

# LLM-Assisted Iterative Topic Modeling with Multi-dimensional Metadata and Human-in-the-Loop Validation: A Case Study of Metaverse Education

Wan-Je Gil\*, In-Soo Shin\*

\*Professor, Dept. of AI Convergence Education, Graduate School of Education, Dongguk University, Seoul, Korea

## [Abstract]

This study proposes an LLM-assisted iterative topic modeling framework that integrates multi-dimensional metadata structuring, multi-model consolidation, and human-in-the-loop validation. Research subjects, fields, and methods were extracted from academic documents, and LDA and BERTopic were applied in parallel to construct an initial topic set aligned with prior topic structures. The topic system was refined through selective cross-validation with heterogeneous LLMs, progressive sampling, and expert review. The finalized topic set was applied to the full corpus for distribution and trend analysis. The results show that metaverse education research has expanded around stable core themes rather than abrupt paradigm shifts. This study reconceptualizes topic modeling as an iterative process of semantic convergence and offers a structured empirical classification of metaverse education research.

▶ **Key words:** Topic Modeling, Large Language Models, Metadata Structuring, Multi-Model Integration, Metaverse Education

## [요 약]

본 연구는 다차원 메타데이터 구조화, 다중 모델 통합, 휴먼 인 더 루프 검증을 결합한 LLM 보조 반복적 토픽모델링 프레임워크를 제안한다. 학술 문헌에서 연구대상, 연구분야, 연구방법을 추출한 뒤 LDA와 BERTopic을 병렬 적용하여 기존 메타버스 교육 토픽모델링 연구와 정렬된 초기 토픽 집합을 구성하였다. 이후 이중 LLM 기반 선택적 교차검증, 단계적 표본 확대, 전문가 검토를 통해 토픽 체계를 정교화하였다. 최종 토픽 집합을 전체 코퍼스에 적용한 결과, 메타버스 교육 연구는 급격한 전환보다 안정적 핵심 주제를 중심으로 확장된 것으로 나타났다. 본 연구는 토픽모델링을 반복적 의미 수렴 과정으로 재개념화하고, 메타버스 교육 연구의 구조화된 실증 분류 체계를 제시한다.

▶ **주제어:** 토픽모델링, 대형언어모델, 메타데이터 구조화, 다중모델 통합, 메타버스 교육

- First Author: Wan-Je Gil, Corresponding Author: In-Soo Shin
- Wan-Je Gil (jaygil8755@gmail.com), Dept. of AI Convergence Education, Graduate School of Education, Dongguk University
- In-Soo Shin (9065031@hanmail.net), Dept. of AI Convergence Education, Graduate School of Education, Dongguk University
- Received: 2026. 03. 13, Revised: 2026. 05. 22, Accepted: 2026. 05. 26.
- This paper is a revised and supplemented version of the one presented at The Korea Society of Computer and Information Winter Conference 2026.

## I. Introduction

최근 텍스트 마이닝 분야는 전통적 확률 기반 모형에서 임베딩 기반 및 대형언어모델(LLM)을 활용한 의미 중심 분석을 포함하는 방향으로 확장되고 있다[1-5]. 잠재 디리클레 할당(LDA, Latent Dirichlet Allocation)으로 대표되는 기존 토픽모델링은 문서-단어 동시출현(co-occurrence) 구조에 기반하여 대규모 텍스트의 잠재 주제를 추출하는 데 기여해 왔으나[1], 문맥적 의미를 충분히 반영하지 못하고 해석 가능성이 제한된다는 비판을 받아왔다[2][3]. 특히 단어 빈도 분포에 의존하는 접근은 연구 설계 맥락, 개념적 층위, 서지 메타정보와 같은 다차원적 의미 구조를 직접적으로 통합하는 데 한계를 가진다[2][4]. 이러한 문제의식 속에서 최근 연구들은 임베딩 기반 군집화, 프롬프트 기반 토픽 생성, 통계 모형과 LLM의 하이브리드 결합 등 다양한 방법론을 제안하며 토픽모델링의 확장을 시도하고 있다[2-5].

LLM은 문장의 의미를 맥락적으로 파악하고 응집력 있는 주제 설명을 생성할 수 있다는 점에서 기존 Bag-of-Words 기반 접근을 보완한다[2][5]. 그러나 다수의 연구가 여전히 단일 모델 중심의 생성 결과에 의존하거나, 직접적인 LLM 적용에서 불완전한 토픽 포괄성, 토픽 정렬 문제, 비효율성 등의 한계를 보고하고 있다는 점에서, 통계적 안정성과 해석적 적합성을 동시에 확보하는 통합 설계가 여전히 중요한 과제로 남아 있다[4][5].

이에 본 연구는 토픽모델링을 단순한 군집 산출 과정이 아니라, 구조화된 메타데이터 생성, 다중 모델 통합, 이중 LLM 교차검증, 그리고 인간 전문가 검증을 통해 반복적으로 정교화되는 의미 수렴 과정으로 재구성하고자 한다. 특히 문헌의 원문을 직접 투입하는 대신, 연구대상, 연구분야, 연구방법과 같은 다차원 메타정보를 LLM을 통해 구조화하고, 이를 기반으로 LDA와 BERTopic을 병렬 적용한 후 기존 연구[8-12]의 토픽 구조와 통합하는 프레임워크를 설계하였다. 이후 선택적 교차검증과 단계적 표본 확대를 통해 토픽 체계를 반복적으로 정교화함으로써, 재현성과 해석 가능성을 동시에 확보하고자 하였다.

이러한 방법론의 적용 가능성과 타당성을 검증하기 위하여 본 연구는 메타버스 교육 분야를 사례 연구 대상으로 설정하였다. 메타버스 교육은 기술 적용, 교수-학습 설계, 학습 효과, 도입 요인과 제약 등 다양한 연구 주제가 교차하는 복합적 영역으로 보고되어 왔으며[6][7], 토픽 구조의 다층성과 변화를 분석하기에 적합한 사례이다. 따라서 본 연구는 메타버스 교육을 하나의 사례로 활용하여, 다차원

메타데이터와 휴먼 인 더 루프 검증을 결합한 LLM 보조 반복적 토픽모델링 프레임의 방법론적 의의와 확장 가능성을 실증적으로 제시하고자 한다.

## II. Preliminaries

### 1. Related works

#### 1.1 Metaverse Education Topic Modeling: A Review of Prior Studies

메타버스 교육 분야의 연구 동향을 체계적으로 파악하기 위하여 토픽모델링을 적용한 기존 연구는 국내외에서 지속적으로 축적되어 왔다[8-12]. 이 중 대표적인 다섯 편의 연구를 중심으로 연구 범위, 방법론, 주요 결과를 종합적으로 검토한다.

국내 연구로는 Lim et al.(2022)가 2010년부터 2021년까지 발표된 메타버스 및 가상세계 활용 교육 관련 논문 418편을 대상으로 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 기반 토픽모델링과 시계열 회귀분석을 병행하여 수행한 연구가 있다[8]. 해당 연구는 가상환경 기반 교육이 초기의 기술 중심 논의를 넘어 ‘사회적 변화 및 문화언어교육 활용’, ‘안전교육훈련’ 등으로 주제가 확장되고 있음을 실증적으로 제시하였다[8]. 또한 Han and Kim(2023)은 2020년부터 2022년까지의 국내 메타버스 교육 연구를 분석하여, 메타버스 기반 교육 연구가 고등교육 뿐 아니라 초등학교, 복지·종교, 간호·예술·언어 분야 등으로 확장되고 있음을 보고하였다[9].

해외 문헌을 대상으로 한 연구 역시 대규모 데이터에 기반한 동향 분석을 시도하였다. Kim(2023)는 2013년부터 2022년까지 해외 학술지에 게재된 가상현실 교육 연구 1,986편을 분석하여, 예비 의료인 교육과 3차원 환경에서의 공간-운동 능력 향상 관련 주제가 주요 토픽으로 부상하고 있음을 밝혔다[10]. Kim and Im(2022)은 1992년부터 2022년까지 30년간 축적된 6,755편의 가상현실 기반 교육 연구를 대상으로 LDA 기반 토픽모델링을 수행하여, 연구의 전개 과정을 네 시기로 구분하여 제시하였다[11]. 한편, Ozyurt and Ozyurt(2025)는 2003년부터 2022년까지 발행된 교육용 증강현실(AR) 관련 문헌 3,718편을 대상으로 LDA 기반 토픽모델링과 서술적 분석을 수행하여, 문화유산 교육, 의학교육 및 환자 돌봄, 식품 소비 안전 등 11개의 주요 연구 주제를 도출하고 그 변화 양상을 분석하였다[12]. 기존 연구[8-12]들의 분석 대상 데이터 수집 방법에 대해서는 Table 1에서 정리하였다.

Table 1. A Summary of Prior Studies - Metaverse Education Topic Modeling

Author	Search Terms	Data Collection Period	Final Number of Articles	Databases
Lim et al.(2022)	(Metaverse, Virtual World, Virtual Reality) + (Teaching/Professor, Classroom, Education, Class, School, Student, Learning)	2010~2021	418	RISS, DBpia, KCI, e-article
Han & Kim (2023)	(Metaverse, Virtual World, etc.) + (Education, Learning, etc.)	2020~2022	365	RISS
Kim (2023)	(Virtual Reality, VR) + (Learning, Education)	2013~2022	1,986	Web of Science
Ozyurt and Ozyurt (2025)	(augmented reality) + (education, teaching, learning etc.)	2003~2022	3,718	Scopus
Kim & Im(2022)	(virtual reality, augmented reality, mixed reality, extended reality, metaverse) + (education, training, learning etc.)	1992~2022	6,755	Web of Science

참고로 Search Terms는 원문에 실린 단어 그대로 인 용하였다. 이들 기존 연구[8-12]는 방대한 학술 텍스트를 정량적으로 분석하여 핵심 주제를 도출하고, 메타버스 및 확장현실 기반 교육의 활용 가능성과 연구 확산 양상을 규명하였다는 점에서 중요한 학문적 기여를 지닌다. 그러나 대부분의 연구가 전통적 확률 기반 토픽모델링 기법인 LDA 또는 LDA 중심 분석에 주로 의존함으로써, 문헌의 맥락적 의미 구조나 서지 메타정보(연구대상, 연구분야, 연구방법 등)와의 유기적 연계를 충분히 반영하지 못하였다는 한계를 보인다. 즉, 연구 주제의 분포와 변화 양상을 제시하는 데에는 성공하였으나, 토픽 간 의미적 관계의 정교한 재구성이나 반복적 해석 과정을 통한 심층 분석으로는 충분히 확장되지 못하였다.

