# 서지 메타데이터 자동 생성 성능 비교 연구*

## - 국내외 대규모 언어모델(Large Language Model)을 중심으로 -

# A Comparative Study of Automatic Bibliographic Metadata Generation Performance: Focusing on Domestic and International Large Language Models (LLMs)

김 선 욱 (SeonWook Kim)**
이 혜 경 (Hyekyung Lee)***

## 초 록

본 연구는 국내 소버린 AI와 글로벌 LLM을 비교하여 도서관 메타데이터 자동 생성을 위한 활용 가능성 파악하는 것을 목적으로 한다. 이를 위해 GPT, Gemini, Grok, HyperCLOVA, EXAONE, A.X 총 6종의 생성형 대규모 인공지능 언어모델을 대상으로 국내외 도서 40권의 MARC 레코드를 생성하게 하고, 완전성, 정확성, 규칙성의 세 가지 척도에 따라 필드 단위 성능을 평가하였다. 분석 결과, 첫째, GPT, Gemini, Grok의 글로벌 LLM 3종은 필드 누락이 적고 지시기호, 식별기호 등 형식 요소를 비교적 안정적으로 처리하여, 전반적으로 국내 소버린 AI 모델보다 높은 성능을 보였으나, 국내 도서로 전환될 경우, 필드 구성과 작성방식 등에서 오류를 보이며 성능이 저하되는 경향이 확인되었다. 둘째, HyperCLOVA, EXAONE, A.X의 국내 소버린 AI 모델은 MARC21 및 KORMARC 기술 모두에서 전반적인 성능 수준이 낮았고, 국내 도서에 대해서도 뚜렷한 성능 향상을 보이지 못하였다. 셋째, 필드별로는 표제와 책임표시사항(245)처럼 대부분의 모델이 비교적 안정적으로 생성하는 영역이 있는 반면, 총서사항(490/830)이나 기본표목의 설정 등 규칙 의존도가 있는 필드에서 모델 간 성능 격차 및 MARC21의 총서 처리 방식을 KORMARC에 기계적으로 적용하는 등 서지 작성 규칙 구조에 대한 이해 부족을 드러냈다. 이에 따라 현시점에서 생성형 인공지능을 도서관 메타데이터 업무에 도입할 때, 전면적인 자동목록 도구로의 전환 보다, 서지 레코드 초안 생성과 오류 탐지, 보완을 지원하는 보조 도구로 활용하는 것이 타당함을 시사하며, 아울러 국내 소버린 AI의 성능 안정성을 확보하기 위해서는 KORMARC를 포함한 국내 서지 데이터를 기반으로 한 체계적인 학습이 필요할 것으로 보였다. 또한 도서관용 소버린 AI를 구축하기 위해서는 학습 데이터의 선별이 주요한 과제로 요구된다.

## ABSTRACT

This study aims to examine the feasibility of using domestic sovereign AI models and global large language models (LLMs) for automated creation of library metadata by comparing their performance in MARC record generation. To this end, six generative AI models (GPT, Gemini, Grok, HyperCLOVA, EXAONE, and A.X) were used to generate MARC records for 40 domestic and foreign monographs, and their field-level performance was evaluated using three criteria: completeness, correctness, and rule compliance. The analysis showed, first, that the three global LLMs (GPT, Gemini, Grok) generally outperformed domestic sovereign AI models, with fewer missing fields and more stable handling of formal elements such as indicators and codes. However, their performance tended to decline when the cataloguing target shifted from English-language to Korean books, as errors increased in field configuration and statement of responsibility. Second, the domestic sovereign AI models (HyperCLOVA, EXAONE, A.X) exhibited relatively low overall performance in both MARC21 and KORMARC, and did not show clear performance gains even for Korean books. Third, at the field level, most models generated relatively stable results for title and statement of responsibility (245), whereas rule-dependent fields such as series statements (490/830) and the choice of main entry showed large performance gaps between models and revealed structural misunderstandings of cataloguing rules for example, mechanically transferring MARC21 practices for series treatment to KORMARC. These findings suggest that, at present, generative AI should be introduced into library metadata workflows primarily as an assistive tool for generating draft records and supporting error detection and correction, rather than as a fully automated cataloguing system. The results also indicate that, in order to ensure stable performance of domestic sovereign AI models, systematic training on Korean bibliographic data, including KORMARC records, is required. Furthermore, the careful selection and curation of training data emerges as a key task in building sovereign AI systems for library applications.

