JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 7, pp. 141-150, July 2026
https://doi.org/10.9708/jksci.2026.31.07.141

A High-Density Compressed Text Topic Modeling Scheme Based on Large Language Models: Focusing on Classical Chinese Poetry

Byeong-chan Lee*

*Lecturer, Dept. of Sino-Korean Literature, Chungnam National University, Daejeon, Korea

[Abstract]

To resolve over-segmentation in modeling highly compressed texts like Classical Chinese poetry, we propose High-Density Compressed Text Topic Modeling (HDCTM). Unlike traditional models that extract keywords post-clustering, HDCTM utilizes a Large Language Model (LLM) at the pre-embedding stage. It extracts structured 3-axis metadata (Subject, Type, Emotion) as semantic anchors, which are fused with raw texts to construct a knowledge-augmented embedding space. The architecture executes a hybrid two-track process: top-down structural integration via entry-point sample extraction, and bottom-up contextual discovery driven by micro-level multi-sampling and macro-level geometric centroid verification. Evaluated on 5,247 classical poems, HDCTM successfully resolves semantic fragmentation, establishing a robust, explainable analytical framework for computational humanities.

▸ **Key words**: Topic Modeling, Large Language Models, Classical Chinese Poetry, Compressed Text, Multi-sampling, Centroid Verification, HDBSCAN, Computational Humanities

[요 약]

본 논문에서는 고전 한시와 같이 고도로 압축된 텍스트 분석 시 발생하는 기존 토픽 모델링의 파편화 문제를 해결하기 위해 '고밀도 압축 텍스트 토픽 모델링(HDCTM)' 기법을 제안한다. 본 기법은 분석 초기 단계에 대형 언어 모델(LLM)을 배치하여 '주제, 유형, 정서'의 3축 키워드를 추출하고, 이를 한시 원문과 결합하여 풍부한 문맥을 보존하는 '지식 증강형(Knowledge-Augmented) 임베딩' 공간을 구축한다. 이후 구조화된 3축 스키마 기반의 하향식(Top-Down) 구조 통합 경로와, 미시적 다중 표본 추출(Multi-sampling) 및 거시적 무게중심(Centroid) 검증에 기반한 상향식(Bottom-Up) 자율 군집 경로를 병행 전개하는 하이브리드 투트랙 방식을 적용한다. 김창흡의 한시 5,247편 전체를 대상으로 검증한 결과, 제안 기법은 기존 통계적 방식의 한계를 극복하고 설명력을 확보하여 전산인문학 분석 도구로서의 타당성을 입증하였다.

▸ **주제어**: 토픽 모델링, 대형 언어 모델, 고전 한시, 압축 텍스트, 다중 표본 추출, 무게중심 검증, 밀도 기반 군집화, 전산 인문학

• First Author: Byeong-chan Lee, Corresponding Author: Byeong-chan Lee
*Byeong-chan Lee (lbcde@hanmail.net), Dept. of Sino-Korean Literature, Chungnam National University
• Received: 2026. 06. 09, Revised: 2026. 06. 26, Accepted: 2026. 07. 07.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

토픽 모델링 기법의 고도화 과정에서 자연어 처리의 중심축은 단순 단어 빈도 기반 분류에서 문맥적 관계를 보존하는 고차원 벡터 변환 기술로 이동해 왔다. 잠재 디리클레 할당(LDA)과 같은 전통적 확률 기반 기법들은 코퍼스 내 어휘의 통사적 분포를 통해 잠재적인 의미 구조를 추론해 왔으며, 최근의 딥러닝 기반 기법들은 사전 학습된 임베딩 표현력을 군집 알고리즘과 결합하여 대규모 문서군을 분류하는 방식을 취하고 있다[1], [2]. 이러한 기법들은 일반적인 장문의 평문 텍스트 환경에서는 의미론적 일관성을 지닌 주제망을 형성하는 데 일정한 효율성을 입증하였다.

그러나 고전 한시(Classical Chinese Poetry)와 같이 제한된 수의 표의문자 내에 복합적인 비유와 역사적 고사(Allusion)를 내포하는 '고밀도 압축 텍스트'를 다룰 때는 기존 모델들의 한계가 뚜렷하게 드러난다. 첫째, 텍스트가 짧고 단어 공생(Co-occurrence) 빈도가 희소하여, 통계 기반 모델은 고유 의미 정보를 노이즈로 오인해 유실시킨다. 둘째, 사전 의미 보강 없이 문맥 임베딩의 수치적 거리에만 의존하는 밀도 기반 군집화(HDBSCAN 등)[3]를 적용할 경우, 기하학적 차원 축소 시 은유적 관계가 평면화되면서 데이터가 이상치(Outlier)로 소실되거나 과도하게 분할되는 파편화(Fragmentation) 현상이 발생한다.

무엇보다 분석 대상 문헌의 의미론적 경계가 모호한 상태에서 분석가가 임의로 토픽 개수(K)를 강제하는 방식은 학술적 객관성을 확보하기 어렵다. 또한 생성형 거대 언어 모델(LLM)을 군집화 완료 이후 레이블 명명 단계에만 제한적으로 도입하는 사후적(Post-hoc) 보완 기법 역시, 초기 군집 단계에서 이미 유실된 다의적 은유 맥락을 근본적으로 복원하지 못한다는 한계를 지닌다.

이러한 파편화를 방지하고 분석의 일관성을 확보하기 위해, 본 연구에서는 거대 언어 모델의 사전 의미 해석 특징(주제·유형·정서)을 고속 추출하여 기하학적 분석과 유기적으로 바인딩하는 고밀도 압축 텍스트 토픽 모델링(HDCTM) 아키텍처를 제안한다. 1단계로 코퍼스 원문 단독 소스([제목] + [본문]) 기반의 기초 임베딩을 구동함과 동시에, LLM의 심층 정독 필터링을 병행하여 3축 개념 범주를 입체적으로 수립한다. 이때 도출된 핵심 키워드 메타데이터는 임베딩 벡터 공간의 수치적 정렬을 지지하는 의미론적 앵커(Semantic Anchor) 역할을 동적으로 수행한다.

반면, 정성적 '추출 근거(Reasoning)' 데이터는 토큰 오버헤드와 기하학적 무작위 변동 변수를 제어하기 위해 임베딩 연산에서는 의도적으로 배제(Stable Isolation)하되, 최종 단계에서 서지학적 검증을 지원하는 설명 가능한 AI(XAI) 분석 도구로 활용할 수 있도록 별도 저장한다. 이후 제어 자유도(Degree of Freedom, DoF) 수준에 따라 상향식(Bottom-Up) 자율 군집 경로와 하향식(Top-Down) 스키마 통합 경로가 독립적으로 작동하는 하이브리드 투트랙 방식을 실증한다.