### 1.2 LLM-Enhanced Topic Modeling - A Novel Topic Modeling Approach

최근 텍스트 마이닝 연구에서는 단어의 동시 등장 빈도에 의존하여 문맥을 충분히 반영하지 못하는 전통적 잠재 디리클레 할당(LDA)의 한계를 보완하기 위해, 대형언어모델(LLM)을 결합한 다양한 방법론이 제안되고 있다. LITA (LLM-assisted Iterative Topic Augmentation)는 대규모 코퍼스에서 LLM 활용 토픽모델링의 비용과 확장성 문제를 해결하기 위해 제안된 프레임워크[4], 사용자 제공 seed 단어와 문서 임베딩 기반 군집화를 바탕으로 초기 토픽 구조를 형성한 뒤, 군집 경계에 위치한 모호한 문서에 한해서만 LLM을 선택적으로 호출하여 토픽 재할당 또는 신규 토픽 생성을 수행하는 반복적 정제 구조를 갖는다. 이 접근은 모든 문서에 대해 LLM을 호출하지 않으면서도 토픽 coherence와 군집 품질을 개선할 수 있다는 점에서 효율적 LLM 기반 토픽정제의 대표 사례로 볼 수 있다. LLM-ITL(LLM-Integrated Topic Learning)은 Neural Topic Model(NTM)과 LLM을 결합하여 토픽 해

석력과 문서 표현 품질을 동시에 개선하고자 한 접근으로, NTM이 학습한 토픽 분포와 LLM이 제안한 토픽 단어 집합을 Optimal Transport 기반 목적함수로 정렬하고, LLM 제안의 신뢰도를 가중치로 반영함으로써 노이즈 영향을 줄이도록 설계되었다[5]. 이는 LLM의 의미 생성 능력과 NTM의 분포 학습 능력을 상호보완적으로 결합한 모델 중심 접근이라는 특징을 가진다. LiSA(LLM-Guided Semantic-Aware Clustering)는 문서 의미공간과 토픽 의미공간 간의 불일치가 토픽모델 성능 저하의 핵심 원인이라는 문제의식에서 출발하여, LLM으로 후보 토픽 단어와 설명을 생성한 후 이를 클러스터링하여 토픽 수준 의미 공간을 구성하고, 문서 임베딩 군집과의 불일치를 탐지한 뒤, 로컬 점검과 협력적 개선을 통해 두 의미공간을 정렬하는 방식을 제안하였다[13]. 이 방법은 단순 토픽 생성이 아니라 문서-토픽 정렬 자체를 체계적으로 다룬다는 점에서 의미공간 정렬 중심의 접근으로 이해할 수 있다. 한편, 순위기반 통합(rank-based aggregation) 연구는 서로 다른 토픽모델 또는 LLM 기반 추출 결과 간의 불일치를 줄이기 위한 앙상블 전략으로서, 다양한 모델이 생성한 토픽 후보를 단순 평균이 아니라 순위 기반으로 집계하여 출력 편향과 변동성을 완화하고, 토픽의 안정성과 해석가능성을 높이고자 하였다[14]. 특히 이 접근은 AI 생성 텍스트와 인간 작성 텍스트가 혼합된 데이터셋에서도 일관된 토픽 구조를 유지하는 데 효과적인 방법으로 제시되었다.

이에 비해 본 연구는 특정 토픽모델의 내부 성능 향상이나 토픽 후보의 앙상블 자체를 목표로 하지 않는다. 본 연구는 LITA와 같이 선택적 LLM 호출을 활용하지만, 그 목적은 임베딩 군집 품질 개선이나 신규 토픽 확장이 아니라, LDA, BERTopic, 기존 연구[8-12] 토픽, 구조화 메타 데이터 사이에서 발생하는 불일치를 조정하고 최종 토픽 할당의 합의를 수렴시키는 데 있다. 또한 LLM-ITL이 NTM 내부의 분포 정렬을 수행하는 모델 중심 접근이라

면, 본 연구는 서로 다른 생성 원리를 가진 토픽 산출물들을 기존 연구[8-12] 기반 상위 의미를 안에서 재배열하는 “토픽 체계 간 정렬”을 수행한다는 점에서 문제 설정이 다르다. LiSA와 같이 불일치 지점을 중심으로 개입한다는 점에서는 문제의식을 공유하지만, LiSA가 문서-토픽 의미 공간 정렬 자체를 핵심 목표로 하는 반면, 본 연구는 기존 연구[8-12] 토픽 체계를 기준으로 비교 가능한 분류 구조를 확정하는 데 초점을 둔다. 특히 본 연구에서는 연구대상, 연구분야, 연구방법과 같은 구조화된 메타데이터가 토픽 할당 판단의 핵심 근거로 작동하며, 이는 임베딩 기반 의미공간 정렬 중심의 접근과 구별되는 지점이다. 마지막으로 rank-based aggregation이 토픽 후보 수준의 집계를 통해 안정성을 확보하는 전략이라면, 본 연구는 확정된 통합 토픽 체계 내에서 문헌 단위의 최종 할당을 반복적 검토와 선택적 교차검증을 통해 확정한다는 점에서 접근 수준과 분석 단위가 다르다. 요컨대, 본 연구의 차별성은 LLM을 토픽 생성의 대체 수단으로 사용하는 것이 아니라, 다중 근거를 연결하고 불일치를 조정하는 검증-정렬-합의 수렴의 도구로 제한적으로 배치한다는 데 있다.

### III. The Proposed Scheme

본 연구의 전체 연구 절차는 Fig. 3에 제시하였다. Fig. 3은 메타버스 교육 문헌의 수집 및 전처리 단계에서부터 LLM 기반 구조화 메타데이터 생성, LDA와 BERTopic의 병렬 적용, 기존 연구[8-12] 기반 토픽 통합, 이종 LLM을 활용한 반복적 교차검증, 최종 토픽 집합 확정 및 전체 코퍼스 적용에 이르는 전 과정을 통합적으로 보여준다. 본 연구의 핵심은 단일 토픽모델의 산출 결과를 그대로 수용하는 것이 아니라, 구조화된 메타데이터, 전통적 확률 기반 토픽모델, 임베딩 기반 토픽모델, 기존연구의 토픽 체계, 그리고 인간 전문가 검토를 반복적으로 결합하여 최종 토픽 체계를 안정화하는 데 있다. 이후 각 절에서는 Fig. 3에 제시된 절차를 단계별로 상세히 설명한다.

#### 1. Data Collection and Preprocessing

본 연구는 메타버스 교육 연구의 동향을 체계적으로 분석하기 위하여 국내 주요 학술 데이터베이스인 RISS와 KCI를 활용하여 학술 문헌을 수집하였다. 자료 수집은 2025년 11월 초에 수행되었으며, 연도에 대한 제한은 두지 않아 연구 초기 단계부터 최근까지의 전체 연구 흐름을 포괄하도록 하였다. 본 연구에서 수집 및 전처리한 데이터

셋의 전체 구성은 Table 2에 요약하였다.

Table 2. Dataset Schema and Column Definitions

Category	Column Name	Description
Bibliographic Information (RISS/KCI)	Title	Standardized title of the paper or thesis
	Author	Name(s) of the author(s)
	Publication Year	Year of publication for time-series analysis
	Source / Affiliation	Journal name or university for degree papers
Structured Metadata LLM Extraction (GPT-5.2)	Abstract	Standardized abstract text in Korean
	Research Subject	Categorized target groups (e.g., Learners, Teachers)
	Research Field	Applied educational domains or technologies
	Research Method	Primary methodology used in the research
	Evidence Phrases	Key phrases from the abstract used as classification basis

초기 수집 단계에서 RISS에서는 학술지 논문 1,177편과 학위논문 595편이 검색되었으며, KCI에서는 학술지 논문 494편이 수집되었다. 이에 따라 초기 원자료는 총 2,266편으로 구성되었다. RISS 자료는 파이썬 기반 크롤링 기법을 활용하여 수집하였으며, 검색 결과 페이지와 상세 페이지의 서지 정보를 구조화하여 정리하였다. KCI 자료는 서지 정보 내보내기 기능을 활용하여 표준화된 형식으로 추출하였다. 학술지 논문의 경우 제목, 저자, 발행연도, 학회명, 저널명, 초록 정보를 수집하였으며, 학위논문의 경우 저널명 대신 학교명과 학위 구분(석사/박사)을 포함하였다. 이후 모든 문헌을 하나의 통합 데이터셋으로 병합하였다.

병합된 데이터셋에 대해서는 다단계 전처리와 적합성 필터링을 수행하였다. 먼저 제목은 대소문자 차이, 불필요한 공백, 특수문자를 제거하여 표준화한 뒤 비교하였고, 표준화된 제목과 저자 정보가 동일한 경우 중복 문헌으로 판단하였다. 필요 시 발행연도 및 학술지 정보를 보조 기준으로 활용하여 중복 여부를 재확인하였다. 또한 연구 주제의 적합성을 확보하기 위하여 명확한 포함 기준을 적용하였다. 첫째, 제목에 “메타버스”, “Metaverse”, “가상세계” 등 메타버스 관련 핵심어가 최소 하나 이상 포함되어야 한다. 둘째, 제목 또는 초록에 “교육”, “학습”, “학교”, “learning”, “teaching” 등 교육 관련 핵심어가 최소 하나 이상 포함되어야 한다. 이 두 조건을 모두 충족하지 않는 문헌은 분석 대상에서 제외하였다. 이 과정을 거쳐 중복 문헌과 주제 관련성이 낮은 문헌을 제외한 결과, 1차 분석 대상은 1,293편으로 정리되었다. 이후 텍스트 마이닝

의 신뢰성을 확보하기 위하여 초록이 제공되지 않은 문헌 119편을 추가로 제외하였고, 최종 분석 대상은 1,174편으로 확정하였다. 초록이 한국어로 제공된 경우에는 그대로 사용하였으며, 한국어와 영문 초록이 함께 제공된 경우에는 한국어 초록만을 활용하였다. 영문 초록만 제공된 경우에는 파이썬 기반 자동 번역 도구를 활용하여 한국어로 번역한 후 분석에 포함하였다. 이를 통해 전체 코퍼스의 언어적 일관성을 유지하였으며, 최종적으로 확정된 문헌 집합을 본 연구의 LLM 보조 반복적 토픽모델링 분석의 기초 데이터로 활용하였다.

이와 같은 데이터 수집 및 전처리 과정은 Fig. 1에 순서대로 제시하였다.

