITS 콘텐츠를 개발하는 데 약 200–300시간의 전문가 작업이 필요하다고 보고하였으며, 이는 ITS가 이론적으로는 효과적이지만 실무적으로는 확장에 제약이 있음을 시사한다. 이러한 문제는 실제 시스템 개발 사례에서도 확인된다. OATutor의 경우도, 25명 이상의 콘텐츠 제작자가 약 3년에 걸쳐 교과서 기반 학습 콘텐츠를 구축한 것으로 보고되었다[3]. 이는 ITS의 확산과 지속적 운영을 위해 콘텐츠 제작 과정의 자동화 및 효율화가 필수적임을 보여준다.

최근 대규모 언어모델(LLM, Large Language Model)의 발전은 이러한 콘텐츠 제작 병목을 완화할 수 있는 새로운 가능성을 제시하고 있다. LLM은 자연어 이해와 생성 능력을 바탕으로 문제, 힌트, 설명과 같은 교육 콘텐츠를 자동으로 생성할 수 있는 잠재력을 지닌다. Z. A. Pardos and S. Bhandari[5]의 무작위 대조 실험 연구에서는 ChatGPT가 생성한 수학 학습용 도움말이 인간 튜터가 작성한 도움말과 비교하여 통계적으로 유의미한 학습 성과 차이를 보이지 않는 것으로 나타났다. 국내에서도 S. M. Lim et al.[6]의 사회과 자동 문항 생성 연구, M. H. Kim et al.[7]의 수능 국어 문법 문제 생성 시스템 연구 등 LLM을 활용한 교육 콘텐츠 자동 생성 가능성을 탐색하는 연구가 점차 증가하고 있다. 이러한 연구들은 LLM이 교육 콘텐츠 생성 과정에서 일정 수준의 교육적 타당성을 확보할 수 있음을 시사한다.

그럼에도 불구하고, 단일 LLM에 기반한 콘텐츠 생성 접근법은 여러 한계를 지닌다. 선행연구에 따르면 LLM은 사실과 다른 내용을 생성하는 할루시네이션(Hallucination), 출력 간 일관성 부족, 맥락적 오류 등의 문제를 보일 수 있으며, 이는 교육 콘텐츠의 신뢰성과 타당성에 직접적인 위협이 될 수 있다[8]. 실제로 LLM이 생성한 교육용 콘텐츠 중 일부는 전문가 검토 과정에서 수정 또는 폐기되는 사례도 보고되고 있으며[5], 이는 단일 모델 호출만으로는 적응형 튜터링 시스템에 즉시 활용할 수 있는 수준의 콘텐츠를 안정적으로 생산하기 어렵다는 점을 시사한다.

한편, 기존의 LLM 기반 교육 콘텐츠 생성 연구는 문제 생성이나 힌트 생성과 같은 단일 단계 작업에 초점을 맞추거나, 자동화된 평가 지표에 의존한 성능 비교에 머무르는 경향이 있다. 교재로부터 스킬을 체계적으로 추출하고, 이를 기반으로 연습 문제와 힌트를 생성하여 ITS에 즉시 활용이 가능한 형태로 연결하는 End-to-End 파이프라인에 대한 실증 연구는 아직 제한적이다. 특히 각 단계의 산출물에 대해 인간 전문가 평가를 통해 교육적 품질을 검증한 연구는 국내외를 막론하고 찾아보기 어렵다.

이에 본 연구는 다중 LLM(Multi-LLM) 파이프라인을 활용하여 적응형 튜터링 시스템용 교육 콘텐츠를 자동 생성하고, 그 품질을 인간 전문가 평가를 통해 체계적으로 검증하는 것을 목적으로 하였다. 이를 위해 교재로부터 학습 스킬을 추출하는 Phase 1(Skill extraction)과 추출된 스킬 기반으로 연습 문제를 생성하는 Phase 2(Problem generation)로 구성된 2단계 파이프라인을 설계하였다. 아울러 단일 LLM과 Multi-LLM 조합을 포함한 다양한 파이프라인 구

성을 비교함으로써, 생성자와 검증자로서 모델 역할 분담이 콘텐츠 품질에 미치는 영향을 분석하였다.

II. 이론적 배경 및 관련 연구

2.1 적응형 튜터링 시스템

H. S. Nwana[9]는 ITS의 핵심 구성요소를 교과 지식 표현(Domain model), 학습자 상태 추정(Student model), 교수 전략 선택(Tutor model), 상호작용(User interface)로 제시하였다. 이 중 Student model은 학습자의 학습 진행 상황을 정교하게 추적하고 적응적 피드백을 제공하는 핵심 모듈로 기능한다.

Student model을 효과적으로 구현하기 위해서는 교과 지식을 세분화된 단위로 구조화하는 것이 필수적이며, 이러한 단위를 KC(Knowledge Component) 또는 스킬(Skill)이라 한다. K. R. Koedinger et al.[10]는 KC를 학습자가 특정 과제를 수행하기 위해 요구되는 지식 또는 인지적 절차의 최소 단위로 정의하며, 교수·학습·평가를 연결하는 이론적 틀을 제시하였다. ITS에서는 KC 단위로 학습자의 숙달도를 추적하고, 이를 기반으로 다음에 제시할 과업을 선택한다.