키워드: 생성형AI, 소버린AI, 메타데이터 자동 생성, 한국문헌자동화목록형식, KORMARC, MARC21
Generative AI, Sovereign AI, Automatic Metadata Generation, Korean Machine Readable Cataloging Format, KORMARC, MARC21

---

* 이 논문은 2025학년도 한남대학교 학술연구비 지원에 의하여 연구되었음.
** 경북대학교 사회과학대학 문헌정보학과 강사(sewokim@gmail.com) (제1저자)
*** 한남대학교 문과대학 문헌정보학과 조교수(keilee@hnu.kr) (교신저자)
논문접수일자 : 2025년 11월 24일 논문심사일자 : 2025년 11월 25일 게재확정일자 : 2025년 12월 8일
한국비블리아학회지, 36(4): 303-331, 2025. http://dx.doi.org/10.14699/kbiblia.2025.36.4.303

# 1. 서 론

ChatGPT를 시작으로 한 생성형 AI 서비스가 등장한 지 만 3년이 되지 않았으나, "5년, 늦어도 10년 이내에 인간 수준의 인공 일반지능(Artificial General Intelligence)이 등장할 것"이라는 업계의 전망은 또 다른 격변을 예고하고 있다(Brown, 2025; Nellis, 2024).

생성형 AI 서비스의 출발점이 대규모 언어모델(Large Language Model)이었기에, LLM과 생성형 인공지능을 사실상 동일시하는 경향이 있는 것도 사실이므로 용어와 범주는 구분할 필요가 있다. 언어모델은 동작 방식에 따라 생성형과 판별형으로 나뉘며, 규모의 관점에서 말하는 '대규모'는 학습 데이터, 모델 파라미터, 연산 자원이 매우 큰 경우를 가리킨다. 이때 대규모라고 해서 반드시 생성형인 것은 아니다. 예컨대 BERT는 대규모 언어모델이지만 본질적으로 판별형 모델에 속한다. 따라서 '생성형/판별형'과 '대규모/소규모'는 서로 다른 축의 구분이다.

마찬가지로 AGI를 "지각·추론·계획·맥락 전이 등 광범위한 지적 능력을 수행하는 시스템"으로 정의하면, 현재 우리가 사용하는 생성형 인공지능과 미래의 AGI 사이에는 여전히 기술적 격차가 존재한다(Mumuni & Mununi, 2025).

그럼에도 불구하고 생성형 인공지능이 초래한 변화는 이미 예상치를 넘어서는 속도로 확산되고 있다. 대표적 지식 산업 직군인 소프트웨어 개발 영역만 보더라도, 코드 생성·리뷰·테스트 자동화 등 다수의 업무 단계가 재편되고 있으며, 일부 과업은 자동화로 대체되는 추세가 관찰된다. 대면 서비스나 수작업 중심 업무는 예외로 남을 것이라는 전망도 있으나, 기술의 목적이 인간의 편의를 확대하는 데 있다는 점을 고려하면 산업 전반과 조직의 다양한 수준에서 구조적 변화가 이어질 가능성이 크다. 문헌정보학 분야도 예외는 아니며, IFLA는 2023년 8월 네덜란드 로테르담에서 열린 '88th IFLA World Library and Information Congress'에서 사서의 목록작성 업무에 인공지능이 어떤 영향을 미칠 것인가에 대한 세션을 진행한 바 있다(Visser et al., 2023).

하지만 이러한 논의 과정에서 간과해서는 안 될 사실은 현재 세계 시장을 주도하는 상용 LLM들이 주로 영미권 데이터를 기반으로 학습된 상태로 비영어권 언어와 문화적 맥락에 대한 이해도가 상대적으로 낮다는 점이다(Huo et al., 2025; Yan et al., 2024). 이를테면 문헌정보학 분야에서의 한국어 서지 데이터가 가진 고유한 특성이나 KORMARC와 같은 국내 표준을 처리하는 데 있어 정합성이 떨어지거나 환각을 일으킬 위험을 내포하고 있다. 더불어 국가적 지식 자산인 도서관 데이터를 해외 기업의 서버에 의존하여 처리하는 것은 미래의 데이터 안보와 프라이버시 측면에서도 우려할 만하다. 이에 따라 단순히 성능이 뛰어난 모델을 사용하는 것을 넘어, 자국의 데이터 주권(data sovereignty)을 확보하고 문화적·언어적 맥락을 온전히 반영할 수 있는 인공지능 인프라의 필요성이 제기되고 있다(Dale, 2025). 바로 이러한 맥락에서, 글로벌 빅테크 기업에 종속되지 않고 국가 단위의 자주적 인공지능 구축 및 운용 능력을 보유하려는 움직임이 소버린 AI(Sovereign AI)로 구체화되고 있다.