본 알고리즘의 유효성 검증을 위해 조선 시대 문인 중 한시 창작 편수가 가장 많고 질적 완성도가 높은 삼연 김창흡의 한시 5,247편 전수를 분석 대상으로 선정하였다. 이는 대규모 고밀도 압축 텍스트를 처리하는 제안 프레임워크의 성능을 객관적으로 평가하기에 최적화된 데이터셋이다. 파이프라인 구현 코드와 전체 분석 결과는 공개 저장소(https://github.com/Byoung-chan/HDCTM---)에 개방한다.

본 논문의 구성은 다음과 같다. 2장에서는 토픽 모델링의 한계와 LLM 임베딩 특성을 고찰하고, 3장에서는 HDCTM의 세부 설계(3축 특징 추출, 지식 증강형 임베딩, DoF 기반 분기 경로)를 제시한다. 4장에서는 김창흡 한시 전수 대상의 실증 실험 및 상·하향식 트랙 간 정합성 검증 결과를 논의하며, 5장에서 결론을 맺는다.

II. Preliminaries

2.1 Latent Dirichlet Allocation (LDA)

잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)은 대규모 문서 집합 기저에 깔린 잠재적 단어-토픽-문서 간의 결합 확률 분포를 베이시안 추론으로 파악하는 생성적 확률 모델이다[1]. 대형 사전학습 언어 모델의 연산 복잡도와 비교하여 뛰어난 스크리닝 속도를 제공하기 때문에 문헌학 및 전산인문학 전반에서 코퍼스의 거시 트렌드를 규명하는 기법으로 널리 차용되고 있다[4].

그러나 LDA는 기본적으로 단어의 출현 빈도와 표면적 확률 정렬에만 의존하기 때문에, 문장의 길이가 극도로 짧고 단어 간의 공생 관계 정보를 충분히 학습할 수 없는 희소 데이터셋(Sparse Dataset)에서는 의미론적 일관성(Coherence)이 급격히 붕괴한다. 특히 품사 변용(詞類活用)과 다의성을 지닌 고전 한시 도메인에서 이러한 한계가 심화되며, 이는 단어의 단순 빈도만으로는 한자 본연의 복합적 의미망을 포착할 수 없어 이를 별도의 지식 구조(Ontology)로 규명하고자 했던 선행 연구[5]의 문제의식과도 궤를 같이한다. 빈도 기반 방식은 대구(Parallelism)와

압운(Rhyme) 등 문맥을 유실하고, '산(山)', '강(江)' 등 고빈도 어휘 중심으로 토픽을 과도하게 일반화하는 한계가 있다.

2.2 Embedding-based Topic Models

딥러닝 기술의 발전과 함께 고정 빈도 기법의 문맥 소실 한계를 극복하기 위해 BERTopic 등의 신경망 기반 토픽 모델이 제안되었다[2]. BERTopic은 BERT 계열의 인코더 모델을 활용하여 문장 혹은 문서 단위의 동적 벡터 공간을 구축한 뒤, UMAP을 활용한 매니폴드 차원 축소 및 HDBSCAN 밀도 기반 군집화를 통해 노이즈를 배제한 순수 군집을 추출한다[3],[6].

신경망 기반 토픽 모델은 인코더 표현력으로 단어 빈도 한계를 극복하며, Bianchi 등[7]은 이를 결합해 토픽 일관성을 높였다. 그러나 밀도 기반 군집 모델은 어휘 왜곡 우려가 있고, 군집 형성 후 c-TF-IDF로 키워드를 추출하는 사후적(Post-hoc) 한계로 인해 다의적 맥락 복원이 불가하다[2].

따라서 본 연구는 군집 전 단계(Pre-embedding Stage)에 LLM 해석 자질을 결합하여 기하 공간의 평면화를 방지한다.

2.3 Transition to LLM-Augmented Topic Models

LLM을 전체 코퍼스에 직접 투입했을 때 발생하는 문제점(비용 문제와 성능 문제)을 두 문장으로 명확히 분리하여, 본 연구(HDCTM)의 제안 배경을 더욱 강하게 부각하는 방식이다. 최근 대형 언어 모델(LLM)을 활용해 추론 성능을 이식하려는 시도가 있다. Pham 등[8]의 TopicGPT는 프롬프트 기반 토픽 설명력을 격상했고, Mu 등[9]은 LLM의 전통적 토픽 모델 대안 가능성을 실증했다. 그러나 이러한 제로샷 기반의 전수 분류 방식을 대규모 코퍼스에 직접 투입할 경우, 막대한 토큰 처리 비용이 발생하여 효율성이 급감한다[10]. 또한 생성 모델 특유의 환각(Hallucination) 및 출력 무작위성으로 인해 전체 문서에 대한 레이블링 일관성을 훼손하는 치명적인 단점이 있다[11].

2.4 Recent Trends and Limitations

최근 전산인문학 및 정보검색 분야에서는 단문(Short text) 및 고전 시가 도메인의 의미론적 한계를 극복하기 위해 대형 언어 모델과 기존 토픽 모델링을 결합하려는 연구가 다각도로 전개되고 있다. 대표적으로 GLMTopic [12]은 사전 학습된 생성형 언어 모델과 공동체 강화 그래프 임베딩(Community-enhanced Graph Embedding) 아키텍처를 하이브리드로 결합하여 비영어권 단문 텍스트의 토픽 일관성을 개선하고자 하였다. 그러나 현대 백화문 중심의 범용 모델은 고대 한어의 고유한 통사 구조와 고사(Allusion)를 처리할 때 심각한 도메인 불일치 문제를 일으키며 의미론적 잡음을 양산한다.

이를 해결하기 위해 Yao 등은 고전문헌 특화 거대 언어 모델인 WenyanGPT를 제안하였다. 이 연구는 방대한 고전 문헌 코퍼스로 모델을 미세 조정하고 전용 평가 벤치마크를 구축함으로써, 기존 백화문 인코더 모델이 고전 시가 환경에서 보인 정보 손실 한계를 극복하고 의미 해석의 정확성을 유의미하게 향상시켰다[13].