본 연구에서 활용한 OATutor는 학습 콘텐츠를 Problem → Steps → Hints/Scaffolds의 계층 구조로 구성하고, 각 Step을 KC에 연결하여 스킬별 숙달도를 추적한다. 시스템은 학습자가 레슨에 포함된 스킬의 목표 숙달 임계값에 도달할 때까지 적응적으로 문제를 선택·제시한다. 이러한 구조는 콘텐츠 자동 생성 시 KC 추출과 문제 생성이 분리된 단계로 설계되어야 함을 시사하며, 본 연구의 2단계 파이프라인 설계의 이론적 근거가 된다.

2.2 LLM 기반 교육 콘텐츠 자동 생성

자동문항생성(AIG, Automatic Item Generation)은 대규모 문제은행 구축에 드는 시간과 비용을 줄이기 위한 접근으로 오랫동안 연구되어 왔다. 전통적 AIG는 문항 구조와 변인을 사전에 정의한 템플릿 기반 생성 방식이 주류였으나[11], 문항의 다양성과 맥락적 자연스러움 측면에서 한계가 있었다. 최근 LLM의 등장은 템플릿에 덜 의존하면서도 자연스러운 문항 생성 가능성을 열었고, 이에 따라 관련 연구가 빠르게 확장되고 있다. G. Kurdi et al.[12]는 교육 목적의 자동 질문 생성 연구를 종합적으로 정리하며, 생성된 문항의 정확성, 교육적 적절성, 검증 절차의 중요성을 강조하였다. Z. Wang et al.[13]는 GPT-3를 활용한 교과서 기반 질문 생성 연구를 통해 LLM의 교육 콘텐츠 생성 잠재력을 보여주었다. 국내에서도 S. H. Son et al.[14]이 LLM을 활용한 맞춤형 학습 지원 시스템을 개발하여 적용 가능성을 제시하였다. 그러나 기존 연구들은 개별 기능 또는 단일 과업 중심의 탐색적 접근이 대부분이며, 교재 기반 콘텐츠를 ITS에 즉시 투입할 수 있는 구조화된 산출물로 연결하는 End-to-End 파이프라인 연구는 찾아보기 어렵다.

2.3 Multi-LLM 협업 시스템

단일 LLM 기반 접근은 할루시네이션, 일관성 부족, 자기 출력 과대평가(Self-Enhancement bias)와 같은 한계를 가질 수 있으며, 교육 콘텐츠 생성에서는 이러한 문제가 학습자 경험과 학습 성과에 직접적인 영향을 미칠 수 있다. 이러한 한계를 완화하기 위한 대안으로 Multi-LLM 협업 접근이 주목받고 있다.

L. Zheng et al.[15]는 LLM-as-a-Judge 패러다임을 통해 LLM을 평가자로 활용하는 접근의 가능성과 함께, 자기 모델 출력에 유리한 평가 편향이 발생할 수 있음을 지적하였다. 이를 보완하기 위해 이종 모델 간 교차 검증(Cross-Model verification)이나 복수 판정자 구성과 같은 전략이 제안되고 있다. A. Madaan et al.[16]는 Self-Refine 접근을 통해 생성 - 피드백 - 수정의 반복적 정제 과정이 단일 생성 방식 대비 출력 품질을 유의미하게 향상시킬 수 있음을 보였다. 이러한 연구들은 생성 단계와 검증·정제 단계를 분리하고 역할을 분담하는 파이프라인 설계가 단일 모델의 한계를 보완할 수 있음을 시사한다.

본 연구는 이러한 선행연구를 바탕으로, 이종 Multi-LLM 역할 분담(GPT, Claude, Gemini)을 통한 End-to-End 파이프라인을 설계하고, 인간 전문가 평가를 통해 생성된 콘텐츠의 교육적 품질을 검증한다는 점에서 기존 연구와 차별성을 갖는다.

III. 연구 방법

3.1 연구 설계 개요

본 연구에서 구현한 콘텐츠 생성 파이프라인과 실험 프레임워크를 도식화하면 그림 1과 같다.