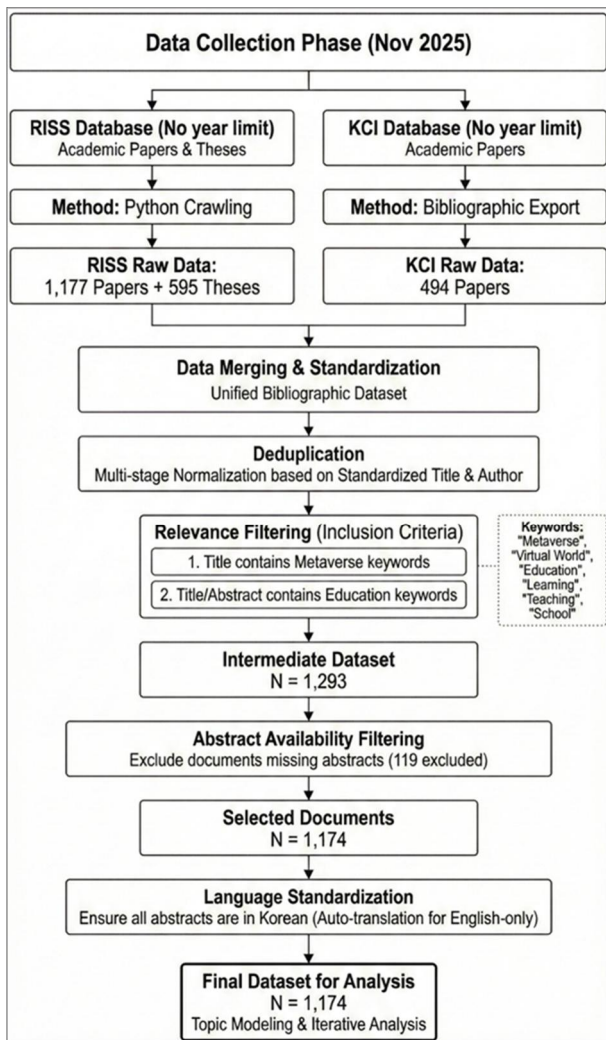


Fig. 1. Data Collection and Preprocessing Flow

## 2. LLM-Based Metadata Structuring Grounded in Prior Research

본 연구는 메타버스 교육 연구의 주제 구조를 정밀하게 도출하기 위하여 원문 텍스트 전체를 직접 토픽모델링에

투입하는 방식 대신, 기존 메타버스 교육 토픽모델링 연구 [8-12]에서 반복적으로 등장한 주제 범주를 기반으로 구조화된 메타데이터를 생성하는 절차를 설계하였다. 이를 위해 메타버스 교육 분야의 기존 토픽모델링 연구들을 분석하여 각 연구에서 제시된 토픽명과 핵심 키워드를 비교·정리하였으며, 반복적으로 출현하는 개념을 연구대상, 연구분야, 연구방법의 세 차원으로 재구성하였다. 구체적으로, 기존 메타버스 교육 토픽모델링 연구[8-12]에서 제시된 토픽명과 핵심 키워드를 수집한 뒤, 이를 교육 연구의 분석 구조에 맞추어 연구대상(Who), 연구분야(What), 연구방법(How)의 세 범주로 재분류하였다. 본 연구에서 연구대상, 연구분야, 연구방법의 세 차원을 설정한 이유는 메타버스 교육 연구가 기술 자체에 대한 논의라기보다 교육적 적용 맥락 속에서 수행되는 연구이기 때문이다. 교육 분야 논문은 일반적으로 “누구를 대상으로”, “어떤 교육 내용 또는 맥락에서”, “어떤 방법으로” 연구가 수행되었는가에 따라 연구의 성격이 구분된다. 따라서 연구대상은 해당 연구에서 분석, 학습, 참여, 평가의 직접 대상이 되는 집단을 의미하며, 유아, 초등, 중등, 고등, 대학, 대학원, 교사, 예비 교사, 학습자, 사용자/참여자, 환자 등이 이에 해당한다. 연구분야는 연구가 적용되거나 논의되는 교육 영역, 교과, 기술 환경 또는 실천 맥락을 의미하며, 학교교육, 고등교육, 의료교육, 안전교육, 언어/영어, 과학, 예술/문화예술, VR/가상현실, AR/증강현실, 플랫폼/시스템 등이 이에 해당한다. 연구방법은 연구자가 사용한 주요 연구 설계, 교육 적용 방식, 자료 수집 및 분석 절차를 의미하며, 실험/효과검증, 설계/개발, 시뮬레이션/시뮬레이터, 설문조사, 인터뷰/FGI, 사례연구, 문헌연구 등이 포함된다.

이러한 세 차원은 기존 메타버스 교육 토픽모델링 연구 [8-12]에서 반복적으로 확인된 주제 범주를 교육 연구의 분석 구조에 맞추어 재구성한 것이다. 기존 연구들에서는 메타버스, 가상현실, 증강현실과 같은 기술 환경뿐 아니라 학교급, 학습자 집단, 교과 영역, 의료·간호·안전·문화예술 등 적용 분야, 실험·설계·사례·문헌분석과 같은 연구방법이 주요 토픽 또는 핵심 키워드로 반복적으로 제시되었다. 본 연구는 이러한 반복 출현 범주를 그대로 나열하는 대신, 교육 연구의 핵심 구조인 대상, 분야, 방법의 세 축으로 정리하여 구조화된 메타데이터를 생성하였다. 이를 통해 각 문헌은 단순 단어 빈도나 표면적 키워드가 아니라, 교육 연구로서의 맥락적 속성을 반영하는 형태로 변환되었다. 이는 원문 기반 분류에서 발생할 수 있는 불필요한 표현 다양성을 줄이고, 문헌의 핵심 속성을 통제된 라벨 공간으로 정규화함으로써 토픽 할당의 계산 효율과 재현성을 동

시에 확보하기 위한 설계이다.

구체적인 연구절차를 살펴보면, OpenAI의 최신 모델인 GPT-5.2[15]를 API 방식으로 호출하여 전체 문헌의 제목과 초록을 입력값으로 제공하고, 연구대상·연구분야·연구방법을 구조화된 형식으로 추출하였다. 모델은 사전 정의된 라벨 집합 중에서만 선택하도록 제한되었으며, 각 판단에 사용된 핵심 표현을 함께 기록하도록 설계되었다. 모델 실행 시 temperature는 0으로 설정하여 동일 입력에 대해 출력의 일관성을 높이도록 하였고, 출력 형식은 JSON 구조로 통일하여 후속 분석에서의 기계적 처리와 검증이 가능하도록 하였다. 이 과정을 통해 각 문헌은 연구대상, 연구분야, 연구방법 라벨의 조합으로 표현되는 구조화 벡터로 변환되었으며, 이는 이후 토픽모델링의 입력 공간으로 활용되었다.

각 차원에서 라벨은 최대 2개까지 선택하도록 제한하였다. 이는 메타버스 교육 연구가 복합적 성격을 가지기 때문에 단일 라벨만 허용할 경우 연구의 주요 맥락이 과도하게 축소될 수 있는 반면, 라벨 수를 제한하지 않을 경우 초록에 등장하는 주변적 표현이나 부차적 맥락까지 라벨로 선택되어 후속 토픽모델링 입력의 일관성이 낮아질 수 있기 때문이다. 이처럼 최대 2개 라벨은 문헌의 중심 속성과 보조 속성을 함께 반영하되, 과도한 다중 라벨링으로 인한 노이즈를 줄이기 위한 절충적 기준이다. 따라서 본 연구의 라벨링 방식은 문헌의 복합성을 보존하면서도, 후속 LDA 및 BERTopic 분석에 투입되는 메타데이터 벡터의 해석 가능성과 비교 가능성을 유지하기 위한 설계로 볼 수 있다. Fig. 2는 메타데이터 추출을 위한 핵심 프롬프트를 제시한다.

```

INSTRUCTIONS = f"""
너는 메타버스 교육 분야의 논문분석 전문가다.
입력으로 제공되는 제목과 초록만 근거로, 연구대상/연구분야/연구방법을 라벨로 분류하고
각 범주별로 토픽모델링에 적합한 정규화 토큰(evidence_norm)을 생성하라.

출력 규칙(절대 위반 금지):
1) 출력은 오직 JSON만 허용한다(설명 문장 금지).
2) 다음 스키마를 정확히 지켜라.
{{
  "연구대상": {{
    "labels": ["...", "..."],
    "evidence_norm": ["...", "...", "...", "...", "...", "...", "..."]
  }},
  "연구분야": {{
    "labels": ["...", "..."],
    "evidence_norm": ["...", "...", "...", "...", "...", "...", "..."]
  }},
  "연구방법": {{
    "labels": ["...", "..."],
    "evidence_norm": ["...", "...", "...", "...", "...", "...", "..."]
  }}
}}

라벨 규칙:
3) labels는 각 항목 최대 2개.
4) labels는 아래 목록에서만 선택한다.
- 연구대상 라벨: {TARGET_LABELS}
- 연구분야 라벨: {FIELD_LABELS}
- 연구방법 라벨: {METHOD_LABELS}
5) 근거가 약하면 labels는 ["해당없음"]으로 하라.
labels가 ["해당없음"]이면 evidence_norm은 빈 리스트로 하라.
    
```

Fig. 2. LLM Instruction for Extracting Metadata

한편, 분류 결과의 정확성과 타당성을 검증하기 위해 전체 분석 대상 문헌 중 100편을 무작위 표집하여, 교육학 박사 1인과 공학 박사 1인이 동일한 코딩 지침에 따라 각 문헌의 연구대상, 연구분야, 연구방법을 독립적으로 코딩하였다. 이후 두 연구자의 합의 결과를 준거 기준으로 설정하고, 이를 LLM의 분류 결과와 비교하였다. 분석에 앞서 초기 파일럿 검토에서 확인된 일부 오류 유형, 특히 연구대상과 연구분야 간의 용어 혼용 문제를 줄이기 위해 프롬프트를 보완하였으며, 각 범주별 금지어와 우선 판별 기준을 추가하여 최종 프롬프트를 확정하였다. 그 결과, 최종 프롬프트 기준 LLM 분류 결과와 인간 전문가 합의 결과 간의 Cohen's kappa는 연구대상 .96, 연구분야 .92, 연구방법 .96으로 나타났으며, 평균 Cohen's kappa는 .95였다. 이는 Landis and Koch(1977)의 기준에 따르면 almost perfect agreement 수준에 해당하며[16], 본 연구에서 사용한 LLM 기반 메타데이터 구조화가 인간 전문가 코딩과 비교하여 높은 수준의 신뢰도를 확보하고 있음을 시사한다. 아울러 temperature=0 설정에 따른 재현성을 점검하기 위해 동일한 입력 문헌 100편에 대해 동일한 프롬프트와 동일한 파라미터 조건으로 3회 반복 실행한 결과, JSON 스키마 일치율은 100%였고, 최종 라벨 일치율은 연구대상 100%, 연구분야 99%, 연구방법 98%로 나타났다. 소수의 차이는 주로 설명 필드의 서술 순서나 근거 문구의 표현 차이에서 발생하였으며, 최종 분류 결과는 전반적으로 안정적으로 유지되었다.