고전 문헌을 자연어 처리(NLP) 환경에서 다룰 때, 한자는 단순한 문자 기호가 아니라 풍부한 역사적 맥락과 개념적 지식을 내포한 자원이다. S. K. Hsieh(2006)는 이러한 한자의 특성에 주목하여, 한자의 구조적·의미적 속성을 컴퓨터가 이해할 수 있는 지식 구조로 구조화하고 이를 언어 자원(Ontology)인 'HanziNet'으로 구축하는 프레임워크를 제안하였다[5].

또 다른 접근법인 Peng의 연구[14]는 단문 한시 데이터셋의 희소성(Data Sparsity) 문제를 완화하기 위해 외부 웹 서비스 구조 하에서 글자 임베딩(Character Embedding)과 TextRank 알고리즘을 결합한 보조 의미 정보 강화 메커니즘을 제안하였다. 해당 방법론은 고전 한문 코퍼스로 학습된 글자 벡터 간 코사인 유사도를 바탕으로 유사도 행렬(Similarity Matrix)을 구축한 뒤, TextRank 연산을 통해 개별 문서 내 글자의 상호 중요도인 '정렬된 글자 분포(Ranked Character Distribution)'를 도출한다. 이후 이를 글자의 조정된 자질로 정의하고, 잠재 디리클레 할당(LDA) 기반 깁스 샘플링(Gibbs Sampling)의 조건부 확률식에 가중치 파라미터로 직접 통합하여 토픽 추출의 정밀도를 보정하는 프레임워크를 수립하였다. 그러나 이와 같은 웹 서비스 기반 기법은 외부 임베딩과 그래프 기반 중요도를 확률 연산에 주입하였음에도 불구하고, 본질적으로는 개별 단어 및 글자의 출현 빈도와 통계적 결합 분포에 의존하는 단어 주머니(Bag-of-Words) 가설의 한계에서 벗어나지 못한다. [14].

또한, 일부 연구에서는 한시의 함축적 의미를 보존하기 위해 기계 번역 및 역번역(Back-translation) 필터를 활용하여 맥락을 인위적으로 확장한 뒤 토픽 분석을 전개하는 방식을 채택하기도 하였다[15]. 그러나 시적 의도(Poetic Intent)가 기계 번역의 번역 왜곡(Translation Paradox) 과정에서 심각하게 훼손되며, 오히려 번역기에 의해 유도된 일반 문장 형태의 현대어 해설이 고전 한시

고유의 미시적 감정선과 역사적 지시 대상(Referents)을 단순화하거나 왜곡하는 결과를 초래한다[15].

이러한 한계는 고전 문헌을 정보가 누락된 '결손 데이터'로 인식하여 인위적으로 본문을 팽창시키거나 정적인 외부 사전에만 결합하려는 관점에서 비롯된다. 이에 본 연구는 원문 본연의 수사적 압축 구조를 보존(Preservation)하는 동시에, LLM의 사전 추론 능력을 통해 '의미론적 힌트(Semantic Anchor)'인 3축 키워드를 능동적으로 융합하는 '지식 증강형 임베딩(Knowledge Augmented Embedding)' 파이프라인을 제안한다. 이는 원문의 통사적 긴장감을 유지하면서도, 압축 텍스트의 의미 구조를 다차원 분해 및 정렬할 수 있는 구조적 해법을 제공한다.

III. Proposed Hybrid Topic Modeling Architecture

3.1 Phase 1: LLM-Based Keyword Extraction via 3-Axis Framework

본 연구에서 제안하는 HDCTM 아키텍처의 전체 파이프라인은 Fig. 1과 같다.

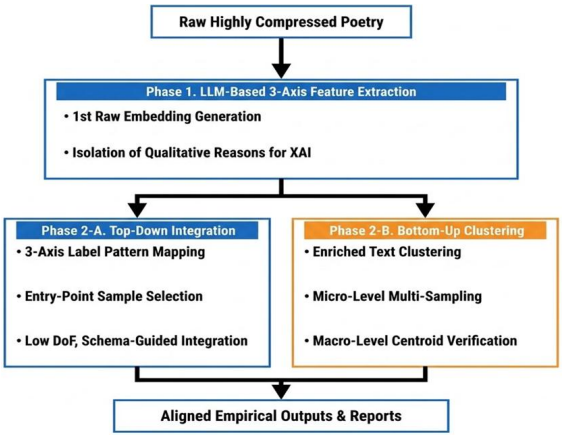

Fig. 1. System Pipeline of HDCTM Strategy

본 제안 HDCTM 아키텍처의 최우선 가설은 기존 통계 기법들의 맹점인 의미 소실과 파편화를 방지하기 위해 임베딩 전 단계에서 LLM을 활용한다. LLM은 원문에서 핵심 키워드와 추출 근거(Reasoning)를 선제적으로 도출한다.

이를 위해 본 시스템은 주제(Subject), 유형(Type), 정서(Emotion)라는 세 가지 독립적 벡터 공간으로 구성된 3축 다차원 스키마(3-Axis Classification Schema)를 사용하며, 복합 맥락 포착을 위해 단일 축 내의 다중 레이블 할당을 허용한다. 또한 대규모 API 호출 과정에서 발생할 수 있는 생성형 AI의 일시적 통신 오류나 JSON 출력 형식 파싱 에러(ValueError)로 인한 데이터 결실을 방지하기 위해 예외 처리 기반의 자동 재시도(Retry) 메커니즘을 통합 탑재하였다. 파싱 및 통신 실패 시 최대 3회(max_retries=3)에 걸쳐 동일 문헌에 대한 재분류 추론을 선제적으로 수행함으로써, 고전 문헌 코퍼스 전체의 탈락 없는 완전성과 데이터 수집의 시스템적 안정성을 확보하도록 설계하였다.

특히 본 아키텍처는 벡터 공간 임베딩 시, 한시 원문이 자아내는 통사적 긴장감과 수사적 구조(대구, 압운 등)를 유실 없이 보존하기 위해 1단계(Phase 1) 환경에서 원문 텍스트 단독 데이터([제목] + [본문])를 소스로 주입하여 1차 기초 임베딩(embedding_vector)을 생성하여 보존한다. 이후, 귀납적 탐색을 가동하는 상향식(Bottom-Up) 트랙의 진입 단계(Phase 2-B)에서 1단계의 LLM 추출 핵심 키워드를 원문과 결합한 '지식 증강형 텍스트(Enriched Text, [제목] + [all_keywords] + [본문])'로 합성하고 이를 대상으로 하여 신규 2차 임베딩(enriched embedding vector)을 동적으로 생성한다. 이는 원문 단독 임베딩 시 발생하는 초고밀도 의미 압축에 따른 파편화 현상을 방지하고, 벡터의 수학적 정렬 과정에서 개념적 일관성을 유지하기 위한 최적의 증강(Enrichment) 전략이다.