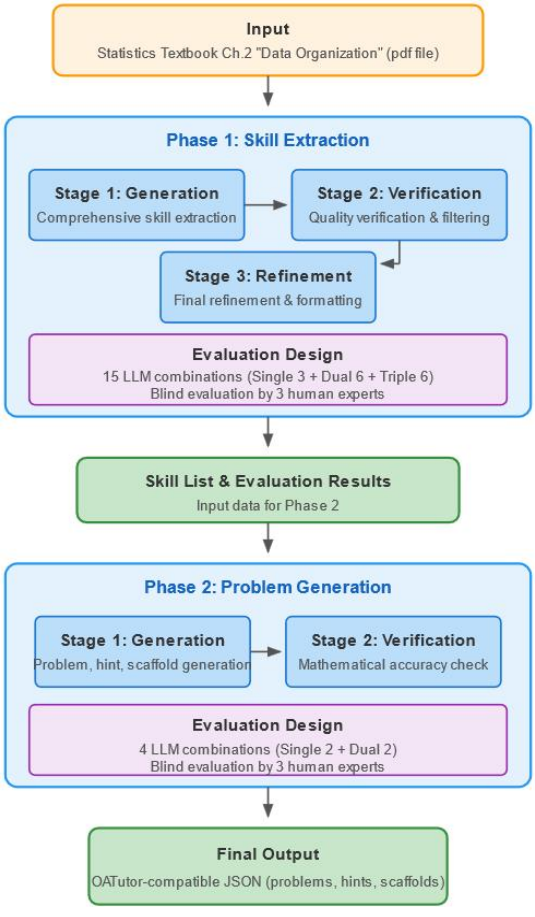


그림 1. 콘텐츠 생성 파이프라인과 실험 프레임워크
Fig. 1. Content generation pipeline and experimental framework

파이프라인은 교재에서 학습 스킬(Knowledge component)을 추출하는 Phase 1(Skill extraction)과, 추출된 스킬을 기반으로 문제와 힌트를 생성하는 Phase 2(Problem generation)로 구성된다. 입력 자료로는 통계 교재[17]의 2장 '자료의 정리 및 요약' pdf 파일을 사용하였으며, 최종 산출물은 OATutor 시스템에 즉시 적용할 수 있는 JSON 형식의 교육 콘텐츠이다.

본 연구에서 '학습 스킬(KC, Knowledge Component)'은 K. R. Koedinger et al.[10]의 정의에 따라 "주어진 과제 상황에서 일정한 연습을 통해 습득 및 적용되는 독립적인 지식 단위"로 조작적 정의하였다. 추출 시에는 "측정 가능성", "독립성", "적정 세분성", "BKT 호환성" 의 네 가지 기준을 적용하였다.

실험에 사용된 LLM 모델은 GPT(gpt-5.2), Claude(claude-opus-4-5-20251101), Gemini(gemini-3-flash-preview)의 세 가지이며, 각 모델은 API(Application Programming Interface) 방식으로 호출되어 파이프라인 내에서 생성자(Generator)나 검증자(Verifier) 또는 정제자(Refiner)의 역할을 수행하도록 설계하였다. 즉, 단일(Single) 파이프라인은 LLM 모델 하나가 모든 역할을, 삼중(Triple) 파이프라인은 세 개 LLM 모델이 역할을 나누어 수행하게 된다. Phase 1에서는 Single(3개), Dual(6개), Triple(6개) 구성의 총 15개 LLM 조합을 비교하였다. Phase 2에서는 문제 및 힌트 생성 단계의 실험 효율성과 비교의 명확성을 높이기 위해 Phase 1의 실험 결과 상위 성능 조합에 포함된 LLM 모델을 이용하여 실험을 수행하였다.

3.2 파이프라인 설계

Phase 1의 Dual 파이프라인에서 입력 데이터는 pdf 파일 형식이며, 별도의 텍스트 추출·변환 과정 없이 각 LLM API의 네이티브 PDF 처리 기능을 활용하여 PDF를 직접 입력하였다. GPT의 경우 API 파일 업로드 기능으로 PDF 첨부하고, 내장 PDF 파서가 텍스트 자동 추출하도록 하였으며, Claude의 경우, Document 타입으로 PDF를 Base64 인코딩 후 전달, 페이지별 인식하도록 하였다. Gemini는 Google AI API의 upload_file() 함수로 PDF 업로드, 멀티모달 입력으로 처리하였다.

생성 단계(Stage 1)의 프롬프트는 그림 2처럼, "가능한 스킬 후보를 포괄적으로 추출하라"는 지시를 통해 재현율(Recall)을 극대화하도록 설계하였으며, 3.1에서 언급한 네 가지 기준을 품질 기준(Criteria)으로 활용하였다. 검증 단계(Stage 2)의 프롬프트는 동일한 품질 기준을 공유하되, 역할을 '엄격한 품질

검증 전문가'로 변경하여 정밀도(Precision) 향상을 목표로 설계하였다. 구체적으로, Stage 1의 출력 JSON을 입력으로 받아 다음 세 가지 기준에 따라 스킬 목록을 평가하도록 지시하였다: (1) 의미적으로 중첩되는 스킬의 병합 또는 제거, (2) 적응형 학습 시스템에서 문항으로 측정할 수 없는 추상적 스킬의 제거, (3) 스킬 명이나 설명이 불분명한 스킬의 명확화 또는 제거. 검증 모델은 수정된 스킬 목록과 함께 각 변경 사유를 출력하도록 하여 검증 과정의 투명성을 확보하였다.