### 3. Initial Topic Extraction and Integrated Design Based on Structured Metadata

구조화된 메타데이터를 바탕으로 본 연구는 서로 다른 이론적 전제를 갖는 두 가지 토픽모델을 병렬적으로 적용하였다. 첫째, LDA를 적용하였다. LDA는 문서-단어 분포를 기반으로 잠재 토픽을 추정하는 확률적 생성모형이다[1]. 다만 본 연구의 LDA는 일반적인 Bag-of-Words 중심 접근과 달리, 기존 메타버스 교육 토픽모델링 연구[8-12] 검토를 통해 정제된 메타라벨과 근거 키워드를 입력 단위로 활용하여 토픽을 추정하였다. 즉, 단순 어휘 빈도에 의존하기보다 연구대상, 연구분야, 연구방법 등 구조화된 정보에 근거하여 주제 분포를 파악하도록 설계하였다.

토픽 수는 기존 메타버스 교육 토픽모델링 연구[8-12]에서 보고된 토픽 수의 분포를 참고하여 후보 범위를 설정한 뒤, coherence(c\_v) 지표를 활용해 의미 응집도를 평가하였다[17]. 기존 연구들은 대체로 5개에서 11개 사이의 토픽을 제시하였으므로, 본 연구에서는 이 범위를 참고하여 후

보 토픽 수를 설정하고 각 후보별 의미 응집도와 해석 가능성을 함께 검토하였다. 또한 여러 random seed를 사용하여 반복 학습을 수행함으로써 토픽 구조의 안정성을 점검하였으며, 평균 coherence가 높고 seed 간 변동이 낮은 값을 보이는 토픽 수를 우선적으로 검토하였다. 최종적으로 7개 토픽은 coherence와 안정성 측면에서 비교적 양호하였고, 기존 연구[8-12]에서 반복적으로 보고된 주요 주제 범주를 과도하게 병합하거나 세분화하지 않는 수준으로 판단되어 초기 토픽 수로 선정하였다. 이러한 절차는 특정 초기값에 따른 우연한 결과를 줄이고, 보다 안정적이며 해석 가능한 초기 토픽 구조를 확보하기 위한 것이다.

LDA 구현에는 gensim을 사용하였다. gensim은 LDA 학습, coherence 계산, passes, alpha, random seed 등 주요 하이퍼파라미터 설정을 명시적으로 지원하므로, 본 연구와 같이 토픽 수 후보를 비교하고 반복 실행을 통해 안정성을 점검하는 분석에 적합하다. 본 연구에서는 LDA의 학습 안정성을 높이기 위해 passes를 20으로 설정하였고, 문헌별 토픽 분포가 지나치게 균일하거나 편향되지 않도록 alpha는 'auto'로 설정하였다.

둘째, BERTopic을 적용하였다. BERTopic은 문헌 임베딩을 기반으로 의미적 유사성을 계산하고, 이를 토대로 문헌 군집을 형성하는 임베딩 기반 토픽모델이다[3]. 본 연구에서는 LDA와 비교 가능성을 확보하기 위해 BERTopic 결과도 동일한 토픽 수를 기준으로 정리하였다. 이는 두 모델의 결과를 기계적으로 동일시키기 위한 것이 아니라, 확률 기반 토픽 구조와 임베딩 기반 의미 군집 구조를 동일한 상위 범주 수준에서 비교하기 위한 절차이다. BERTopic의 결과는 단어 빈도와 확률 분포를 중심으로 토픽을 추정하는 LDA와 달리, 의미 임베딩 공간에서 형성되는 문헌 간 근접성과 군집 구조를 반영한다는 점에서 LDA 결과를 보완한다. BERTopic 구현에는 bertopic과 sentence-transformers를 사용하였다. bertopic은 임베딩 기반 문서 군집화와 토픽 표현 추출을 통합적으로 지원하며, sentence-transformers는 문헌의 의미 표현을 생성하는 임베딩 모델을 적용하는 데 활용하였다. 본 연구의 분석 대상은 한국어 초록을 중심으로 하되 일부 영문 초록을 한국어로 번역한 문헌을 포함하므로, 한국어 문헌의 의미 표현과 다국어 텍스트 처리에 모두 대응할 수 있는 임베딩 모델이 필요하였다. 이에 본 연구에서는 다국어 임베딩 모델인 BAAI/bge-m3를 활용하였다. 이 모델은 100개 이상의 언어를 지원하는 범용 다국어 임베딩 모델로 알려져 있으며, 검색, 문서 유사도 계산, 군집화 등 다양한 텍스트 표현 과제에서 안정적인 성능을 보이는 것으로 보고

되어 본 연구의 분석 목적에 적합하다고 판단하였다[18].

주요 실험 세팅은 Table 3에 제시하였다. Table 3의 설정값은 단순 실행 환경의 나열이 아니라, LDA와 BERTopic의 상호보완적 비교, 토픽 수 결정의 안정성 점검, 한국어 중심 코퍼스의 의미 표현 확보라는 본 연구의 분석 목적에 따라 결정되었다.

Table 3. Experimental Settings

Category	LDA	BERTopic
Execution Environment	Google Colab (Python 3.x)	Google Colab (Python 3.x, T4 GPU)
Library	gensim	BERTopic, sentence-transformers
No. of Topics	7	7
Other Settings	passes = 20, alpha = 'auto'	embedding_model = "BAAI/bge-m3"

그러나 최종 토픽 수의 확정과 초기 토픽 집합의 구성은 두 모델의 내부 지표만으로 결정하지 않았다. 본 연구는 기존 메타버스 교육 토픽모델링 연구 5편에서 반복적으로 제시된 주제 범주를 참조 기준으로 설정하고, 이를 바탕으로 LDA와 BERTopic 결과를 비교·검토하여 초기 통합 토픽 체계를 구성하였다. 이 과정은 자동화된 수학적 연산이라기보다, 서로 다른 모델에서 도출된 토픽들을 기존 연구 [8-12]의 축적된 주제 구조와 대조하면서 상위 범주 수준에서 재배열하는 규칙 기반의 질적 통합 절차에 해당한다.

통합 과정에는 연구자 2인(교육학 박사 1인, 공학 박사 1인)이 참여하였다. 두 연구자는 각 모델의 대표 키워드, 구조화 메타데이터 분포, 해당 토픽에 포함된 문헌 사례를 함께 검토하여, 개별 토픽이 독립된 상위 주제로 유지되어야 하는지, 기존 범주와 통합되어야 하는지, 혹은 지나치게 세분화된 하위 변형에 해당하는지를 독립적으로 판단하였다. 초기 판단이 일치하지 않는 경우에는 관련 문헌의 맥락과 기존 연구[8-12]의 토픽 정의를 다시 검토한 뒤 합의에 이를 때까지 논의를 반복하였으며, 최종적으로 합의된 결과만 초기 통합 토픽 체계에 반영하였다.

통합의 기준은 다음 세 가지였다. 첫째, LDA와 BERTopic에서 핵심 키워드와 의미 범주가 실질적으로 동일하거나 매우 유사한 경우에는 하나의 상위 주제로 병합하였다. 이는 모델 간 표현 방식의 차이로 분리되어 나타난 동일 주제를 재결합하기 위한 절차이다. 둘째, 하나의 상위 주제가 세부 적용 맥락이나 표현 차이로 과도하게 분절되어 나타난 경우에는, 기존 연구[8-12]에서 반복적으로 확인된 상위 범주를 기준으로 재통합하였다. 이는 모델 특성상 발생할 수 있는 과잉 세분화를 완화하기 위한 조정

원칙이다. 셋째, 기존 연구[8-12]에서 지속적으로 보고된 핵심 주제는 가능한 한 독립 범주로 유지하였다. 이는 연구 연속성과 학문적 축적성을 확보하고, 후속 비교 분석에서 해석 가능성을 높이기 위한 기준이다.

본 연구에서 기존 연구[8-12], LDA, BERTopic은 동일한 비중으로 기계적으로 결합된 것이 아니라, 서로 다른 역할을 갖는 참조 근거로 활용되었다. 기존 메타버스 교육 토픽모델링 연구[8-12]의 토픽 구조는 상위 범주의 안정성과 학문적 연속성을 판단하는 1차 기준으로 사용되었고, LDA 결과는 메타라벨 분포에 기반한 통계적 응집성을 확인하는 근거로 활용되었으며, BERTopic 결과는 임베딩 기반 의미 유사성과 문헌 군집의 자연스러운 분화를 점검하는 보완적 자료로 활용되었다. 따라서 본 연구의 통합은 수치적 가중합이 아니라, 기존 연구[8-12]를 우선 기준으로 두고 두 모델의 결과를 상호보완적으로 대조·조정하는 방식으로 이루어졌다.

이와 같은 절차를 통해 도출된 초기 토픽 집합은 두 모델의 결과를 단순 병합한 산출물이 아니라, 기존 연구[8-12] 기반 주제 구조와의 의미적 정합성, 모델 간 공통성, 과도한 세분화 여부를 종합적으로 검토하여 연구진이 합의한 초기 분석 틀이다. 이러한 설계는 단일 모델 의존에서 발생할 수 있는 구조적 편향을 줄이고, 학문적 맥락과의 정합성을 유지하면서도 후속 반복 정제 단계에서 활용 가능한 비교 기준을 제공한다는 점에서 방법론적 의의를 가진다.

#### 4. Iterative Multi-Model Consensus-Based Topic Refinement

초기 토픽 집합은 LDA와 BERTopic 결과를 기계적으로 병합한 것이 아니라, 기존 연구[8-12]에서 제시된 토픽 범주와의 의미적 정합성을 고려하여 구성된 통합 구조이다. 그러나 이 구조 역시 일회적 결과로 확정하지 않고, 단계적 표본 적용과 선택적 교차검증을 통해 안정성을 점검하도록 설계하였다.

우선 전체 데이터 중 일부 표본을 무작위 추출하여 초기 토픽 집합을 적용하였다. 이 단계는 토픽 구조의 구조적 적합성을 점검하기 위한 파일럿 적용 단계로서, 토픽이 실제 문헌을 충분히 설명할 수 있는지, 특정 영역이 과도하게 세분화되거나 반대로 과도하게 통합되어 있지는 않은지를 확인하기 위한 목적을 가진다. 표본 비율은 점진적으로 확대하도록 설계하였으며, 초기 소규모 적용(5%) 이후 중간 규모 적용(10%)을 통해 구조적 안정성을 재확인하는 방식으로 진행되었다.

각 문헌에 대한 토픽 할당은 LLM(GPT-5.2 [15] API)을 활용하여 수행되었으며, 문헌은 현재 토픽 집합 중 하나로 분류되도록 하였다. 이때 모델은 토픽 선택과 함께 분류 확신도(confidence)를 산출하도록 설정하였다. 분류 확신도는 해당 문헌이 기존 토픽 체계 내에서 충분히 설명 가능한지 여부를 판단하는 1차 지표로 활용되었다.