이 과정에서 LLM이 도출한 정성적 '추출 근거(Reasoning)' 데이터는 의도적으로 임베딩 연산 소스에서 완전 격리(Stable Isolation)된다. 서술형 문장 전체를 실시간 벡터 연산에 투입할 경우, 첫째로 기하급수적인 토큰 오버헤드(Token Inflation)가 발생하여 시스템의 연산 경제성을 해치며, 둘째로 생성 모델 고유의 어휘적 변동성(Variance)으로 인해 정적 임베딩 공간의 기하학적 거리 왜곡 및 경계 모호성이라는 또 다른 잡음(Noise)을 자아내기 때문이다.

밀도 기반 군집화 과정에서 이상치로 분류된 아웃라이어(b_cluster_id = -1) 데이터는 거시적 메타 대주제 보고서의 통계적 왜곡을 방지하기 위해 최종 HTML 컴파일러 연산에서 제외되며, 이상치 데이터는 CSV 파일에 분리 저장(Stable Isolation)된다. 연구자는 최종 군집에서 제외된 데이터도 해당 메타데이터를 통해 LLM의 해석 결과를 역추적(Traceback)함으로써 분류 경계를 검증하고 설명력(XAI)을 확보할 수 있다.

3.2 Phase 2-A: Top-Down Integration

추출된 3축 메타데이터 가공이 완료되면, 본 모델은 데이터 제어 자유도(Degree of Freedom, DoF)의 통제 수준에 따라 두 개의 아키텍처 경로로 분기되어 구동된다. 하향식 전문가 분류 경로(Phase 2-A)는 제어 자유도를 최소화(Low DoF)함으로써 데이터의 무작위 변동성을 제어하고 의미론적 일관성의 붕괴를 방지하는 통제형 모델이다.

이 경로에서는 자의적인 문맥 왜곡을 차단하기 위해 텍스트 기저의 정성적 추출 근거를 배제하고, 오직 규격화된 '주제·유형·정서' 레이블의 정형 조합 코드 스트링(topic_str)만을 중간 이정표(Milestone)로 투입한다. 해당 스트링을 매개로 1차 구조화된 소그룹 집단을 형성한 후, 각 소그룹의 문학적 정체성을 신속하게 스크리닝하기 위해 그룹 내 첫 번째 문헌 레코드를 진입점 표본(Entry-point Sample)으로 취하고 문맥적 정합성을 검증할 그룹 내의 작품들을 무작위 샘플링으로 대형 언어 모델(LLM)에 주입한다. 이를 통해 개별 작품군 간의 미시적 변이를 포괄하면서도 소주제별 본질을 명확히 대변하는 고밀도 개념 요약문을 효율적으로 도출한다.

그다음 본 아키텍처는 개별 작품이 지닌 미시적 벡터의 단순 물리적 결합이 아닌, LLM의 전처리로 고차원 추상화가 완료된 도출 요약문(group_summary) 텍스트 자체를 대상으로 신규 문맥 임베딩(summary_embeddings)을 수행하여 요약문 수준의 메타 추상화 공간을 생성한다. 이러한 요약문 기반 재임베딩 방식은 개별 단편 시가 노출하는 극단적인 희소 노이즈(Sparse Noise)를 기하학적으로 평탄화하고 메타 군집의 수렴 성능을 격상시키는 효과를 제공한다. 최종적으로 생성된 요약문 벡터 공간을 UMAP 알고리즘을 통해 3차원(n_components=3)의 저차원 공간으로 투영한 후, HDBSCAN 알고리즘을 적용하여 계층적 메타 군집화(Meta Clustering)를 완성함으로써 거시적 타당성을 지닌 메타 대주제(Macro-categories)로의 정밀한 계층적 수렴 정렬(Hierarchical Reintegration)을 수행한다.

3.3 Phase 2-B: Bottom-Up Clustering

반대로 상향식 데이터 주도 군집 경로(Phase 2-B)는 제어 자유도를 극대화(High DoF)하여 데이터 내재적 특징으로부터 사전에 정의되지 않은 미지의 미시 세부 군집을 탐색적으로 도출하는 도구이다. 이 트랙은 한시 원문과 1단계에서 추출된 3축 키워드가 유기적으로 결합되어, 2-B단계 진입 시 신규 임베딩 처리된 고밀도 지식 증강형 벡터(enriched_embedding_vector) 공간을 기반으로 작동한다. 초고밀도로 압축된 증강 벡터 공간을 UMAP(Uniform Manifold Approximation and Projection) 알고리즘을 통해 5차원으로 정밀하게 차원 축소한 후, HDBSCAN 알고리즘을 구동하여 사전 토픽 개수(K)의 임의 강제 없이 데이터 내재적 밀도 분포망에 따른 초기 미시 군집을 형성한다. 초기 미시 군집 형성이 완료되면, 군집 내부의 사상적 변이(Variance)와 복합적인 은유 맥락을 왜곡 없이 포착하기 위해 군집별 내부 분포에서 다중 표본 추출(Multi-sampling) 프레임을 가동한다. 추출된 다면적 작품 맥락은 대형 언어 모델(LLM)에 주입되어 군집 수준의 개념 요약문(group_summary)과 주요 키워드를 형성한다. 이러한 귀납적 텍스트 전처리는 단편 시가 지닌 정보의 희소성을 보완하고 의미론적 선명도를 확보하는 기반이 된다. 이후 도출된 미시 군집 요약문들의 텍스트 임베딩(summary_embeddings) 연산을 거쳐 UMAP 기반의 3차원 축소(n_components=3) 및 HDBSCAN에 의한 계층적 메타 군집화(Meta Clustering)를 수행함으로써, 거시적 타당성을 지닌 메타 대주제(Macro-categories)로의 수렴 정렬을 완성한다. 본 군집화의 5차원 공간과 달리 메타 추상화 공간에서는 거시적 수렴 성능을 극대화하기 위해 차원을 3차원으로 조정한다.

최종적으로 정렬된 메타 대주제 구조의 수렴 타당성(Anomalous Variance Control)과 통계적 객관성을 정량적으로 교차 실증하기 위해, 메타 군집 소속 전체 고차원 벡터들의 수학적 평균값인 기하학적 무게중심(Centroid)을 산출한다. 무게중심 유도 공식은 아래와 같다.