Triple 파이프라인의 정제 단계(Stage 3)에서는 '교육과정 설계 최종 검토자' 역할의 LLM이 Stage 2의 출력을 입력으로 받아, 의미적으로 중첩되는 스킬의 통합, 용어 표준화, 서술 형태의 일관성 확보를 Prompt Chaining 방식으로 수행하고, 각 변경 사유를 함께 반환하도록 설계하였다.

```
# 역할: 교육 콘텐츠 생성 전문가
# 작업: 통계학 교재 2장에서 가능한 모든 스킬 후보 추출
#       (검증 전 단계이므로 포괄적으로)

{criteria}

# 출력 형식 (JSON):
{
    "stage": "generation",
    "model": "{model_name}",
    "chapter": "2장",
    "candidate_skills": [
        {
            "temp_id":          "cand_001",
            "skill_name_ko":    "...",
            "description":      "...",
            "confidence":       "high|medium|low"
        }
    ]
}

# 입력 내용:
{content}
```

그림 2. Phase 1 생성 단계 프롬프트 템플릿 예시
Fig. 2. Example of phase 1 generation stage (Stage 1) prompt template

Phase 2에서는 각 스킬 당 7개 문제를 생성하되 난이도를 기본(1~3번), 중급(4~5번), 응용(6~7번)에 따라 점진적으로 상향 조정하도록 프롬프트를 구성하였다. 출력 형식은 문제 진술, 정답, 힌트, 스캐폴드를 포함한 JSON 구조를 강제하였다. 힌트는 개념적 안내에서 Bottom-out 힌트까지 다단계로 점진적 구체화를, 스캐폴드는 문제를 7개 하위 단계 질문으로 분해하도록 프롬프트에서 지시하였으며, Dual 파이프라인의 검증자는 수학적 정확성과 함께 힌트 충분성 및 스캐폴드 논리성을 점검하도록 하였다.

Phase 2의 대상 스킬은 Phase 1의 추출 결과로부터 의도적 표집(Purposive sampling)을 통해 선정하였다. 선정 시, Webb의 DOK 프레임워크[18]를 참고하였으며, 개념적 이해, 절차적 계산, 규칙 적용, 복합적 판정 등 인지적 요구 수준이 질적으로 구별되는 4개 유형을 선정하였다. 이는 LLM의 문제 생성 품질이 인지적 요구 유형에 따라 달라지는지를 다각적으로 검증하기 위함이다. 스킬 당 7개 문제는 BKT 기반 숙달도 추적에 필요한 최소 문항 수(5~7개)[2]를 충족하도록 설계하였으며, 총 112개 문제에 대한 336회 평가는 유사 연구[6]의 규모를 웃돈다.

3.3 시스템 구현 및 생성 결과 예시

그림 3은 Claude→GPT 조건에서 생성된 문제와 힌트가 OATutor 시스템에 로딩되어 학습자에게 제시되는 화면 예시이다.

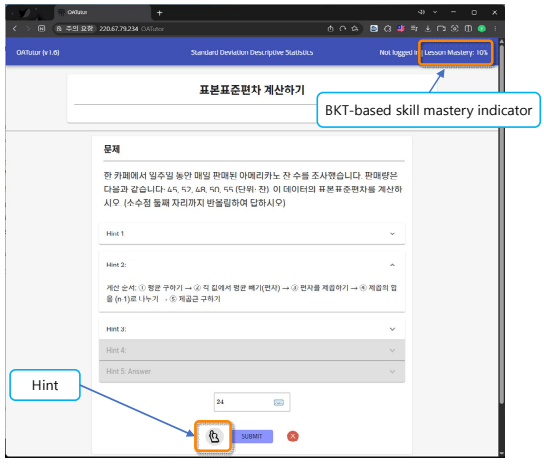


그림 3. 적응형 튜터링 시스템 실행 화면
Fig. 3. Runtime interface of the adaptive tutoring system

학습자가 문제를 풀지 못하면 힌트 버튼을 통해 단계적 안내를 받을 수 있으며, 화면 상단의 BKT 기반 스킬 숙달도(Lesson mastery)가 학습 진행에 따라 실시간으로 갱신된다.

표 1. Phase 2 생성된 문제 목록 예시
Table 1. Example of generated problems

Problem	Difficulty	Problem statement	Answer
P1	Basic	데이터 {2, 4, 4, 6, 9}의 표본표준편차를 구하시오	2.55
P2	Basic	데이터 {-2, 0, 3, 5, 7, 11}의 표본표준편차를 구하시오	4.64
P3	Basic	데이터 {72, 85, 90, 93}의 표본표준편차를 구하시오	9.22
P4	Intermediate	8명의 통학 시간(분) 데이터의 표본표준편차를 구하시오	12.37
P5	Intermediate	두 반의 표본표준편차를 비교하여 산포가 큰 반을 판별하시오	A반
P6	Advanced	평균과 편차 제곱합이 주어진 조건에서 표본표준편차를 역산하시오	4.47
P7	Advanced	새 데이터 추가 시 표본표준편차 변화를 예측하고 계산하시오	2.83


```
{
 "id": "SK1_P1_D1",
 "title": "평균과 중위수의 이상값 민감도 비교하기",
 "lesson": "평균과 중위수의 이상값 민감도 비교하기",
 "courseName": "AI시대의 통계적사고",
 "steps": [
   {
     "id": "SK1_P1_D1a",
     "stepAnswer": ["평균"],
     "problemType": "MultipleChoice",
     "stepTitle": "문제",
     "stepBody": "어느 소규모 회사...(omitted)...은?\nA.
      평균\nB. 중위수",
     "choices": ["평균", "중위수"],
     "hints": {
       "DefaultPathway": [
         {
           "id": "SK1_P1_D1a-h1",
           "type": "hint",
           "dependencies": [],
           "title": "Hint 1",
           "text": "이상값이 있으면 평균은 모든 ...
           (omitted)... 덜 변합니다."
         },
         {
           "id": "SK1_P1_D1a-h2",
           "type": "hint",
           "dependencies": ["SK1_P1_D1a-h1"],
           "title": "Hint 2",
       ...(omitted)...
```