교차검증 단계에서는 모든 문헌에 대해 이중 LLM을 적용하지 않고, 분류 확신도가 낮거나 기존 토픽 체계로 설명하기 어려운 경우에 한하여 이중 LLM인 Gemini 2.5 Flash API[19]를 호출하도록 설계하였다. 즉, 분류 확신도가 낮거나 기존 토픽 체계로 설명하기 어려운 경계 사례에 대해서만 이중 LLM을 추가로 호출하는 선택적 교차검증 구조를 적용하였다. 이러한 설계는 계산 비용을 통제하면서도 경계 사례에 대한 의미적 검증을 강화하기 위한 것이다.

이중 LLM이 제안한 신규 토픽 후보는 즉시 확정하지 않고, 의미 수준에서 재검토하였다. 구체적으로, 신규 후보가 기존 토픽과 의미적으로 중복되는지, 기존 토픽의 세분화에 해당하는지, 또는 독립적 주제를 형성하는지를 판단하였다. 이는 기존 연구[8-12]의 토픽구조, 초기 LDA 및 BERTopic 결과, 그리고 표본 적용 과정에서의 분류 패턴을 종합적으로 고려하여 수행되었다.

이 과정은 단일 표본 적용으로 종료되지 않았으며, 표본 비율을 점진적으로 확대하면서 동일한 절차를 반복하였다. 즉, 소규모 표본에서 구조적 적합성을 점검한 이후 중간 규모 표본으로 확장하여 토픽 체계의 안정성을 재확인하였다. 각 단계에서 더 이상 의미적 재구성이 요구되지 않는 상태, 다시 말해 신규 후보가 기존 구조 안에서 설명 가능하고 분류 불일치가 구조적 결함으로 이어지지 않는 수준에 도달했을 때 토픽 구조를 확정하였다.

이와 같은 반복적 재구성 절차는 토픽모델링 결과를 일회적 군집 산출물로 간주하지 않고, 표본 확대와 교차검증을 통해 점진적으로 안정화되는 합의 구조로 이해한다는 점에서 본 연구 설계의 핵심 특징을 이룬다.

#### 5. Finalization of the Topic Set and Full-Corpus Application

정교화 과정을 거쳐 확정된 최종 토픽 집합은 전체 문헌에 대해 전수 적용되었다. 이 단계에서는 새로운 토픽 생성을 허용하지 않고, 모든 문헌이 확정된 토픽 집합 내의 범주 중 하나로 반드시 할당되도록 설계하였다. 이는 연구 동향 분석이라는 목적상 모든 문헌을 동일한 비교 구조 내에 위치시키기 위함이다. 토픽 할당은 LLM을 활용하여 수행되었으며, 실행 시 temperature를 0으로 설정하여 재

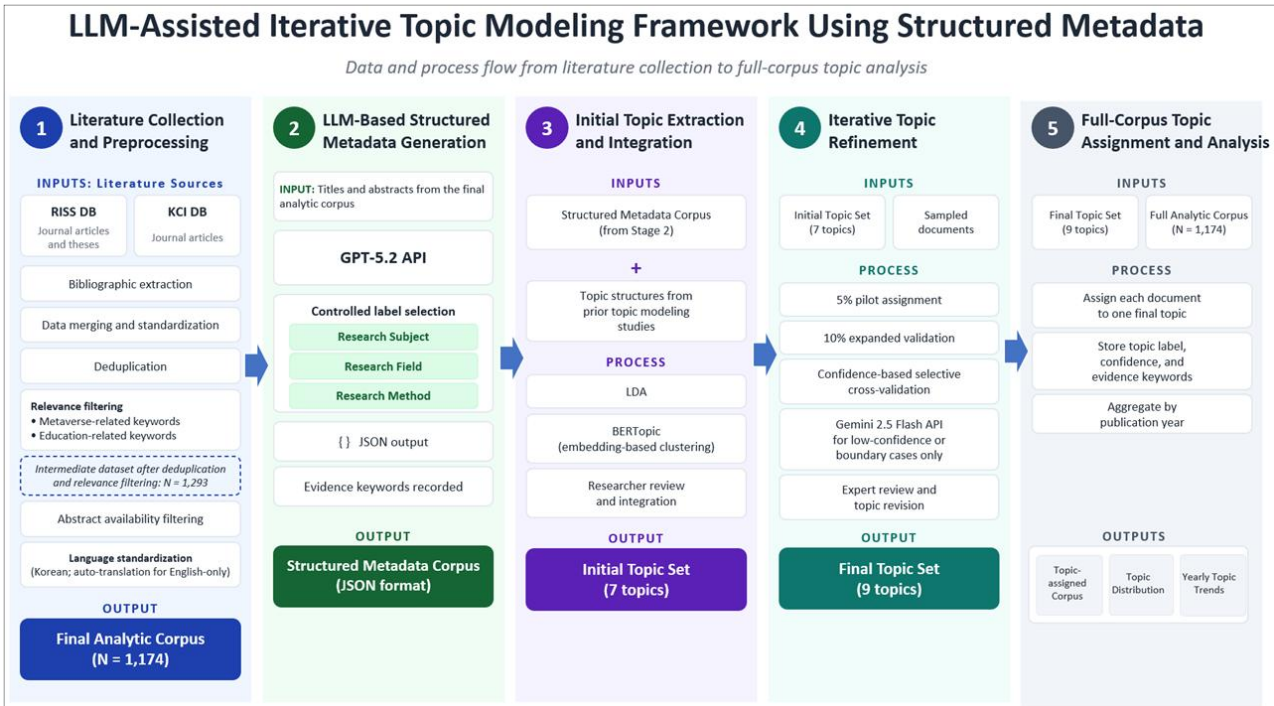


Fig. 3. Data and Process Flow of the LLM-Assisted Iterative Topic Modeling Framework

현성을 높이도록 하였다. 각 문헌의 최종 토픽, 모델 확신도, 판단에 사용된 근거 정보는 함께 저장되었다.

전수 적용 이후에는 문헌의 발행 연도를 기준으로 토픽 분포를 재구성하여 연도별 추이를 분석하였다. 구체적으로, 각 연도에 속한 문헌 집합을 기준으로 토픽별 빈도와 상대 비율을 산출하고, 이를 시계열 구조로 배열하여 특정 토픽의 증가, 감소, 정체 여부를 비교하였다. 이 과정은 단순 빈도 비교를 넘어, 메타버스 교육 연구의 관심 축이 시기별로 어떻게 이동하였는지를 구조적으로 파악하기 위한 설계이다. 또한 토픽 간 상대적 비중 변화를 통해 특정 시기에 부상한 주제와 점진적으로 축소된 주제를 구분하고, 연구 생태계의 확장 및 재편 양상을 해석할 수 있도록 하였다.

이와 같은 전수 적용 단계는 탐색적 토픽 생성과 구분되는 확정 단계로서, 분포 분석, 시기별 변화, 비중 비교, Hot/Cold Topic 식별을 가능하게 하는 기반 데이터셋을 구축한다[20].

## IV. Results

### 1. Results of LLM-Based Metadata Structuring Grounded in Prior Research

기존 연구[8-12]를 바탕으로 설계한 구조화 메타데이터 체계를 전체 문헌에 적용한 결과, 메타버스 교육 연구의 주요 관심 축이 연구대상, 연구분야, 연구방법의 세 차원

에서 비교적 뚜렷하게 나타났다. 연구대상 차원에서는 ‘학습자’가 612회로 가장 높게 나타났으며, 대학(240회), 사용자/참여자(196회), 교사(196회), 일반인/성인(146회), 초등(140회)이 뒤를 이었다. 이는 메타버스 교육 연구가 전반적으로 학습자 중심의 성격을 가지는 동시에, 대학 교육 맥락에서의 적용 연구가 상대적으로 강하게 축적되어 왔음을 보여준다. 아울러 사용자 경험을 중심으로 한 활용 연구와 교사의 수업 적용 및 실행 맥락을 다루는 연구도 중요한 비중을 차지하고 있음을 확인할 수 있다.

연구분야 차원에서는 VR/가상현실이 777회, 플랫폼/시스템이 506회로 가장 높은 빈도를 보였다. 메타버스는 불용어로 처리하여 빈도순위에서 제외하였다. 이는 메타버스 교육 연구가 기술 기반 환경과 시스템 구현 맥락을 중심으로 전개되어 왔음을 시사한다. 동시에 학교교육(201회), 예술/문화예술(174회), 고등교육(143회), 언어/영어(127회) 등의 범주가 뒤를 이은 점은, 초기의 기술 중심 논의가 실제 교육 적용 영역으로 점차 확장되고 있음을 보여준다. 즉, 메타버스 교육 연구는 기술 인프라 자체에 대한 탐색을 넘어, 교과 및 학습 맥락과 결합하는 방향으로 연구 범위를 넓혀 온 것으로 해석할 수 있다.

연구방법 차원에서는 설계/개발(384회), 실험/효과검증(331회), 문헌연구(319회), 사례연구(318회), 설문조사(300회)가 상위에 위치하였다. 이러한 결과는 메타버스 교육 연구가 단순한 개념 제안이나 기술 소개에 머무르지 않고, 실제 교수-학습 환경에서의 설계와 구현, 효과 검증, 사례

축적을 병행하는 응용 연구 단계로 발전하고 있음을 보여 준다. 특히 설계/개발과 실험/효과검증이 함께 높은 비중을 차지한다는 점은, 메타버스 교육 연구가 기술적 구현과 교육적 효과 검증을 동시에 중시하는 실천적 연구 경향을 갖고 있음을 시사한다.

Fig. 4는 LLM 기반 구조화 메타데이터 생성 과정의 실제 산출 예시를 보여준다. 각 문헌은 GPT-5.2 기반 메타데이터 추출 과정을 통해 연구대상, 연구분야, 연구방법의 라벨과 함께, 해당 판단의 근거가 된 핵심 표현 (evidence\_norm)으로 변환되었다. 이는 단순 라벨 부여에 그치지 않고, 분류 결과의 해석 가능성과 검토 가능성을 함께 확보하였음을 보여준다. 특히 동일 문헌에 대해 범주별 라벨과 근거 표현이 함께 제시됨으로써, 후속 토픽 통합 과정에서 분류의 타당성을 재검토할 수 있는 기반이 마련되었음을 확인할 수 있다.