Centroid Equation (무게중심 유도 공식)

$$\overrightarrow{C_k} = \frac{1}{|S_k|} \sum_{v \in S_k} \vec{v}$$

여기서 S_k는 군집 k에 배정된 작품들의 고차원 임베딩 벡터 세트이며, $\overrightarrow{C_k}$는 해당 군집의 기하학적 중심 노드(*Centroid*)이다.

이 무게중심 벡터와의 코사인 유사도(Cosine Similarity) 거리가 가장 가까운 작품을 결정론적(Deterministic) '최종 대표 표본'으로 선정하고, 거리가 인접한 주변 이웃 노드(Neighboring Nodes)들을 연쇄적으로 추출한다. 추출된 표본 맥락은 LLM 기반의 최종 심층 분석 보고서 컴파일러로 이송되어 정성적 분석의 설명력(XAI)을 입증하는 문헌학적 검증(Philological Grounding) 도구로 엄밀하게 작동한다. 이는 수리적 통계 공간과 인문학적 실제 문헌 데이터 간의 정합성을 최종 검증하는 다중 스케일(Multi-scale) 통제 메커니즘이다.

IV. Experiments and Performance Evaluation

본 장에서는 제안한 고밀도 압축 텍스트 토픽 모델링(HDCTM) 아키텍처의 성능을 검증하고자 연구자가 가공한 삼연 김창흡의 한시 전체를 대상으로 실증 분석을 수행한다. 분석 대상인 한시는 은유와 전고(典故)의 밀도가 높아 기존 통계적 분류 기법을 적용하기 까다로운 코퍼스이다. 본 실험에서는 파이썬 알고리즘 구현 과정에서 활용된 주요 연산 매개변수와 도출된 통계적 구조를 논문의 세부 절과 연계하여 상세히 기술함으로써, 실험의 재현 가능성과 주제 분류의 정확도를 객관적으로 검증한다.

4.1 Experimental Setup and Baseline Collapse

제안 아키텍처의 효용성을 입증하기 위해, 은유와 철학이 고도로 응축된 김창흡(金昌翕)의 한시 5,247편 전체를 전수 가공하여 실험 데이터셋을 구축하였다. 이 작품들 대부분 최소 20자에서 56자 내외의 극단적으로 짧은 통사적 길이를 지니며, 내부에는 대구, 압운, 복합적 메타포 및 조선 중기 이후의 역사적 정치 상황에 얽힌 고사가 중첩되어 일반적인 단어 주머니(BoW) 방식으로는 분석 불가능하다. 3축(주제, 유형, 정서) 기반의 LLM 추출기와 UMAP 및 HDBSCAN 연산을 수행한 구체적인 시스템 사양과 알고리즘 튜닝 파라미터는 Table 1과 같다.

Table 1. Technical Environment & Algorithmic Parameters

Classification	Specific Metric / Library	Configured Value (Parameter)
Dataset Specs	Data Source	Institute for the Translation of Korean Classics
	Target Corpus	Classical Chinese Poetry of Samyeon Kim Chang-heup
	Total Volume	5,247 Poems
	Text Attributes	Highly compressed short text (Avg. 20 or 56 characters per poem)
Hardware Env	Computing System CPU	Intel Core i9-10900KF, 128GB RAM
Software Env	Language & Core Libraries	Python 3.11, PyTorch 2.1.1, scikit-learn 1.3.0, scipy 1.11.3
Cloud API	Gemini 2.5 Pro (Vertex AI) text-multilingual-emb-002 (768-dim)	Config: Temp=0.2, Top-P=1.0
Embedding Engine	Multilingual Encoder Model	text-multilingual-embedding-002(768-dimensions)
Dimensionality	UMAP (v0.5.3)	UMAP (v0.5.3) | n_neighbors = 15, n_components = 5 (Phase 2-B main) / 3 (Phase 2-A & Meta), metric = 'cosine'
Cluster Engine	HDBSCAN (v0.8.29)	min_cluster_size = 15 (Configured via GUI), min_samples = None (Implicitly coupled), metric = 'euclidean'
Optimization	MD5 Hashing Cache, Multi-threading Batch	Enabled (via ThreadPoolExecutor)

아울러 본 연구의 Phase 1에서 사용된 3축 분류 스키마(Subject, Type, Emotion)는 삼연 김창흡에 관한 기존 연구 성과를 직접 참조하여 설계된 것이 아니라, 고전 한시 일반에 적용 가능한 주제·유형론적 범주를 바탕으로 구성되었다. 즉, 본 연구의 메타 주제 도출은 특정 작가론의 내용을 미리 반영한 schema-guided fitting이 아니라, 일반론적 한시 분류 틀을 매개로 한 독립적 연산 결과이다. 이러한 독립적 출발점은 이후 4.5절에서 수행되는 선행 연구와의 비교가 순환 논증이 아니라 수렴 타당성(Convergent Validity) 점검으로 기능하는 방법론적 전제가 된다.

4.2 Baseline Evaluation: Over-generalization and Over-segmentation

제안 기법(HDCTM)의 도입 당위성을 입증하기 위해, 초기 탐색모델에서 LDA(K=10)와 GuwenBERT를 연동한 표준형 BERTopic(n_neighbors=30, min_cluster_size=8)을 통제 변수로 설정하여 대조 재실험을 수행하였다. 의도적으로 파라미터 기준을 낮추어(min_cluster_size=8) 최소 단위의 미시적 군집이라도 형성되도록 유도했음에도, 두 모델(LDA·BERT) 모두 정성적 의미 전개에서 심각한 구조적 붕괴를 보였다.

빈도 기반의 LDA는 '오늘(今日)', '산천(山川)' 등 고빈도 범용 형태소에 휩쓸려 개별 토픽의 고유한 문맥을 상실하고 과일반화(Over-generalization)되었다. 반면, 표준형 BERTopic 기법은 군집 수가 168개에 달할 정도로 군집 구조가 과도하게 세분화(Over-segmentation)되는 한계를 보였다. 더욱이 추출된 핵심 키워드 역시 '_天水', '不達_', '黃卷_', '君_不見' 등 서브워드(Subword) 단위로

단편화되거나 전체 문맥을 대변하지 못하는 지엽적 형태소로 조각나는 과파편화 현상을 보였다.