그림 4. Phase 2 문제와 힌트 생성 출력 예시(JSON)
Fig. 4. Example of phase 2 problem and hint generation output (JSON format)

표 1은 최적 파이프라인 Claude→GPT 조건에서 "표본표준편차 계산" 스킬에 대해 생성된 7개 문제의 목록 예이다. 각 문제에는 5단계 힌트와 하위 단계 스캐폴드가 포함된다. 그림 4는 시스템이 출력한 JSON 구조를 보여준다. 문제 진술, 정답과 함께 단계별 스캐폴드(Steps)와 점진적 힌트(Hints)가 OATutor 호환 형식으로 구조화되어, 시스템에 직접 적재 가능하다.

3.4 평가 방법 및 분석 절차

표 2. Phase 2 문제 생성 평가 루브릭 요약
Table 2. Summary of phase 2 Problem generation evaluation rubric

Criterion	5 points	3 points	1 point
Mathema-tical accuracy	Problem, solution, and answer are all correct	1~2 minor errors in solution (answer is correct)	Problem itself is mathematically invalid
Skill alignment	Measures only the target skill accurately	Measures target skill but other skills are mixed	Unrelated to the target skill
Difficulty appropria-teness	Perfectly suitable for introductory college level (DOK 2~3)	High school or intermediate level	Beyond the target learner's capability
Hint quality	3+ progressive hints including Bottom-out hint	1~2 hints with insufficient progression	No hints or hints cause confusion
Scaffold quality	2~4 sub-steps with clear learning objectives each	Sub-steps exist but lack logical structure	No scaffolds provided

추출된 스킬 셋과 생성된 문제, 힌트들은 3명의 통계 전문가에 의해 평가되었다. Phase 1의 스킬 추출 평가는 완전성(Completeness), 적절성(Appropriateness), 세분성(Granularity), 명확성(Clarity), 활용성(Usability)의 5개 기준으로 각 5점, 총 25점 체계를 적용하였으며, Phase 2의 문제 생성 평가 루브릭은 표 2와 같다. 인간 전문가 평가는 Likert 5점 척도 기반의 분석적 루브릭을 사용하

여 수행하였다. 각 평가 차원에 대해 5점(완벽)~1점(부적합)의 구체적 판단 기준을 명시하고, 평가자에게는 블라인드 ID가 부여된 평가 대상과 함께 점수 기재와 판단 근거 서술란이 포함된 구조화된 평가 양식을 배포하였다. 난이도 적절성은 Webb(1997)의 DOK 프레임워크에서 Level 2(기술/개념 적용)~Level 3(전략적 사고)을 대학 통계학 입문 수준의 적합 범위로 설정하여 평가하였다[18]. 실험 규모는 Phase 1에서 15개 조건 × 3명 평가자 = 45회, Phase 2에서 4개 조건 × 4개 스킬 × 7개 문제 × 3명 평가자 = 336회 평가로 구성되었다.

통계 분석에서는 집단 간 평균 차이 검정을 위해 일원배치 분산분석(One-way ANOVA) 또는 비모수적 대안인 Kruskal–Wallis H 검정을 시행하였으며, 평가자 간 신뢰도는 급내상관계수(ICC, Intraclass Correlation Coefficient)를 산출하여 평가 일치도를 검증하였다.

Ⅳ. 연구 결과

4.1 Phase 1: 스킬 추출 실험 결과

Phase 1에서는 15개 LLM 파이프라인 조합의 스킬 추출 품질을 비교하였다. 인간 전문가 3인의 블라인드 평가 결과(만점 25점), (GPT→Claude) 파이프라인이 평균 22.67점(SD=3.21)으로 최고 성능을 기록하였으며, GPT 단독 조건이 22.00점(SD=2.65)으로 2위를 차지하였다. 반면, Gemini가 생성자 역할을 수행한 조합들은 하위권에 분포하여 (Gemini→GPT→Claude)가 17.33점으로 최저 성능을 보였다.

파이프라인 유형별 분석에서 일원배치 분산분석을 실시한 결과, 표 3과 같이 Single(M=21.33), Dual(M=21.06), Triple(M=19.56) 유형 간 평균 차이는 통계적으로 유의하지 않았다($F=2.079$, $p=0.138$). 그러나 효과 크기($\eta^2=0.090$)는 중간 수준으로, 현재 표본 크기에서는 통계적 유의성이 확보되지 않았으나 표본을 확대하면 유의한 차이가 검출될 가능성을 시사한다. 특히 Single과 Triple 간 비교에서 Cohen's d=0.69로 중간에서 큰 효과 크기가 관찰되어, Triple 파이프라인의 복잡성 증가가 오히려 성능 저하를 유발할 수 있음을 보여주었다.