연구대상 _labels	연구분야 _labels	연구방법 _labels	연구대상 _evidence_norm	연구분야 _evidence_norm	연구방법 _evidence_norm
[일반인/성인, 사용자/참여자]	[플랫폼/시스템, 사회문화/디지털포용]	[데이터마이닝/로그분석, 혼합연구]	[메타버스, 메타버스플랫폼, 비대면서비스, 가상공간, 3D가상공간, 메타버스리터러시]	[비정형빅데이터분석, 텍스트마이닝, 키워드빈도분석, TFIDF, 연결중심성분석, C...	
[학습자]	[VR/가상현실, 플랫폼/시스템]	[실험/효과검증, 사용자경험(UX)/경험분석]	[메타버스시스템, 확장현실의 시스템, 원격교육, 비대면교육, 온라인학습, 사용자경험]	[현장실험, 비교실험, 효과검증, 경험분석, 재미의도 측정, 재방문의도측정]	
[학습자, 교사]	[플랫폼/시스템, VR/가상현실]	[설계/개발]	[학생, 강사, 교사]	[메타버스교육, 비대면교육플랫폼, 멀티반플랫폼, 게임기반학습, 전자질문, 확장현실의수업]	
[교사]	[플랫폼/시스템, VR/가상현실]	[사용자경험(UX)/경험분석, 혼합연구]	[교수자]	[메타버스교육플랫폼, 웹기반가상공간, 가상현실, 메타버스, 온라인강의, 아바타]	[사용성평가, 온라인 설문조사, 심층면접, 만족도조사, 필요성 분석, 경험분석]

Fig. 4. Sample Output of LLM-Based Metadata Extraction

Table 4는 전수 적용된 구조화 메타데이터를 집계한 결과로서, 연구대상, 연구분야, 연구방법의 각 차원에서 상대적으로 높은 빈도를 보인 상위 항목들을 시각적으로 제시한다.

## 2. Results of Topic Modeling Based on Structured Metadata

기존 연구[8-12]에서 반복적으로 확인된 주제들을 참조하고, LDA와 BERTopic의 개별 결과를 비교·통합한 결과, 초기 통합 토픽 체계는 총 7개의 상위 토픽으로 수렴하였다. 이는 LDA의 확률적 토픽 구조와 BERTopic의 임베딩 기반 군집 구조를 상호 대조하면서, 기존 연구[8-12]에서 지속적으로 보고된 핵심 주제를 상위 범주 수준에서 재배열한 결과이다[1][3]. LDA와 BERTopic의 개별 결과는 각각 Table 5와 Table 6에 제시하였다. LDA 결과는 문서-단어 분포에 기반한 확률적 응집 구조를 보여주며, BERTopic 결과는 임베딩 기반 의미 근접성과 군집 구조를 보다 직접적으로 반영한다. 두 결과를 비교하면, 공통

Table 4. Top Frequency Labels in Structured Metadata Across the Full Corpus

	Top 10 Words	Counts
Research Field	VR/Virtual Reality	777
	Platform/System	506
	School Education	201
	Arts/Culture & Arts	174
	Higher Education	143
	Language/English	127
	Vocational Education/Training	51
	Socio-culture/Digital Inclusion	47
	Immersive Learning/Presence	45
	Religion/Christianity	34
Research Method	Design/Development	384
	Experiment/Effectiveness Verification	331
	Literature Review	319
	Case Study	318
	Survey	300
	Interview/FGI	135
	Mixed Methods Research	110
	UX/Experience Analysis	95
	Performance/Outcome Evaluation	82
	System/Platform Construction	37
Research Subject	Learners	612
	University/College	240
	Users/Participants	196
	Teachers	196
	General Public/Adults	146
	Elementary School	140
	Not Applicable	119
	Middle School	101
	Pre-service Teachers	51
	High School	46

적으로 플랫폼·가상현실·수업·학습·효과 검증과 같은 핵심 키워드들이 반복적으로 나타나지만, BERTopic은 간호, 종교, 언어 교육과 같은 응용 맥락을 보다 분화된 형태로 드러내는 경향을 보였다. 반면 LDA는 설계·개발, 효과 검증, 사용자 경험, 사회·디지털 포용과 같은 상위 주제 범주를 상대적으로 응집된 형태로 제시하였다. 따라서 본 연구에서는 두 모델의 결과를 기계적으로 결합하지 않고, 각 모델이 산출한 대표 키워드, 구조화 메타데이터의 의미 범주, 포함 문헌의 사례, 기존 연구[8-12]의 토픽 구조와의 대응성을 종합적으로 검토하여 초기 통합 토픽 집합을 구성하였다.

Table 5. Results of LDA

No.	Topic Keywords
1	Teacher, Arts, Culture, Class/Instruction, Case Study, Development, Design, Student, Program, School
2	Literature, Design, Case Study, Development, Learning, System, Class/Instruction, English, Language, University/College
3	Society, Digital, Children, Culture, Interview, Inclusion, Safety Education, FGI (Focus Group Interview), Qualitative Research
4	Survey, Adult, Participant, User, General Public, Digital, System, Service, Avatar, Structure
5	Effectiveness, Verification, Experiment, Survey, Student, Elementary School, Learning, Group, Test, Post-test
6	User, Experience, Expert, Participant, Learning, System, Immersion/Presence, Interview, Evaluation
7	Vocational, Training, Subject/Target, Special Education, Disability, Student, Media, Performance, Development, Setting

Table 6. Results of BERTopic

No.	Representative Topic Keywords
1	Analysis, Virtual Reality, Design, VR, Development, Learning, Platform, Class/Instruction, Learner, Case Study
2	Teacher, Learning, Platform, Effectiveness, Survey, Verification, Virtual Reality, Learner, Class/Instruction
3	User, Vocational, Participant, Training, Adult, Platform, Analysis, General Public, System
4	Literature, Analysis, Virtual Reality, Case Study, VR, Platform, Review, Reflection, Digital, Culture, Concept
5	Christianity, Religion, Literature, Church, Concept, Review/Reflection, Theory, Generation, Worship, Virtual Reality
6	Design, Language, English, Class/Instruction, Platform, Learning, Model, Development, Case Study, Learner
7	Nursing, Verification, Effectiveness, Post-test, College Student, University, Control Group, Test, Experiment

통합 과정은 3장 2절에서 제시한 통합 기준을 따라, 플랫폼·가상환경·시스템 설계 관련 토픽은 “메타버스 교육 플랫폼 및 시스템 설계”로, 의료·간호·재활·시뮬레이션 관련 토픽은 “보건의료 및 재활 교육”으로, 장애·직업·기술 훈련 관련 토픽은 “특수교육 및 직업·기술 훈련”으로 정리하였다. 또한 “교육 효과 검증 및 학습 성과 연구”와 “사회문화적 맥락 및 디지털 포용”은 기존 연구에서 반복적으로 보고된 핵심 주제이므로 별도의 상위 토픽으로 유지하였다.

구체적으로 Topic 1인 “Design of Metaverse Educational Platforms and Systems”는 LDA-2의 design/system 관련 키워드와 BERTopic-1 및

BERTopic-6의 VR/design/platform 관련 키워드가 플랫폼, 시스템, 설계, 개발, 가상현실이라는 의미 범주에서 수렴하여 통합하였다. Topic 2인 “Verification of Educational Effectiveness and Learning Outcome Research”는 LDA-5와 BERTopic-2 및 BERTopic-7에서 effectiveness, verification, experiment, survey, pre-/post-test 등의 키워드가 공통적으로 나타나 별도의 상위 토픽으로 유지하였다. Topic 3인 “User Experience (UX) and Immersion-Based Learning Research”는 LDA-6과 BERTopic-3이 user, participant, experience, immersion, perception 관련 키워드를 공유하여 사용자 경험과 몰입 기반 학습이라는 상위 범주로 통합하였다.

Topic 4인 “Application of School-Based and Subject-Specific Instruction”은 LDA-1의 teacher/class 관련 키워드와 BERTopic-2 및 BERTopic-6의 class/learning/language 관련 키워드가 학교 수업, 교과 적용, 언어·예술 등 교과 기반 수업 맥락에서 수렴하여 구성하였다. Topic 5인 “Health-Medical and Rehabilitation Education”은 BERTopic-7에서 nursing, medical, verification, experiment 관련 키워드가 비교적 명확하게 분화되었고, 기존 연구에서도 의료, 재활, 수술 훈련, 시뮬레이션 관련 토픽이 반복적으로 보고되어 독립적인 응용 분야 토픽으로 유지하였다. Topic 6인 “Special Education and Vocational/Technical Training”은 LDA-7과 BERTopic-3에서 vocational, training, disability, special education, engineering, technology 관련 키워드가 함께 나타나 특수교육 및 직업·기술 훈련이라는 상위 범주로 통합하였다. Topic 7인 “Socio-Cultural Contexts and Digital Inclusion”은 LDA-3의 society/digital/inclusion 관련 키워드와 BERTopic-4 및 BERTopic-5의 culture/concept/religion 관련 키워드가 사회문화적 맥락과 디지털 포용이라는 상위 의미 범주에서 연결되어 하나의 통합 토픽으로 구성하였다.

이와 같이 본 연구의 초기 토픽 통합은 LDA와 BERTopic의 토픽 번호를 일대일로 대응시키는 방식이 아니라, 대표 키워드의 의미적 수렴, 구조화 메타데이터의 범주적 유사성, 포함 문헌의 사례, 기존 연구[8-12]의 토픽 구조와의 대응성을 종합적으로 검토한 절차이다. 따라서 두 모델은 기계적으로 결합된 것이 아니라, 초기 통합 토픽 체계를 구성하기 위한 상호보완적 참조 근거로 활용되었다. 최종 초기 통합 토픽 집합은 Table 7과 같다.