이러한 베이스라인 모델들의 품질 붕괴는 임베딩 인코더 모델(GuwenBERT vs Gemini) 간의 매개변수 규모 차이에서 기인한 '통제 변수 오염'의 영향이 아니다. 이는 초기 군집 형성 단계에서 학술적 지침(Schema Guide)이 부재한 채, 군집 형성 이후에 c-TF-IDF 등을 통해 사후적으로(Post-hoc) 키워드를 추출·배정하는 전통적 메커니즘 자체의 기하학적 한계이다. 특히 시(詩) 데이터 특유의 짧고 함축적인 문장 구조는 사전 정의된 가이드 없이 단순 임베딩 공간의 밀도와 수학적 거리에만 의존하면 이처럼 극심한 군집 분절을 불러오기 쉽다. 초기 군집 자체가 이미 무의미하게 조각난 상태에서는, 군집 형성 이후에 LLM을 단순 연동하여 토픽을 요약하게 만드는 '사후적 LLM 증강 모델(Post-hoc LLM-Augmented BERTopic)'을 적용하더라도 요약 기능이 정상적으로 작동할 수 없다. 본 연구가 전체 파이프라인을 전면 개편하여, 임베딩 전 단계(Pre-embedding Stage)에서 3축 스키마를 선제 추출해 '의미론적 앵커(Semantic Anchor)'로 결합하는 지식 증강형(Knowledge-Augmented) 구조를 제안한 근본적인 이유가 여기에 있다.

4.3 Phase 2-A: Top-Down Schema-Guided Integration

하향식 전문가 스키마 가이드 통합 트랙(Phase 2-A)은 1단계에서 규격화되어 추출된 주제·유형·정서의 86개 고유 조합인 '3축 범주형 템플릿 스트링(topic_str)'을 활용해 거시적 사상 정렬을 전개한다. 본 아키텍처는 원리상 주제, 유형, 정서의 다차원 3축 메타데이터를 입체적으로 조합하여 고도화된 스키마 분류를 수행할 수 있는 연역적 유연성을 보유하고 있다. 다만, 본 연구에서는 귀납적 무감독 상향식(Phase 2-B) 자율 군집 결과와의 정교한 비교 분석을 전개하고 두 경로 간의 제어 자유도(DoF) 차이를 명확하고 직관적으로 대조하기 위해, 3축 중 핵심 사상적 뼈대를 관장하는 '주제(Subject)' 차원만을 핵심 스키마의 중추로 삼아 최종 거시 메타 주제를 복원하였다.

본 경로는 데이터의 연역적 엄밀성을 고수하여 5,247편 전수를 아웃라이어 결락 없이, 전문가 스키마에 의해 분기된 소그룹 내부의 진입점 표본(Entry-point Sample) 및 관련 서사들을 상호 참조적으로 추출하여 도출한 소그룹 LLM 요약문 임베딩 공간의 HDBSCAN 계층 메타 군집화 과정을 통해 4개의 거시적 대주제(meta_cluster_id 0~3)로 수렴 정렬하였다.

삼연 김창흡 문학 세계의 가장 거대한 뿌리인 'Meta-Theme 3'(대자연 귀의와 은일 수양)은 총 3,243편으로 전체의 과반인 61.8%를 점유하는 것으로 나타났다. 이는 기사환국 이후 세속의 명리를 잊고 설악산, 벽운동 등 깊은 대자연 속에 은거하며 안빈낙도와 심성 수양에 전념했던 김창흡의 실제 역사적 삶의 궤적 및 문헌학적 Ground Truth와 정밀하게 부합하는 지표이다. 또한, 전체의 5.2%(271편) 수준으로 안전하게 격리 복원된 'Meta-Theme 2'(사회 현실 고발과 풍자)는 작가가 지녔던 조선 후기 당대 현실 비판적 자아를 왜곡 없이 수학적 공간에 표출해 냈음을 객관적으로 반증한다.

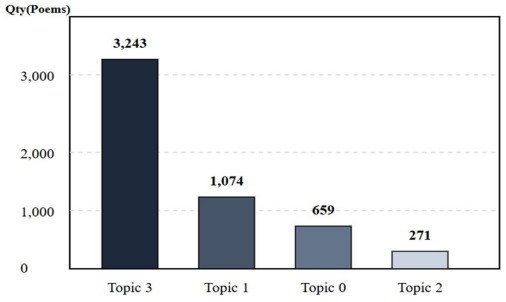

Fig. 2. Quantitative Distribution of the Top-Down Approach

4.4 Phase 2-B: Bottom-Up Data-Driven Clustering and Centroid-Based Validation

상향식 데이터 주도 군집 트랙(Phase 2-B)은 3축 스키마 제약 없이 작품 본연의 수사 구조를 보존하면서, 1단계에서 도출된 3축 키워드를 공간적 힌트인 '의미론적 앵커'로 결합한 'enriched_text' 신규 지식 증강형 임베딩(enriched_embedding_vector) 공간을 기반으로 작동한다. HDBSCAN 구동 결과, 5,247편의 한시는 1차적으로 248개의 자율적 미시 군집을 형성하였다. 미시 군집의 의미론적 선명도를 위해, 군집별 내부 분포에서 최대 10편의 작품을 다중 표본 추출(Multi-sampling)하여 LLM에 주입한다. 최종적으로 이 군집 요약문들의 텍스트 임베딩(summary_embeddings) 메타 군집화 과정을 거쳐 최종 7개의 독창적인 메타 대주제(meta_cluster_id 0~6)로 수렴 정렬되었다.

본 아키텍처의 최종 분류 정합성을 검증하기 위해, 초기 미시 군집 레벨이 아닌 최종 거시 메타 대주제 공간을 대상으로 소속 임베딩 벡터의 평균값인 '수학적 무게중심(Centroid)'과 코사인 거리를 연산하였다. 이는 수천 편 규모로 도출된 대형 군집의 정중앙에 실제로 일관된 인문학적 기류가 흐르고 있는지를 교차 실증하기 위한 다중 스케일(Multi-scale) 통제 메커니즘이다. 이때 도출된 대표작은

개별 작품 자체의 문학적 예술성을 대변하는 것이 아니라 해당 메타 군집에 배정된 작품들의 분류 정확도(Classification Accuracy)를 정량적·정성적으로 교차 검증하기 위한 의미론적 진단용 프로브(Diagnostic Probe)로 작동한다. 예컨대 기하학적 메타 무게중심과 가장 인접한 표본 노드로 식별된 '朝哭(ID: 88, Cosine Sim: 0.892)'을 기준 프로브로 설정하고, 해당 메타 군집 내에서 코사인 거리가 인접한 주변의 이웃 작품(Neighboring Nodes)들을 연쇄 추출하여 대조 분석한 결과, 이를 통해 가족 상실 및 자아 성찰과 관련된 주제적 의미와 설명적 일관성이 메타 군집 내부에서 유지되고 있음을 확인하였다.