표 3. ANOVA: 파이프라인 유형별
Table 3. ANOVA: Pipeline type

Source	SS	df	MS	F	p	η^2
Between	27.56	2	13.78	2.079	0.138	0.09
Within	278.67	42	6.63			
Total	306.22	44				

생성자(Generator) 모델별 분석에서는 표 4, 그림 5와 같이 유의한 차이가 확인되었다($F(2, 42)=4.302$, $p=0.020$, $\eta^2=0.170$). Bonferroni 보정을 적용한 사후분석 결과, GPT(M=21.67)가 Gemini(M=19.07) 대비 유의하게 우수한 성능을 보였으며($t=2.791$, $p=0.009$, Cohen's d=1.02), GPT와 Claude(M=20.80) 간 차이는 유의하지 않았다($p=0.354$). 이는 스킬 생성 단계에서 GPT가 풍부한 후보 스킬을 생성하는 능력이 우수함을 의미한다.

평가자 간 신뢰도 검증을 위해 수행된 급내 상관계수(ICC) 분석 결과, ICC(3, k) 값은 0.881(95% CI: 0.72~0.96)로 Koo와 Li[19]의 기준에 따라 'Good Reliability' 수준에 해당하였으며, 이는 3인 평가자의 판단이 통계적으로 일관되었음을 보여준다($F(14, 28)=8.435$, $p<0.001$).

표 4. ANOVA: 생성자 모델별
Table 4. ANOVA: Generator model

Source	SS	df	MS	F	p	η^2
Between	52.04	2	26.02	4.302	.020*	0.17
Within	254.13	42	6.05			
Total	306.18	44				

*p<.05

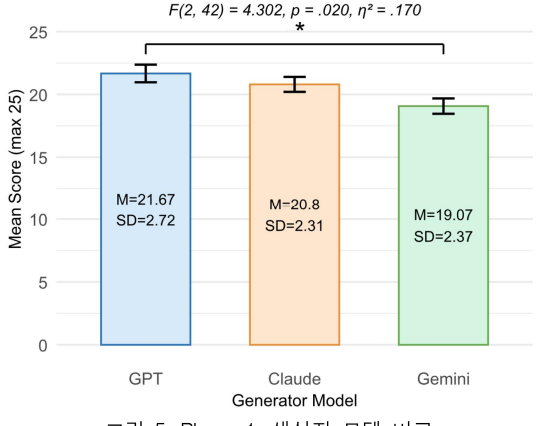


그림 5. Phase 1: 생성자 모델 비교
Fig. 5. Phase 1: Generator model comparison

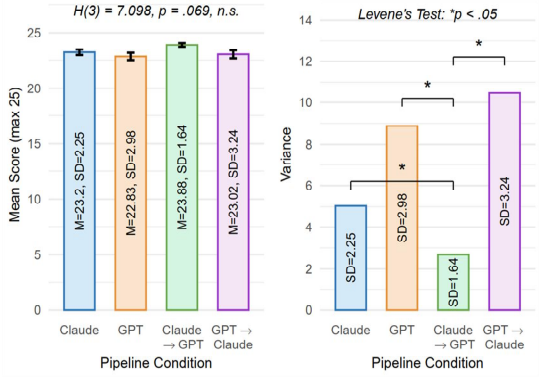
4.2 Phase 2: 문제 생성 실험 결과

Phase 2에서는 Phase 1의 최적 스킬 목록을 기반으로 4개 LLM 조합의 문제 생성 품질을 평가하였다(25점 만점). 총 112개 문제에 대한 336회 평가 결과, (Claude→GPT) 파이프라인이 평균 23.88점(SD=1.64)으로 최고 성능을 기록하였다. (Claude 단독)은 23.20점(SD=2.25), (GPT→Claude)는 23.02점(SD=3.24), (GPT 단독)은 22.83점(SD=2.98)을 기록하였다.

Kruskal-Wallis 검정 결과, 표 5, 그림 6(a)와 같이 4개 조건 간 평균 차이는 통계적으로 유의하지 않았다(p=0.069). 이는 최신 LLM들의 문제 생성 능력이 높은 수준으로 상향 평준화되었음을 시사한다. 그러나 품질 안정성 측면에서는 표 6, 그림 6(b)와 같이 조건 간 차이가 확인되었다. Levene 검정(Brown-Forsythe) 결과, (Claude→GPT) 조건의 분산은 (Claude 단독)(W=5.136, p=0.025), (GPT 단독)(W=4.561, p=0.034), (GPT→Claude)(W=4.121, p=0.044) 대비 모두 유의하게 작았다. 이는 (Claude→GPT) 파이프라인이 평균 점수뿐 아니라 품질 균일성 측면에서도 우수함을 보여준다.