소버린 AI는 이미 EU와 사우디아라비아는 인공지능의 전략적 중요성을 인식하고 소버린

AI를 통한 국가적 역량을 키우려는 노력을 기울이고 있으며, 우리나라도 최근 정권 교체 이후 소버린 AI를 국가 차원에서 지원하는 정책을 본격적으로 추진하기 시작하였다.

이와 같은 뉴노멀에 대비하기 위해서는, 인공지능 서비스를 관리할 수 있는 능력을 갖추고 이를 수행할 숙련된 전문가를 체계적으로 양성하기 위한 교육과 현업의 변화가 요구된다. 문헌정보학 분야 역시 사서들이 주제 전문가(domain specialist)로서 인공지능을 구조적으로 이해하고 국가도서관 체계 내 적재적소에 인공지능 기반 서비스를 배치 및 운영할 수 있는 역량을 기를 수 있도록 다양한 접근과 해석이 병행되어야 한다.

따라서 본 연구는 국내 소버린 AI 모델인 EXAONE, A.X, HyperCLOVA 3종과 글로벌 상용 LLM인 GPT, Gemini, Grok 3종을 동일한 제로 샷 및 공통 프롬프트 조건에서 비교 및 평가함으로써, KORMARC와 MARC21 작성 규칙에 부합하는 서지 메타데이터 생성 품질의 현황과 한계를 파악하는 것을 목적으로 한다. 나아가 이러한 연구의 결과는 궁극적으로 도서관 현장에서의 합리적인 생성형 AI 모델 선택과 업무 프로세스 개발을 위한 지침을 마련하는 데 기초자료를 활용될 수 있을 것으로 기대한다.

## 2. 선행연구

문헌정보학 분야에서 메타데이터를 자동 생성과 관련된 연구는 레코드의 자동 변환과 구분에서 시작되었다. 또한, 전통적인 메타데이터 품질 평가 방안은, 해당 분야의 전문가가 올바르게 메타데이터가 작성되었는지 직접 살펴보는 행위라고 정의할 수 있다.

이미화(2015)는 RDA(Resource Description and Access) 도입 과정에서 기존 서지 레코드에 RDA 요소를 추가하는 업무가 인간이 처리할 수 없는 대량의 메타데이터를 생성이 요구되는 상황임을 인식하고, 이를 해결하려는 방안으로 Library of Congress-Program for Cooperative Cataloging의 하이브리드화 레코드 관련 지침을 파악하였으며, 태그별 구분과 식별, 자원유형 변환을 위한 매핑표 작성등의 요구사항이 필요함을 피력하였다.

도서관이 취급해야 하는 정보의 규모가 확장됨에 따라, 연구자들은 실제 도서관에 적용 가능한 사례로 자동 주제색인에 깊은 관심을 가져왔으며, 통계기법과 머신러닝을 활용한 자동 주제색인 생성에 관한 연구를 진행하였다(Golub, 2019; Suominen, 2019). 특히 독일 국립도서관(DNB)은 2019년부터 2022년까지 자동 주제 목록화 시스템(EMa)을 실제로 개발하여 도입하였고, 2022년에는 자연어처리를 활용한 자동 목록화 시스템 도입에 관한 연구를 수행함으로써 국가도서관 수준에서 메타데이터 자동 생성의 대표적인 모범 사례로 간주되고 있다(Poley et al., 2025).

미국 의회도서관은 2022년부터 'Exploring Computational Description'이라는 일련의 실험 연구를 통해 자동화된 메타데이터 생성의 실현 가능성을 탐색하였다. 특히 2023년부터 2024년까지 진행된 2단계 연구에서는 프로토타입을 구축하기 위한 실험을 수행하면서 약 23,000건의 전자책과 기존 MARC 레코드를 기

반으로 다양한 머신러닝 모델을 테스트하였다. 그 결과 제목・저자・식별자 등 핵심 서지 요소의 자동 추출에서 높은 성능을 확인했지만, 데이터의 편중, 다중 주제 분류 처리의 어려움, 훈련 데이터 부족, 저작권 및 개인정보 보호와 같은 규제 준수 문제 등 실무적 난점도 드러났다(Brador, 2024).