상향식 데이터 주도 군집 트랙(Phase 2-B)은 전체 코퍼스를 빠짐없이 분류하는 완전 포괄형 분류기가 아니라, 지식 증강형 임베딩 공간에서 밀도 기반으로 응집도가 높은 핵심 의미군을 탐색하는 정밀 탐색 트랙으로 설계되었다. HDBSCAN 수행 결과, 5,247편 중 1,899편(36.2%)은 고밀도 핵심 군집으로 수렴하였고, 3,348편(63.8%)은 특정 거대 메타 주제에 강제 배분하지 않는 경계 사례로 보존되었다.

이러한 63.8%의 데이터 보존은 알고리즘의 분류 실패나 통계적 노이즈가 아니다. 이는 억지스러운 통계적 강제 수렴(Forced convergence)이 초래할 고전 시가의 다의적 은유 왜곡을 차단하기 위한 결정론적 조치이다. 코퍼스 전체의 아웃라이어 없는 거시 통합(100%)은 앞선 하향식 트랙(Phase 2-A)이 이미 완벽하게 전담하고 있으므로, 이 원화된 두 트랙은 서로를 보완하는 완벽한 상보적 시너지(Complementary Synergy)를 달성한다.

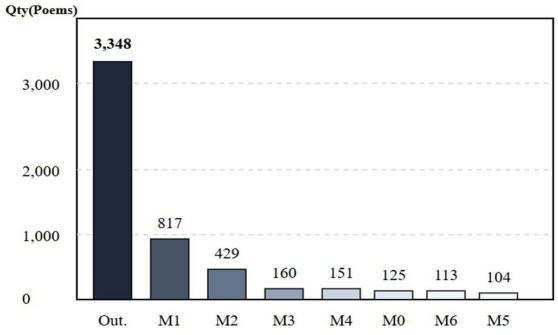

Fig. 3. Quantitative Distribution of the Bottom-Up Approach

4.5 Complementary Synergy and Concordance Analysis

Table 2는 선행 인문학 연구를 본 연구의 출력에 대한 사후적 정답표로 사용하는 검증표가 아니라, 고전 한시 일반의 주제·유형 분류론에서 출발해 독립적으로 도출된 HDCTM의 상·하향식 메타 주제 체계가 삼연 김창흡 연구의 주요 해석 테제의 수렴 타당성 기반 정합성 매핑표(Convergent Validity-based Concordance Mapping Matrix)이다.

Table 2. Convergent Validity Mapping: Independently Derived HDCTM Meta-Themes and Prior Humanities Studies on Kim Chang-heup

Concordance Mapping with Prior Humanities Studies	Phase 2-A: Top-Down (Structural Integration)	Phase 2-B: Bottom-Up (Contextual Discovery)
Expert A [16] Tension of cultivation & duty	Deductively structures conflict via Meta 2 (Social Critique) vs. Meta 3 (Nature/Cultivation).	Inductively captures tension via Meta 4 (Social Critique) & Meta 1 (Nature Hermitage).
Expert B [17] Existential conflict of hermitage & life	Categorizes macro-boundaries via Meta 1 (Rustic Peace) & Meta 2 (Social Critique).	Reconstructs conflict via Meta 2 (Reclusive Ideals) & Meta 4 (Social Critique).
Expert C [18] Fusion of travelogue & retrospection	Formally merges Meta 0 (Historical Reflection) & Meta 3 (Nature Journey).	Discovers contextual nuances via Meta 5 (Historical Space/Ruins) & Meta 1.

여기서 핵심 전제는, 본 연구의 분류 스키마와 메타 주제가 김창흡 연구의 선행 성과를 반영하여 구성된 것이 아니라는 점이다. 따라서 Shin[16], Choi[17], Kim[18]이 서로 다른 시기와 기관에서 독립적으로 제시한 김창흡 시학의 핵심 논점들이 본 연구의 메타 주제 범주 안에 안정적으로 수렴된다는 사실은, HDCTM의 분류 체계가 특정 연구 프레임을 재현한 것이 아니라 작가의 실제 시학 구조를 포괄적으로 포착했음을 시사한다.

이때 상향식(Bottom-Up)과 하향식(Top-Down)은 서로 다른 기능을 수행한다. 하향식 경로는 일반적 스키마를 바탕으로 작품군을 거시 범주로 구조화하여 시세계의 지배적 축을 안정적으로 통합한다(Structural Integration).

반면 상향식 경로는 개별 작품의 미시적 맥락 차이를 예민하게 포착하여, 동일한 거대 주제 내부에서 분화되는 세부 의미망을 드러낸다(Contextual Discovery). 두 경로가 공통적으로 김창흡 연구의 주요 성과를 수용 가능한 범위 안에 포괄한다는 점은, 분류 체계의 상호 보완성과 방법론적 정당성을 동시에 지지한다.

특히 본 연구에서 설정된 4개 및 7개의 거대 메타 주제는, 인문학 논문이 본질적으로 해당 작가의 핵심 주제 축을 중심으로 논의를 조직한다는 점을 고려할 때, 선행 연구의 논점들을 상위 범주 차원에서 흡수하는 것이 오히려 자연스러운 귀결이다.

따라서 선행 연구의 주요 테제가 이 메타 주제 체계에 포괄된다는 사실은 약점이 아니라, 해당 메타 주제 분류가 작가 연구 수준에서 충분한 내용 포괄성(Content Coverage)을 확보하고 있음을 보여주는 긍정적 지표이다. 반대로, 만약 본 연구의 메타 주제 체계가 김창흡 시학의 실제 구조를 제대로 반영하지 못했다면, 기존 인문학 연구의 핵심 논점들 가운데 일부는 이 범주 바깥으로 이탈하거나 부적절하게 환원되었을 것이다. Table 3의 결과는 그러한 체계적 이탈이 관찰되지 않음을 보여주며, 이는 HDCTM 분류 체계의 수렴 타당성(Convergent Validity)을 지지하는 중요한 근거가 된다.