표 5. Kruskal-Wallis 검정
Table 5. Kruskal-Wallis Test

H	df	p
7.098	3	0.069



(a) (b)

그림 6. Phase 2: 파이프라인 유형별 비교
(a) 평균 비교 (b) 분산 비교
Fig. 6. Phase 2: Comparison by pipeline condition
(a) Mean comparison (b) Variance comparison

표 6. Levene 검정 (Brown-Forsythe)
Table 6. Levene's Test (Brown-Forsythe)

Comparison	W	p
Claude→GPT vs Claude (only)	5.136	.025*
Claude→GPT vs GPT (only)	4.561	.034*
Claude→GPT vs GPT→Claude	4.121	.044*

*p<.05

평가자 간 신뢰도는 ICC(3, k)=0.934(95% CI: 0.91~0.95)로 'Excellent Reliability' 수준에 해당하였다(F(110, 220)=15.188, p<0.001).

V. 결론 및 향후 과제

본 연구는 Multi-LLM 파이프라인을 활용하여 적응형 튜터링 시스템의 교육 콘텐츠를 자동으로 생성하는 방법론을 제안하고, 인간 전문가 평가를 통해 그 효과를 검증하였다. 연구는 교재로부터 학습 스킬을 자동 추출하는 Phase 1(Skill extraction)과, 추출된 스킬을 기반으로 연습 문제를 생성하는 Phase 2(Problem generation)로 구성된 2단계 End-to-End 파이프라인으로 설계되었다.

콘텐츠 품질 평가 결과, Phase 1에서는 생성자 역할을 수행한 GPT가 Gemini 대비 통계적으로 유의하게 우수한 성능을 보였으며(p<0.05), GPT→Claude 파이프라인이 25점 만점 기준 평균 22.67점(SD = 3.21)으로 가장 높은 성능을 기록하였다. Phase 2에서 파이프라인 4개 조건 간 평균 차이가 통계적으로 유의하지 않았으나(H(3)=7.098, p=0.069), Levene 검정 결과 Claude→GPT 파이프라인의 분산(SD=1.64)이 다른 모든 조건 대비 유의하게 작았다(p<0.05). 따라서 해당 조합의 강점은 평균 점수의 우위보다 품질 균일성에 있다. 인간 전문가 3인의 평가에 대한 신뢰도 분석 결과는 ICC(3, k) 값이 Phase 1에서 0.881로 '우수(Good)' 수준, Phase 2에서 0.934로 '매우 우수(Excellent)' 수준으로 나타나, 본 연구의 평가 결과가 충분한 객관성과 일관성을 확보하고 있음을 확인하였다.

본 연구 결과, Multi-LLM 파이프라인의 단순한 단계 확장은 품질 향상을 보장하지 않았으나(Phase 1: p=0.138), 이종 모델 간 생성-검증 역할 분담은 품질 안정성 측면에서 유의한 효과를 보였다

(Levene's test, $p<0.05$). 이는 Multi-LLM의 효과가 단계 수가 아닌 모델 간 역할 배치 때문에 결정됨을 시사하며, 이종 LLM 간 역할 분담에 기반한 생성-검증 구조의 효과를 실증적으로 보여준다. 나아가, 적응형 학습 플랫폼(OATutor)에 즉시 적용 가능한 한국어 통계교육 콘텐츠 자동 생성 파이프라인을 제시하였다는 점에서 학술적·실무적 의의를 지닌다.

본 연구는 단일 교재의 한 개 챕터를 대상으로 하였으며, Phase 1에서 조건당 평가 수와 Phase 2의 스킬 수가 제한적이어서 통계적 검정력에 한계가 있다. 이는 Multi-LLM 파이프라인의 역할 분담 구조를 탐색하는 초기 연구(Exploratory study)로서 방법론 제안과 실증적 가능성 확인에 초점을 둔 데 기인한다. 또한 생성자 모델 간 성능 차이가 통계학 도메인의 단일 교재에서 확인된 결과이므로, 다른 도메인으로의 일반화에는 후속 검증이 필요하다. 그러나 본 연구의 파이프라인은 "포괄적으로 추출하라"와 같이 역할 기반(Role-based) 프롬프트로 설계되어 통계학 고유의 지식이 프롬프트에 내장되어 있지 않으며, 생성-검증 역할 분담의 효과는 특정 도메인이 아닌 LLM 간 협업의 구조적 특성에서 비롯된 것이다. 아울러 Phase 2에서 평가한 4개 스킬이 개념형, 계산형, 규칙형, 판정형을 포괄하고 있어, 단일 챕터 내에서도 다양한 스킬 유형에 걸쳐 파이프라인의 일관된 성능이 확인되었다. 이러한 점에서 입력 교재를 교체하여 다른 교과 영역에 적용할 수 있는 구조적 확장 가능성을 갖는다.

향후 연구에서는 동일 교재의 다른 장이나 수학, 프로그래밍 등 다양한 교과로의 확장을 통해 일반화 가능성을 검증하고, Pre-Post Test 및 A/B Test를 활용한 학습 효과 검증, 그리고 오픈소스 LLM 적용을 통한 비용 효율적 파이프라인 설계 등을 주요 연구 과제로 제안한다.