생성형 AI의 등장 이후부터는 이를 적극적으로 도서관에 도입하려는 연구가 이어졌다. 이용구(2023)는 대규모 언어모델인 BERT를 이용한 주제명을 자동 분류 가능성을 타진하고 실제 주제명이 부여된 KDC 분류체계와 주제명 범주에 대한 성능을 평가하였다. 김선욱 외(2023)는 국내 도서를 대상으로 ChatGPT가 자동 생성한 더블린 코어 메타데이터의 품질을 평가하여, 생성형 AI가 실제 도서관 메타데이터 작성에 적용될 가능성을 검토하였다. Taniguchi(2024)는 ChatGPT(GPT-4)를 활용하여 RDA 기준에 따른 MARC 21 서지 레코드 생성 가능성을 검증하기 위해, Maxwell's Handbook for RDA의 105건의 서지 레코드와 정보원 데이터를 기반으로 제로-샷 방식을 적용, MARC21 레코드를 생성한 후 ChatGPT 자체 평가 결과와 전문가 평가 결과를 비교하였다.

D'Souza et al.(2025)는 독일 국립과학기술도서관 (TIB)의 공개 카탈로그(TIBKAT)에 수록된 약 12만 건의 제목과 초록을 입력 데이터로 활용하여, 독일 통합 전거파일(GND) 주제어를 top-k로 추천하는 대규모 언어모델(LLM) 기반의 자동 주제색인 과제를 설계하고, 이를 정성적인 방법과 정량적인 방법으로 평가하는 방안을 제안하였다.

한편, 앞서 살펴본 생성형 AI 기반 선행연구들은 주로 대중적으로 널리 활용되고 있는 GPT 계열 모델의 성능을 검증하고 도입 가능성을 탐색하는 데 초점을 두고 있다. 그러나 최근에는 GPT 이외에도 Gemini, Grok 등 다양한 대규모 언어모델 기반 생성형 AI가 등장하고 있음에도 불구하고, 이들 후발 모델에 대한 체계적인 비교 연구는 여전히 부족한 실정이다. 특히 국내에서 개발된 생성형 AI 모델을 대상으로 도서관 메타데이터 업무에의 적용 가능성을 검토한 연구는 거의 전무하다. 이에 본 연구는 GPT를 포함한 총 6종의 국내외 생성형 AI 모델을 대상으로 서지 데이터생성 성능을 비교・평가함으로써, 다양한 생성형 AI의 현 수준을 파악하여, 향후 도서관 업무에서의 합리적인 모델 선택과 활용 방향을 논의하는 데 기초자료를 제공한다는 점에서 의의가 있다.

## 3. 이론적 배경

### 3.1 대규모 언어모델의 발전

2022년 11월, Open AI 社가 ChatGPT를 조용히 공개(open)했을 때, 기존에 클로즈(close) 베타 서비스 접근 승인을 기다리던 일부 연구자들은 완전 개방 정책을 크게 환영했으나, 여전히 일반 대중에게 ChatGPT는 인식 지평 밖의 존재였다. 이듬해인 2023년에 들어서야 언론의 집중적 보도와 다양한 실용적 사례의 등장으로 ChatGPT가 혁신의 도구로 소개되었고, 이를 통해 일반 대중도 비로소 인공지능의 시대에 접어들었음을 체감하게 되었다.

그러나 불과 수년 동안, 생성형 AI는 전례

없는 확산 속도로 우리 삶을 바꾸고 있다(한국은행, 2025). 이 과정에서 다양한 산업 분야와 학계에서의 실제 적용 사례가 공개되면서, 기존의 AI 리터러시에 대한 개념적 접근은 실제적 도구로서 연구되기 시작했다(Laupichler et al., 2022; Southworth et al., 2023).