Table 7. Initially Integrated Topic Set Derived from LDA, BERTopic, and Prior Topic Modeling Studies

No	Topic Name	Representative Keywords	Model Basis (LDA & BERTopic)	Related Prior Topic Modeling Studies (Authors & Topics)
1	Design of Metaverse Educational Platforms and Systems	Metaverse, Platform, System, Design, Development, Virtual Reality, VR, Environment	LDA-2 (Design/System); BERTopic-1 (VR/Design), 6 (Design/Platform)	Han & Kim (2023) Topics 1, 2, 4 (Space/Class/Content Dev); Kim & Im (2022) Topics 1, 4 (Platform/VR)
2	Verification of Educational Effectiveness and Learning Outcome Research	Effectiveness, Verification, Experiment, Survey, Pre- and Post-test, Examination, Learning Outcomes, Group	LDA-5 (Effectiveness/Verification); BERTopic-2 (Effectiveness/Survey), 7 (Verification/Experiment)	Kim & Im (2022) Topic 6 (Effects); Kim So-hee (2023) Topic 4 (Learning Effects)
3	User Experience (UX) and Immersion-Based Learning Research	User, Experience, Immersion, Participant, Survey, Perception, Presence	LDA-6 (User/Experience); BERTopic-3 (User/Participant)	Ozyurt and Ozyurt (2025) Topic 4 (Learning Experience); Kim So-hee (2023) Topic 5 (Immersive/Engagement)
4	Application of School-Based and Subject-Specific Instruction	Class/Instruction, School, Elementary, Learning, Science, Language, Arts, Curriculum	LDA-1 (Teacher/Class); BERTopic-2 (Class/Learning), BERTopic-6 (Language/Class)	Lim et al. (2022) Topic 6 (School Class); Han & Kim (2023) Topics 2, 5 (Subject/Elementary); Kim So-hee (2023) Topic 1 (School VR Utilization)
5	Health-Medical and Rehabilitation Education	Nursing, Medical, Rehabilitation, Patient, Practice/Practicum, Simulation, Surgical, Training	BERTopic-7 (Nursing/Verification)	Ozyurt and Ozyurt (2025) Topics 2, 7, 8 (Medical/Surgical/Rehab); Kim & Im (2022) Topics 2, 3 (Rehab/Medical Ed); Lim et al. (2022) Topic 2 (Rehab Training)
6	Special Education and Vocational/Technical Training	Disability, Special Education, Vocation, Job Task, Training, Engineering, Technology, Autism	LDA-7 (Vocational/Training); BERTopic-3 (Vocational/Training)	Ozyurt (2025) Topic 9 (ASD); Lim et al. (2022) Topic 3 (Disability/Vocational Ed); Kim So-hee (2023) Topic 3 (Technical Training)
7	Socio-Cultural Contexts and Digital Inclusion	Society, Digital, Inclusion, Religion, Christianity, Elderly, Children, Welfare	LDA-3 (Society/Digital); BERTopic-4 (Culture/Concept), BERTopic-5 (Religion)	Han & Kim (2023) Topic 3 (Welfare/Religion); Ozyurt (2025) Topic 1 (Cultural Heritage); Kim So-hee (2023) Topic 2 (Social Context)

**3. Cross-Validation Results of the Initial Topic Set Using Heterogeneous LLMs**

확정된 7개 토픽을 초기 토픽 집합으로 정의한 뒤, 전체 데이터에 적용하기 전에 5% 표본을 대상으로 파일럿 검증을 수행하였다. 이 단계는 이미 구성된 토픽 집합에 문헌을 단순히 재분류하여 기존 토픽 중 하나로 배정되는지를 확인하기 위한 절차가 아니다. 본 연구에서 5% 파일럿 검증의 목적은 첫째, 초기 7개 토픽 체계가 실제 문헌의 주제적 다양성을 어느 정도 포괄하는지 확인하는 것, 둘째, 토픽 할당 과정에서 낮은 confidence 또는 evidence-label 불일치 사례가 발생하는지 점검하는 것, 셋째, 기존 토픽 체계로 설명하기 어려운 NEW\_TOPIC 후보가 나타나는지 확인하는 데 있다. 즉, 이 단계는 초기 토픽 체계의 적용 가능성과 경계 사례를 사전에 점검하기 위한 예비 검증 절차로 설계되었다.

5% 파일럿 적용 결과, GPT-5.2 API[15]를 활용한 1차 토픽 할당에서는 모든 표본 문헌이 기존 7개 토픽 범주 내에서 분류되었으며 신규 토픽(NEW\_TOPIC)은 관찰되지 않았다. 이는 초기 토픽 집합이 최소 표본 범위에서는 문헌의 주요 주제 구조를 포괄하고 있음을 보여준다. 그러나 본 연구는 이 결과를 전체 코퍼스에 대한 충분한 안정성의 근거로 해석하지 않았다. 5% 표본은 초기 체계의 명백한 결함 여부를 확인하기 위한 파일럿 단계에 해당하므로, 후속 단계에서는 표본 비율을 10%로 확대하고, 신뢰도 임계값을 0.85로 설정하여 보다 보수적인 검증을 수행하였다.

10% 표본 검증은 모든 문헌에 대해 다중 모델을 일괄 호출하는 방식이 아니라, 1차 모델의 판단이 불확실하거나 기존 토픽 체계만으로 설명하기 어려운 사례에 한해 2차 모델인 Gemini 2.5 Flash API[19]를 호출하는 선택적 교차검증 방식으로 설계되었다. 이러한 설계는 전체 문헌을

반복적으로 재분류하기보다, 낮은 confidence 또는 경계 사례에 검토 자원을 집중하는 방식이다. 최근 LLM 보조 토픽 모델링[4]에서도 모호한 문서에 한정해 LLM을 선택적으로 호출하여 토픽 재할당 또는 신규 토픽 생성을 검토하는 방식이 제안된 바 있다.

10% 표본 실험에서 2차 모델이 생성한 NEW\_TOPIC 후보는 총 14건이었다. 그러나 본 연구에서는 NEW\_TOPIC 후보의 생성 여부만으로 즉시 토픽 수를 증가시키지 않았다. 각 후보에 대해 기존 토픽과의 의미 중복 여부, 기존 토픽의 하위 적용 맥락인지 여부, 독립적인 상위 범주로 기능할 수 있는지 여부를 검토하였다. 그 결과 후보 명칭은 서로 달랐으나, 의미 수준에서는 반복적으로 나타나는 소수의 상위 개념으로 수렴하였다. 특히 기존 7개 토픽 체계가 주로 적용·실천 중심의 연구를 포괄하는 반면, 10% 검증 과정에서 “연구동향 및 지식구조 분석”과 “이론·개념 및 정책 연구”는 기존 토픽의 하위 변형으로 흡수하기 어려운 독립 범주로 판단되었다. 이에 따라 본 연구는 두 범주를 최종 토픽 체계에 추가하였다.

다만 10% 표본 이후 표본을 더 확대할 경우 세부적인 NEW\_TOPIC 후보가 추가로 생성될 가능성을 완전히 배제할 수는 없다. 따라서 본 연구는 10% 표본 검증을 전체 코퍼스에 대한 이론적 포화의 확정으로 해석하지 않고, 초기 토픽 체계의 구조적 보완을 위한 단계적 검토 절차로 한정하였다. 최종 분석 단계에서는 새로운 토픽 생성을 더 이상 허용하지 않고, 확정된 9개 토픽 체계를 전체 코퍼스에 일관되게 적용하였다. 이는 연구 동향 분석에서 모든 문헌을 동일한 비교 틀 안에 배치하기 위한 것으로, 본 연구의 초점은 토픽 수의 무제한 확장이 아니라 기존 연구 [8-12], LDA, BERTopic, 이종 LLM 검토, 연구자 판단을 종합하여 해석 가능하고 비교 가능한 토픽 체계를 구성하는 데 있다.

#### 4. Finalization of the Topic and Its Application to the Entire Corpus

따라서 초기 7개 토픽 체계는 교차검증 결과를 반영하여 최종적으로 9개 토픽 체계로 확장하여 확정되었다. 최종 토픽 체계를 확정된 이후에는 전체 문헌을 대상으로 전수 적용을 수행하였다. 이 단계에서는 새로운 토픽 (NEW\_TOPIC)의 생성을 허용하지 않고, 모든 문헌이 확정된 9개 토픽 중 하나에 반드시 할당되도록 설계하였다. 이 단계의 목적은 추가적인 토픽 탐색이 아니라, 확정된 토픽 체계를 전체 코퍼스에 일관되게 적용하여 분포를 산출하고 후속 비교 분석이 가능한 기반 데이터를 구축하는 데

있다. 이러한 강제 할당 방식은 전체 분포 계산, 시기별 변화 분석, 토픽 간 비중 비교, 추세 분석을 가능하게 하며, 후속적인 Hot Topic 및 Cold Topic 분석을 수행하기 위한 기초 자료를 제공한다[20]. 또한 토픽 할당 과정에서는 confidence(분류 확신도)와 evidence\_used(분류 근거)를 함께 저장하여, 분류 결과에 대한 품질 점검과 인간 연구자의 사후 검토가 가능하도록 하였다.

최종 전처리 단계에서 확보한 1,174편 중, 초록의 정보량이 부족하여 토픽 할당 근거를 확인하기 어려운 문헌과 추가 확인된 중복 문헌 50편을 제외하였다. 이에 따라 최종 토픽 분포 분석에는 1,124편을 사용하였으며, 그 결과는 Table 8에 제시하였다. 새롭게 추가된 두 개의 토픽, 즉 “메타버스 교육의 이론·개념 및 정책 연구”와 “메타버스 교육 연구 동향 및 지식 구조 분석”은 각각 12.40%와 3.00%를 차지하였다. 이는 초기 7개 토픽 체계가 주로 적용·실천 중심 연구를 포괄하고 있었던 반면, 교차검증 과정을 통해 이론·개념 및 정책 논의, 그리고 연구동향·지식 구조 분석과 같은 연구가 별도의 상위 범주로 분리될 필요가 있었음을 보여준다.

Table 8. Results of Topic Distribution

Topic Name	No. of Papers	Ratio
Application of School-Based and Subject-Specific Instruction	315	28.00%
Verification of Educational Effectiveness and Learning Outcome Research	247	22.00%
Theoretical, Conceptual, and Policy Research on Metaverse Education	139	12.40%
Design of Metaverse Educational Platforms and Systems	135	12.00%
User Experience (UX) and Immersion-Based Learning Research	93	8.30%
Special Education and Vocational/Technical Training	74	6.60%
Socio-Cultural Contexts and Digital Inclusion	57	5.10%
Analysis of Research Trends and Knowledge Structure in Metaverse Education	35	3.00%
Health-Medical and Rehabilitation Education	29	2.60%
	1,124	100

연도별 토픽 분포의 변화 추이는 Fig. 5에 제시하였다.

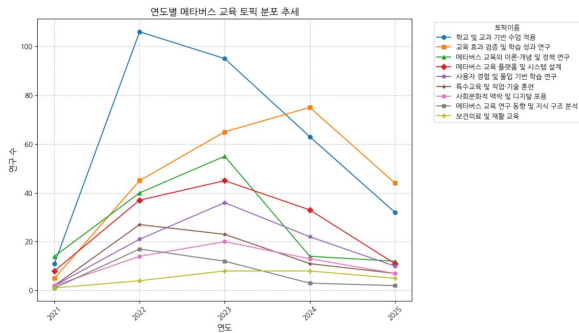


Fig. 5. Yearly Trends in Topic Distribution

연도별 9개 토픽의 분포를 비교한 결과, 특정 토픽이 단 기간에 급격히 부상하거나 급격히 쇠퇴하는 뚜렷한 Hot/Cold 현상은 확인되지 않았다. 상위 비중을 유지한 토픽은 대체로 “학교 및 교과 기반 수업 적용”과 “교육 효과 검증 및 학습 성과 연구”였으며, 연도에 따라 전체 연구 수의 증감은 존재하였으나 상대적 순위 구조 자체가 크게 재편되지는 않았다. 이는 메타버스 교육 연구가 단기적 유행에 의해 급변하기보다는, 비교적 안정적인 핵심 연구 축을 중심으로 점진적으로 확장되어 왔음을 시사한다.