References

[1] J. A. Kulik and J. D. Fletcher, "Effectiveness of intelligent tutoring systems: A meta-analytic review", Review of Educational Research, Vol. 86, No. 1, pp. 42-78, Mar. 2016. https://doi.org/10.3102/0034654315581420.

[2] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge", User Modeling and User-Adapted Interaction, Vol. 4, No. 4, pp. 253-278, Dec. 1994. https://doi.org/10.1007/BF01099821.

[3] Z. A. Pardos, M. Tang, I. Anastasopoulos, S. K. Sheel, and E. Zhang, "OATutor: An open-source adaptive tutoring system and curated content library for learning sciences research", Proc. 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, pp. 1-16, Apr. 2023. https://doi.org/10.1145/3544548.3581574.

[4] T. Murray, "An overview of intelligent tutoring system authoring tools: Updated analysis of the state of the art", Authoring Tools for Advanced Technology Learning Environments, pp. 491-544, 2003. https://doi.org/10.1007/978-94-017-0819-7_17.

[5] Z. A. Pardos and S. Bhandari, "ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills", PLOS One, Vol. 19, No. 5, e0304013, May 2024. https://doi.org/10.1371/journal.pone.0304013.

[6] S. M. Lim, H. W. Cho, J. W. Lee, and H. S. Lee, "Exploring the potential of large language models for automatic item generation in social studies", Journal of Educational Information and Media, Vol. 30, No. 3, pp. 1035-1060, Sep. 2024.

[7] M. H. Kim, Y. J. Lee, N. Y. Seo, H. Y. Cheon, and D. I. Jeon, "A proposal for a grammar question generation system for the Korean CSAT language section using LLM", Journal of Edutainment, Vol. 7, No. 1, pp. 311-327, Mar. 2025. https://doi.org/10.36237/koedus.7.1.311.

[8] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, and P. Fung, "Survey of hallucination in natural language generation", ACM Computing Surveys, Vol. 55, No. 12, pp. 1-38, Dec. 2023. https://doi.org/10.1145/3571730.

[9] H. S. Nwana, "Intelligent tutoring systems: An overview", Artificial Intelligence Review, Vol. 4,

No. 4, pp. 251-277, Dec. 1990. https://doi.org/10.1007/BF00168958.

[10] K. R. Koedinger, A. T. Corbett, and C. Perfetti, "The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning", Cognitive Science, Vol. 36, No. 5, pp. 757-798, Jul. 2012. https://doi.org/10.1111/j.1551-6709.2012.01245.x.

[11] M. J. Gierl and T. M. Haladyna, "Automatic Item Generation: Theory and Practice", Routledge, 2012.

[12] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes", International Journal of Artificial Intelligence in Education, Vol. 30, No. 1, pp. 121-204, Mar. 2020. https://doi.org/10.1007/s40593-019-00186-y.

[13] Z. Wang, J. Valdez, D. B. Mallick, and R. G. Baraniuk, "Towards human-like educational question generation with large language models", Proc. 23rd International Conference on Artificial Intelligence in Education, Durham, UK, pp. 153-166, Jul. 2022. https://doi.org/10.1007/978-3-031-11644-5_13.

[14] S. H. Son, S. H. Lee, W. D. Jeong, and J. H. Cho, "Development of a personalized learning support system using LLM", Intelligence, Information, Convergence and Future Education, Vol. 4, No. 7, pp. 1-8, 2025.

[15] L. Zheng, W. L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, and I. Stoica, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena", Advances in Neural Information Processing Systems, New Orleans, USA, Vol. 36, pp. 46595-46623, Dec. 2023.

[16] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, and P. Clark, "Self-refine: Iterative refinement with self-feedback", Advances in Neural Information Processing Systems, New Orleans, USA, Vol. 36, pp. 46534-46594, Dec. 2023.

[17] Y. S. Choi, "Understanding Statistics with R", Bigbook, 2014.

[18] N. L. Webb, "Criteria for alignment of expectations and assessments in mathematics and science education", Research Monograph No. 6, University of Wisconsin-Madison, Wisconsin Center for Education Research, 1997.

[19] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research", Journal of Chiropractic Medicine, Vol. 15, No. 2, pp. 155-163, Jun. 2016.

저자소개

장 필 식 (Phil-Sik Jang)

1990년 2월 : 서울대학교 조선공학과(공학사)
1992년 2월 : KAIST 산업공학과(공학석사)
1998년 8월 : KAIST 산업공학과(공학박사)
1997년 9월 ~ 현재 : 세한대학교 항공물류학과 교수
관심분야 : HCI, 빅데이터 분석, 알고리즘 트레이딩

이 주 양 (Ju-Yang Lee)

1993년 2월 : 대전대학교 가정관리학과(이학사)
2013년 8월 : 세종대학교 호텔관광경영학과(호텔관광경영학석사)
2016년 8월 : 순천향대학교 관광경영학과(관광경영학박사)
2023년 7월 ~ 현재 : 순천향대학교 SCH 경제경영연구소 연구 부교수
관심분야 : 관광경영, 항공서비스, 고객만족경영