동시기에 인공지능 분야를 연구 중이던 학계와 업계는 ChatGPT의 등장을 계기로 급격한 변화를 보이기 시작하였다. 2023년 2월에 최초 공개된 Meta의 LLaMa는 공개 가중치(Open Weight) 및 광범위 허용 정책과 함께 배포되었으며(Touvron et al., 2023), 이는 더 많은 인공지능 연구자와 기업을 해당 분야로 이끄는 계기가 되었다. 이로 인해 '라마(llama)'를 기원으로 하는 알파카(Alpaca), 비큐냐(Vicuna)와 같은 파생 모델이 등장하였고, 기존의 대규모 언어모델을 '저비용 고효율 미세조정'으로 더 강력하게 만들 수 있음을 증명함으로써 모델 생태계가 구축될 수 있는 시발점이 되었다(Chiang et al., 2023; Dubois et al., 2023). 그 결과 공개 모델의 등장과 파생 모델의 고도화 연구가 급성장하는데 크게 이바지하였다(Chen at al., 2023; Minaee at al., 2024).

한편 대규모 언어모델이 기업에서 통합 운용 중인 대표적인 사례로는 2023년 5월에 공개된 Google의 Gemini를 들 수 있다. Gemini는 ChatGPT처럼 웹을 통해 접근할 수 있을 뿐만 아니라, 기업형 서비스로서 안드로이드 기반 스마트폰에 기본 제공되며, 전 세계적으로 30억 명의 이용자를 보유한 생산성 향상 플랫폼인 Google Workspace와도 통합되어 기업 생산형 향상을 위한 도구의 역할을 강조하고 있다(Pappu, 2024).

### 3.2 대규모 언어모델의 방향

최근까지 이용자는 LLM 모델과 그 버전을 선택하는 단일 모델 선정 방식을 주로 사용해 왔다. 이로 인해 사용 중인 모델에 문제가 발생하면 다른 모델로 교체해야 했으며, 한 도메인에서 우수한 성능을 보인 모델도 다른 도메인에서는 성능이 저하될 수 있으므로 도메인 혹은 업무별로 별도 모델 평가가 필수적이었다. 그 결과 기업은 항상 모델 변경에 따라 발생하는 비용의 증가와 운영 위험을 상식적으로 감수해야 했다. 생성형 AI의 동작을 명확히 예측하기 어려운 상황에서 이와 같은 문제를 해결하려면 더 복잡한 모델이 필요하다는 규모의 법칙이 그동안의 연구 기조였다(Kaplan et al., 2020). 이에 따라 기업과 연구자는 더 많은 파라미터와 더 많은 GPU 연산 자원을 전제로 하는 더 복잡한 모델을 지향하는 연구 및 산업계의 경향으로 나타났다.

그러나 중국의 딥시크(Deepseek) 사례를 통해 더 많은 파라미터와 더 큰 모델이 항상 더 나은 성능을 보장하지 않으며, 오히려 이용자의 도메인에 따라 더 적은 데이터와 낮은 비용으로 구축한 모델이 더 우수한 성능을 보일 수 있음을 보여주었다(Luo et al., 2022; Guo et al., 2024; Wu et al., 2023). 이와 더불어 에너지 소비와 탄소 비용을 고려할 때, 대규모 언어모델을 기반으로 하는 생성형 AI가 과연 지속 가능한 비즈니스인가에 대해 회의적인 시각을 제기하는 견해도 등장하였다(The Economist, 2023).

이러한 논의에 따라 최근의 연구 동향은 대규모 언어모델의 규모를 방대히 하는 것 보다, 효율적인 규모 확장과 도메인 특화를 지향하는 방

향으로 전환되고 있으며(Dettmers et al., 2023; Li et al., 2023; Shao et al., 2024), 기존의 대규모 언어모델과 같은 거대 단일 모델은 파운데이션 모델 (Foundation Model)이라는 용어로 재정의되고 있다(Bommasani et al., 2021).

특히 최근 생성형 AI의 연구에서 주목받는 주제는 라우터(router)의 적용이다. MoE(Mixture of Experts) 구조라고도 불리는 이 아키텍처는 하나의 거대한 단일 모델을 사용하는 대신, 여러 개의 분야별 전문가 하위 모델을 구성하고, 라우터가 입력 토큰을 분석하여 이용자의 의도와 요구 수준을 파악한 뒤 해당 작업을 가장 잘 처리할 수 있는 특정 전문가에게만 연산을 할당하거나 여러 전문가가 상호작용하도록 함으로써 보다 복잡한 추론을 가능하게 한다(Fedus et al., 2021; Jiang et al., 2024).