V. Conclusions

본 연구는 고전 한시와 같은 고밀도 압축 텍스트에서 발생하는 기존 토픽 모델링의 과파편화 한계를 극복하기 위해, 임베딩 전 단계(Pre-embedding)에 대형 언어 모델(LLM)의 3축 지식을 선제적으로 결합한 '고밀도 압축 텍스트 토픽 모델링(HDCTM)' 아키텍처를 제안하고 실증하였다. 이를 통해 기존 사후적(Post-hoc) 키워드 추출 방식이 유발하는 은유 왜곡을 제어하고, 추출 근거(Reasoning)의 물리적 격리를 통해 벡터 공간의 의미론적 선명도를 확보하였다.

삼연 김창흡의 한시 5,247편을 전수 분석한 결과, 본 아키텍처는 '거시적 주제 통합'과 '자유로운 미시 주제 탐색'이라는 연구자의 두 가지 핵심 의도를 완벽한 상보적 시너지(Complementary Synergy)로 구현해 내었다. 하향식(Top-Down) 트랙은 코퍼스 전체를 단 하나의 아웃라이어(0%) 결락 없이 거시적 뼈대로 100% 구조 통합하여 코퍼스 차원의 포괄성을 확보하였다. 반면 상향식(Bottom-Up) 트랙은 전체를 억지로 분류하려는 통계적 강제 수렴을 배제하고, 자유도가 높은 환경에서 자연스럽게 묶이는 1,899편(36.2%)의 고밀도 핵심 의미군만을 정밀하게 발견해 내었다. 이때 분류 보류된 3,348편(63.8%)의 데이터는 알고리즘의 실패가 아니라, 억지스러운 메타 주제 배분에 따른 인문학적 해석 왜곡을 방지하기 위한 결정론적 경계 사례 보존이다.

결과적으로 하향식의 '전체 구조 통합'과 상향식의 '자유로운 맥락 발견'이 이원화되어 작동하면서도 동일한 작가론적 핵심 논점을 안정적으로 포괄한다는 사실은, 본 체계가 단순한 통계적 분류기를 넘어 고전 문헌의 다층적 의미망을 포착하는 수렴 타당성(Convergent Validity)을 갖추었음을 입증한다.

향후 연구로는 대상을 넓혀 고전 한시의 시간적 흐름에 따른 사상사적 변화를 추적하기 위해 시간 변수(Timestamp)를 통합한 '동적 고밀도 압축 텍스트 토픽 모델링(Dynamic HDCTM)'으로의 발전을 도모하고자 한다.

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. DOI: 10.1162/jmlr.2003.3.4-5.993

[2] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022. DOI: 10.48550/arXiv.2203.05794

[3] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," The Journal of Open Source Software, Vol. 2, No. 11, pp. 205, 2017. DOI: 10.21105/joss.00205

[4] C. Galli, C. Cusano, M. Meleti, N. Donos, and E. Calciolari, "Topic Modeling for Faster Literature Screening Using Transformer-Based Embeddings," Metrics, Vol. 1, No. 1, Article 2, 2024. DOI: 10.3390/metrics1010002

[5] S. K. Hsieh, "Hanzi, concept and computation: A preliminary survey of Chinese characters as a knowledge resource in NLP," Ph.D. dissertation, University of Tübingen, Tübingen, Germany, 2006.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, pp. 4171-4186, 2019. DOI: 10.18653/v1/N19-1423

[7] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence," Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (ACL-IJCNLP), pp. 759-766, August 2021. DOI: 10.18653/v1/2021.acl-short.96

[8] C. M. Pham, A. Hoyle, S. Sun, 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) (NAACL), pp. 2956-2984,

June 2024. DOI: 10.18653/v1/2024.naacl-long.164.

[9] Y.Mu, C. Dong, K. Bontcheva, and X. Song, "Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling," Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), pp. 10160-10171, May 2024. Available at: https://aclanthology.org/2024.lrec-main.887/

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al., "Language models are few-shot learners," Advances in Neural Information Processing Systems, Vol. 33, pp. 1877-1901, 2020. DOI: 10.5555/3495724.3495883

[11] E. H. C. Chow, "Retrieval and Multi-Hop Reasoning in 1M-Token Context Windows: Evaluating LLMs on Classical Chinese Text," arXiv preprint arXiv:2605.02173, 2026. DOI: 10.48550/arXiv.2605.02173

[12] W. Chen, W. Hussain, and J. Chen, "GLMTopic: A Hybrid Chinese Topic Model Leveraging Large Language Models," Computers, Materials & Continua, Vol. 85, No. 1, pp. 1559–1583, 2025. DOI: 10.32604/cmc.2025.065916

[13] Y. Yao, M. Wang, B. Chen, and X. Zhao, "WenyanGPT: A large language model for classical Chinese tasks," arXiv preprint arXiv:2504.20609, 2025. DOI: 10.48550/arXiv.2504.20609

[14] L. Peng, "A Topic Modeling Web Service for Classical Chinese Poetry With Auxiliary Semantic Information Enhancement," International Journal of Web Services Research, Vol. 22, No. 1, pp. 1–28, 2025. DOI: 10.4018/ijwsr.390498

[15] W. Li and P. C. Brom, "Paradox of poetic intent in back-translation: evaluating the quality of large language models in Chinese translation," Frontiers of Information Technology & Electronic Engineering, Vol. 26, No. 11, pp. 2176–2203, 2025. DOI: 10.1631/fitee.2500298

[16] G. H. Shin, "A Study on the Ideological Characteristics and Poetic Figuration of Kim Chang-heup's Poetry," Ph.D. dissertation, Changwon National University, Changwon, Korea, 2019.

[17] Y. J. Choi, "A Study on the Philosophical Poetic World of Samyeon Kim Chang-heup," Ph.D. dissertation, Korea University, Seoul, Korea, 2015.

[18] N. G. Kim, "The Life and Poetic World of Samyeon Kim Chang-heup," Journal of Korean Poetry in Classical Chinese, Vol. 13, pp. 27-55, December 2009. DOI: UCI:I410-ECN-0102-2012-300-000216773

Authors

Byeong-chan Lee received his B.A. in Chinese Language and Literature from Chungnam National University in 1991, followed by an M.A. in Korean Literature, specializing in Classical Chinese Literature, from the same institution in February 1994. He earned his Ph.D. in Classical Chinese Literature, with a focus on Sino-Korean poetry, from Dankook University in August 2001. He is currently a lecturer in the Department of Classical Chinese Literature at Chungnam National University. His research interests include digital humanities, with a particular focus on applying big data analysis, n-gram corpora, text mining, digital analytics, and visualization techniques to the study of classical Chinese poetry (Hansi).