한편, “메타버스 교육 플랫폼 및 시스템 설계”나 “보건의료 및 재활 교육”과 같은 일부 응용 영역은 특정 시기에 일시적 증가를 보였으나, 이후 다시 완만한 감소세를 나타내며 연구 구조 전반을 재편할 정도의 전환으로 이어지는 않았다. 중·하위권 토픽들 역시 소폭의 증감은 관찰되었으나, 장기적 소멸이나 급격한 부상으로 해석할 수준의 변화는 나타나지 않았다. 종합하면, 메타버스 교육 연구는 연구 패러다임의 급격한 이동보다는, 수업 적용-효과 검증-플랫폼 설계라는 기본 구조를 유지한 채 세부 응용 영역이 조정되는 안정화 국면에 접어든 것으로 해석할 수 있다.

### V. Conclusion

본 연구는 메타버스 교육 연구를 대상으로, LLM 기반 구조화 메타데이터 생성, LDA와 BERTopic의 병렬 적용, 선행 토픽모델링 연구[8-12]와의 비교·통합, confidence 기반 선택적 이중 LLM 교차검증, 그리고 인간 전문가 검토를 결합한 반복적 토픽모델링 절차를 제안하였다. 이는 단일 토픽모델의 산출 결과를 그대로 수용하는 대신, 교육 연구의 맥락을 반영하는 구조화된 입력과 다중 근거를 활용하여 해석 가능하고 비교 가능한 토픽 체계를 구성하기 위한 설계이다. 전수 적용 결과, 메타버스 교육 연구는 9 개의 상위 토픽으로 분류되었다. “학교 및 교과 기반 수업 적용”과 “교육 효과 검증 및 학습 성과 연구”가 가장 큰

비중을 차지하였으며, “메타버스 교육의 이론·개념 및 정책 연구”, “메타버스 교육 플랫폼 및 시스템 설계”, “사용자 경험 및 몰입 기반 학습 연구”가 주요 토픽으로 확인되었다. 이는 메타버스 교육 연구가 기술 구현이나 플랫폼 활용 논의에 머무르지 않고, 수업 적용, 효과 검증, 이론·정책 논의, 사용자 경험, 특수교육, 직업훈련, 사회문화적 맥락 등으로 다층화되어 왔음을 보여준다. 또한 연도별 추이 분석에서는 급격한 패러다임 전환보다는 수업 적용-효과 검증-플랫폼 설계라는 핵심 축을 중심으로 세부 응용 영역이 점진적으로 조정되는 안정화 양상이 관찰되었다.

본 연구의 기여는 다음 세 가지로 요약된다. 첫째, 문헌의 제목과 초록을 연구대상, 연구분야, 연구방법의 세 차원으로 구조화함으로써, 단어 빈도 중심의 분석에서 포착하기 어려운 교육 연구의 맥락적 속성을 토픽모델링 입력 단계에 통합하였다. 둘째, LDA의 확률 기반 토픽 구조와 BERTopic의 임베딩 기반 의미 군집을 선행 연구의 토픽 체계와 대조함으로써, 단일 모델 의존에서 발생할 수 있는 구조적 편향을 완화하고 학문적 연속성을 유지하는 통합 토픽 체계를 도출하였다. 셋째, 단계적 표본 확대 (5%·10%)와 confidence 기반 선택적 이중 LLM 교차검증, NEW\_TOPIC 후보 검토, 전문가 검토를 결합하여, 토픽모델링을 일회적 군집 산출이 아니라 반복적 의미 수렴 과정으로 확장하는 구체적 절차를 제시하였다.

다만 본 연구에는 몇 가지 한계가 있다. 첫째, 분석 대상이 RISS와 KCI 중심으로 구성되어 국제 데이터베이스 문헌을 충분히 포괄하지 못하였다. 둘째, LLM 기반 메타데이터 생성과 토픽 할당 결과는 프롬프트 설계, 모델 버전, 라벨 체계에 따라 일부 변동이 발생할 수 있다. 셋째, LDA와 BERTopic은 서로 다른 생성 원리를 가지므로 두 모델의 결과를 동일한 기준으로 정량 비교하기에는 제약이 있다. 넷째, 본 연구는 10% 표본 검증 단계에서 NEW\_TOPIC 후보가 소수의 상위 개념으로 수렴하는 시점을 정제 종료 기준으로 삼았으나, 이를 이론적 포화의 확정으로 해석하지는 않았으며 표본 규모를 추가로 확대한 검증은 수행하지 않았다.

향후 연구에서는 Web of Science, Scopus 등 국제 데이터베이스를 포함하여 분석 범위를 확장하고, 다양한 LLM 및 임베딩 모델 간 비교를 통해 본 절차의 일반화 가능성을 검토할 필요가 있다. 또한 인간 전문가 판단과 LLM 판단의 차이를 정량적으로 비교하고, 본 연구에서 제안한 구조화 메타데이터 기반 반복적 토픽모델링 절차를 다른 교육 연구 분야에 적용함으로써 방법론의 확장 가능성과 적용 범위를 검증할 수 있을 것이다.

## ACKNOWLEDGEMENT

This study was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea in 2021 (NRF-2021S1A3A2A01090926).

## REFERENCES

- [1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, Jan. 2003.
- [2] C. M. Pham, A. Hoyle, S. Sun, P. Resnik, and M. Iyyer, "TopicGPT: A Prompt-based Topic Modeling Framework," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2956-2984, Mexico City, Mexico, June 2024. DOI: 10.18653/v1/2024.naacl-long.164
- [3] M. Grootendorst, "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure," *arXiv*, pp. 1-21, Mar. 2022. [arXiv:2203.05794](https://arxiv.org/abs/2203.05794)
- [4] C.-H. Chang, J.-T. Tsai, Y.-H. Tsai, and S.-Y. Hwang, "LITA: An Efficient LLM-assisted Iterative Topic Augmentation Framework," *Advances in Knowledge Discovery and Data Mining*, pp. 545-558, May 2025. DOI: 10.1007/978-981-96-8170-9\_35
- [5] X. Yang, H. Zhao, W. Xu, Y. Qi, J. Lu, D. Phung, and L. Du, "Neural Topic Modeling with Large Language Models in the Loop," *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 1296-1313, July 2025. DOI: 10.18653/v1/2025.acl-long.70
- [6] M. Pradana, E. Elisa, S. N. Syarifuddin, and others, "Metaverse in Education: A Systematic Literature Review," *Cogent Social Sciences*, Vol. 9, No. 1, 2023. DOI: 10.1080/23311886.2023.2252656
- [7] B. Kye, N. Han, E. Kim, Y. Park, and S. Jo, "Educational Applications of Metaverse: Possibilities and Limitations," *Journal of Educational Evaluation for Health Professions*, Vol. 18, pp. 32, Nov. 2021. DOI: 10.3352/jeehp.2021.18.32
- [8] J. H. Lim, J. P. Hong, J. M. Park, and M. R. Ahn, "Analysis of Domestic Educational Research Trends Using Metaverse and Virtual Worlds from 2010 to 2021: Applying LDA-Based Topic Modeling and Time Series Regression," *Journal of Educational Information and Media*, Vol. 28, No. 2, pp. 187-214, June 2022.
- [9] S. Han, and T. Kim, "Analysis of Domestic Metaverse Education Research Trends From 2020 to 2022: Focusing on Topic Modeling," *Culture and Convergence*, Vol. 45, No. 4, pp. 123-141, Apr. 2023.
- [10] S. H. Kim, "Analysis of Overseas Research Trends in Education Using Virtual Reality: Using LDA-Based Topic Modeling," *Journal of Learner-Centered Curriculum and Instruction*, Vol. 23, No. 24, pp. 365-381, Dec. 2023. DOI: 10.22251/jlcci.2023.23.24.365
- [11] D. Kim, and T. Im, "A Systematic Review of Virtual Reality-Based Education Research Using Latent Dirichlet Allocation: Focus on the Topic Modeling Technique," *Applied Bionics and Biomechanics*, Vol. 2022, Article ID 1201852, pp. 1-11, Aug. 2022. DOI: 10.1155/2022/1201852
- [12] H. Ozyurt, and O. Ozyurt, "Decoding Educational Augmented Reality Research Trends: A Topic Modeling Analysis," *Education and Information Technologies*, Vol. 30, No. 1, pp. 57-87, Jan. 2025. DOI: 10.1007/s10639-024-12943-1
- [13] J. Liu, Z. Shang, W. Ke, P. Wang, Z. Luo, J. Liu, G. Li, and Y. Li, "LLM-Guided Semantic-Aware Clustering for Topic Modeling," *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 18420-18435, Vienna, Austria, July 2025. DOI: 10.18653/v1/2025.acl-long.902
- [14] T. Çelikten, and A. Onan, "Topic Modeling through Rank-Based Aggregation and LLMs," *Knowledge-Based Systems*, Vol. 314, Article 113219, Apr. 2025. DOI: 10.1016/j.knsys.2025.113219
- [15] OpenAI, "GPT-5.2 Model," *OpenAI API Documentation* <https://developers.openai.com/api/docs/models/gpt-5.2>
- [16] J. R. Landis, and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, Vol. 33, No. 1, pp. 159-174, March 1977. DOI: 10.2307/2529310
- [17] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 399-408, Shanghai, China, Feb. 2015. DOI: 10.1145/2684822.2685324
- [18] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation," *arXiv*, pp. 1-19, Feb. 2024. DOI: 10.48550/arXiv.2402.03216
- [19] Google Cloud, "Gemini 2.5 Flash," *Generative AI on Vertex AI Documentation* <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash>
- [20] M. Karabacak, H. A. Vishal, R. Khalaf, A. A. Ghanem, A. N. Sebastian, and C. B. Goodwin, "Natural Language Processing Reveals Research Trends and Topic Evolution in The Spine Journal: A Bibliometric Analysis," *The Spine Journal*, Vol. 24, No. 2, pp. 245-254, Feb. 2024. DOI: 10.1016/j.spinee.2023.10.001

## Authors



Wan-Je Gil received the B.A. degree in German Language and Literature from Yonsei University, Korea, in 1995, the M.Ed. degree in Robot Education from Gongju National University of Education, Korea, and the Ph.D.

degree in Computer Engineering from Kongju National University, Korea, in 2023. Since 2020, he has been serving as an adjunct professor in the Graduate School of Education at Dongguk University, majoring in AI Convergence Education. His research interests include AI-integrated education, educational data science, and generative AI.



In-Soo Shin received the B.A. degree in Education from Yonsei University, Seoul, Korea, and the M.S. and Ph.D. degrees in Educational Measurement and Statistics from Florida State University, Florida, USA.

Dr. Shin joined the faculty of the Graduate School of Education at Dongguk University, Seoul, Korea, where he is currently a Professor in the Department of AI Convergence Education. He also serves as a faculty member in the Department of Educational Data Science at the Graduate School of Educational Service Science. His research interests include meta-analysis, AI-integrated education, and educational data science.