### 3.3 소버린 AI의 등장

대형 모델에 기반한 생성형 AI 서비스는 분산 처리와 안정적인 운용을 위해 국경을 초월한 클라우드와 인프라와 네트워크를 통해 구현된다. 이러한 구조를 통해 전 세계적으로 자원을 분산함으로써 국소적인 부하를 완화하고, 자원을 보다 효율적으로 관리할 수 있다는 장점이 있다.

그러나 해외에 위치한 장비에 의존할 경우, 해저 케이블 손상이나 주요 데이터센터 자체에 재난이 발생 시, 국내 이용자는 해외의 제공기관이 문제를 해결할 때까지 서비스 중단을 감수할 수밖에 없다. 재난 상황이 아니더라도, 현재 전 세계 인공지능 관련 핵심기술을 보유하거나 이를 제공하는 일부 빅테크 기업은 막강한 통제 권한을 행사하고 있으면서, 자국의 규제와 수출 통제 등 각종 정책의 영향을 받는다. 이는 곧 해당 인공지능 서비스가 국내법과 국제법의 규제의 이중적 제약을 동시에 받는다는 한계를 의미한다. 이로 인해 국제 정세의 변화에 따라, 해당 기업이 제공하는 인공지능 서비스의 기능이 제한되거나 이용이 차단되는 등의 상황이 발생할 가능성도 존재한다.

이 외에도 대형 모델을 학습하는 데 필요한 대규모 컴퓨팅 장비의 수출을 통제하려는 움직임이 나타나고, 전문 인력의 해외 유출에 대한 우려가 커지는 등 다양한 관점에서 AI 주권의 필요성이 제시되고 있다(Reuters, 2025; The Straits Times, 2025).

따라서 소버린 AI는 단순한 기술 소유가 아니라 국내법 및 국제법의 정합성을 함께 고려하여 데이터, 장비, 모델을 포괄하는 전체 생태계 관점에서 접근해야 한다(Dale, 2025).

2024년 과학기술정보통신부가 '대한민국 인공지능 3대 강국 도약' 계획을 발표한 이후(과학기술정보통신부, 2024), 2025년에는 대통령실 산하에 'AI 미래기획 수석비서관' 직위를 신설하고, 대통령 직속으로 '민·관 합동 국가인공지능전략위원회'를 강화하는 정부조직 개편을 단행함으로써, 국가적 컨트롤타워를 통한 포괄적 선도를 도모하고 있다(정준화, 2025). 그리고 NVIDIA社는 우리나라 소버린 AI 정책을 지원하기 위한 GPU 지원을 약속했다(NVIDIA, 2025).

이러한 추세 속에서 2025년 8월 8일 과학기술정보통신부는 「독자 인공지능 기초 모형(AI 파운데이션 모델)」 사업 대상자로 네이버 클라우드, 업스테이지, SK텔레콤, NC AI, LG AI 연구원 등 5곳을 선정하고, 한국형 인공지능

(K-AI)을 위한 모델 구축 사업에 착수하였다(과학기술정보통신부, 2025). 이 가운데 LG AI 연구원과 네이버 클라우드는 모델 성능에 관하여 대중에게 공개한 바 있다(LG AI Research, 2024; LG AI Research, 2025; NAVER Cloud HyperCLOVA X Team, 2025).

## 4. 연구방법

### 4.1 연구 설계

본 연구는 국내 소버린 AI 3종과 상용 LLM 3종을 대상으로 서지 메타데이터를 생성하게 하여 그 성능과 한계를 파악하고자 하였다. 이를 위하여 본 연구는 〈그림 1〉과 같이 데이터 수집 및 처리, 메타데이터 생성, 평가의 3단계로 진행하였다.

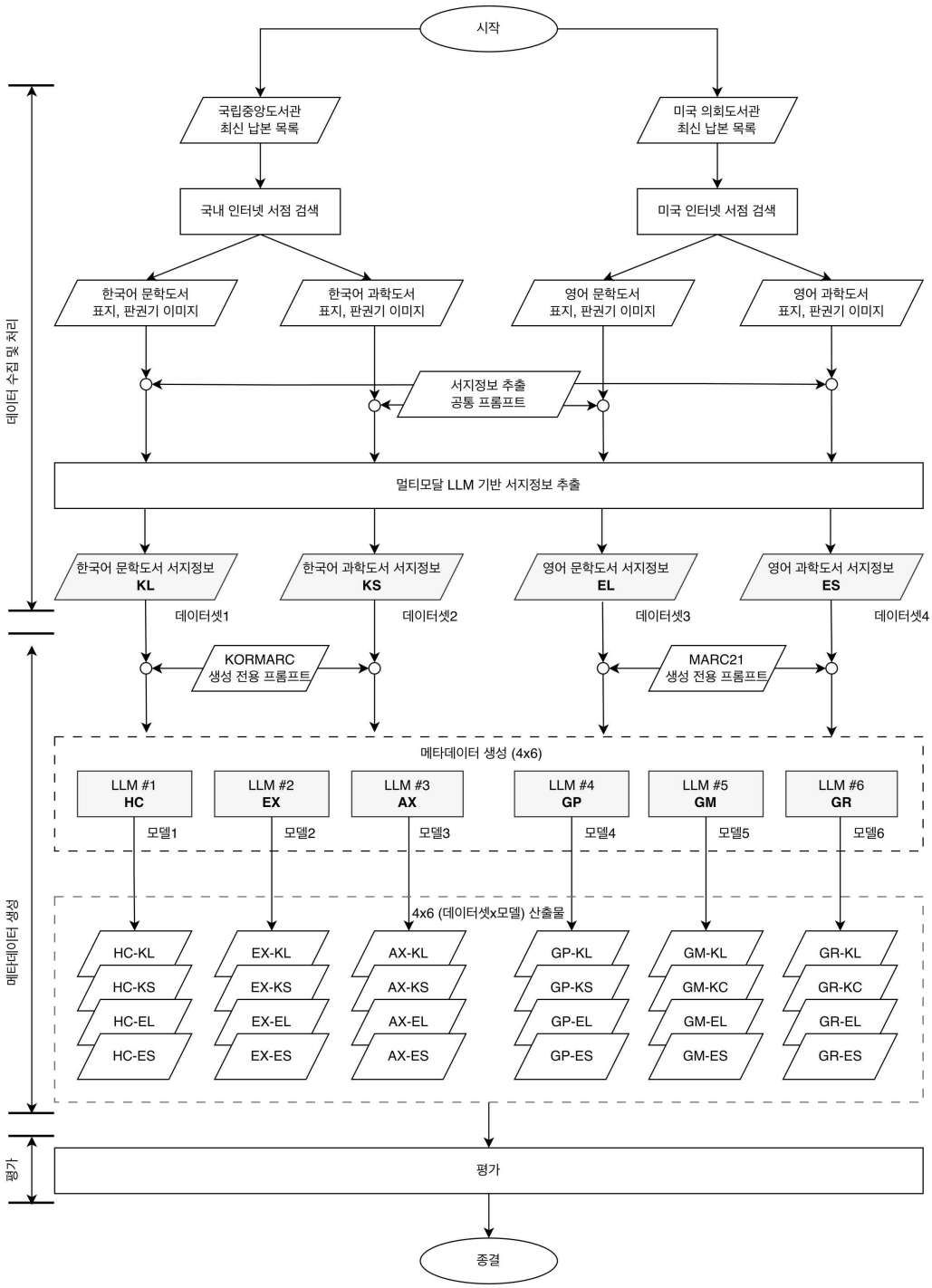
### 4.2 데이터 수집 및 처리

데이터 수집 및 처리의 개괄적인 흐름은 〈그림 1〉의 좌측에 표시된 '데이터 수집 및 처리' 부분에 해당한다. 데이터 수집은 2025년 7월 4주부터 8월 1주 차까지 최신 도서를 대상으로 진행하였다. 또한 Artificial Analysis 社가 제공하는 Artificial Analysis LLM Performance Leaderboard(Artificial Analysis, 2025)에 따르면, 본 연구에서 사용한 모델들은 모두 2025년 상반기에 출시되었으며, 이에 따라 이들 모델이 최신 도서의 서지 정보를 충분히 학습하지는 않았을 것으로 가정하였다. 그리고 최소 수집 단위인 도서 이미지는 선행연구에 따라 표지와 판권기로 한정하였으나(김선욱 외, 2023), 국립중앙도서관과 미국 의회도서관에서는 판권기 이미지를 제공하지 않으므로, 이는 별도의 인터넷 서점에서 추가 수집하였다.

따라서 본 연구에서는 실험 대상 도서의 표지와 판권기 이미지, 그리고 각 국가대표도서관에서 제공하는 MARC 레코드를 수집하였는데, 이 중 MARC 레코드는 생성형 AI 모델의 성능을 검증하기 위한 기준 자료로 활용하기 위함으로, 서지 데이터는 작성자에 따라 내용에 차이가 발생할 수 있으나, 본 연구에서는 성능을 정량적으로 도출하고 모델 간 차이를 비교할 수 있는 공통 기준이 필요하므로, 국가대표도서관인 국립중앙도서관과 미국 의회도서관에서 제공하는 MARC 데이터를 '정답 데이터'로 간주하여 성능 평가를 수행하였다. 이에 따른 데이터 수집 및 처리의 세부 절차는 다음과 같다.

우선 국내 도서데이터는 국립중앙도서관의 최신 납본 목록을 기준으로, 국내 인터넷 서점(교보문고, 알라딘 등)에 표지 및 판권기 이미지가 제공되고 있는 도서만을 대상으로 수집하였다. 국립중앙도서관은 웹사이트 내 신착도서 메뉴를 통해 KDC 10 주류와 '전체'까지 총 11개 분류를 통해 신착자료를 안내하고 있다. 이 가운데 본 연구는 이중 '문학', '자연과학', '기술과학' 신착 자료 중 가장 최신 목록에 해당하면서 도서의 표지와 판권기의 이미지를 모두 확보할 수 있는 도서만을 선정하여 문학 분야 10권, 과학 분야 10권을 선정하고 이들의 이미지와 MARC 레코드를 수집하였다.

한편, 국외 도서데이터는 미국 의회도서관의 최신 납본 목록과 미국 인터넷 서점(아마존, 반


시작
국립중앙도서관
최신 납본 목록
미국 의회도서관
최신 납본 목록
국내 인터넷 서점 검색
미국 인터넷 서점 검색
한국어 문학도서
표지, 판권기 이미지
한국어 과학도서
표지, 판권기 이미지
영어 문학도서
표지, 판권기 이미지
영어 과학도서
표지, 판권기 이미지
서지정보 추출
공통 프롬프트
멀티모달 LLM 기반 서지정보 추출
한국어 문학도서 서지정보
KL
한국어 과학도서 서지정보
KS
영어 문학도서 서지정보
EL
영어 과학도서 서지정보
ES
데이터셋1
데이터셋2
데이터셋3
데이터셋4
KORMARC
생성 전용 프롬프트
MARC21
생성 전용 프롬프트
메타데이터 생성 (4x6)
LLM #1
HC
LLM #2
EX
LLM #3
AX
LLM #4
GP
LLM #5
GM
LLM #6
GR
모델1
모델2
모델3
모델4
모델5
모델6
4x6 (데이터셋x모델) 산출물
HC-KL
HC-KS
HC-EL
HC-ES
EX-KL
EX-KS
EX-EL
EX-ES
AX-KL
AX-KS
AX-EL
AX-ES
GP-KL
GP-KS
GP-EL
GP-ES
GM-KL
GM-KC
GM-EL
GM-ES
GR-KL
GR-KC
GR-EL
GR-ES
평가
종결
데이터 수집 및 처리
메타데이터 생성
평가

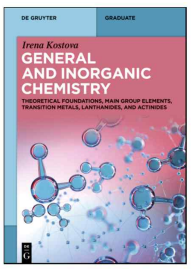
〈그림 1〉 연구 설계

즈 앤드 노블 등)에서 표지 및 판권기 이미지가 제공되는 자료만을 선정하여 수집하였다. 미국 의회도서관은 십진 분류를 사용하지 않고 LC 분류법을 사용하므로, 본 연구에서는 국내 도서 데이터에 상응하는 자료를 수집하기 위하여 의회도서관의 Search Online Catalog 메뉴에서 패싯을 지정하여 자료를 선정하였으며, 그 과정은 다음과 같다.

1. '형식'에서 Book으로 한정함.
2. '언어'를 영어로 한정함.
3. '출판연도'를 2025년으로 한정함.
4. '주제'에서 주제어를 선택함.
   가. 영어 문학 도서: Literature 주제어를 선택함.
   나. 영어 과학 도서: DDC Science(500)의 강목에 해당하는 Science, Mathematics, Astronomy, Physics, Chemistry, Earth science and geology, Biology, Zoology (단, Fossils와 Plants는 검색 결과 등장하지 않아서 제외) 주제어를 선택함.

이 검색 기준에 따라 신착 자료 중 가장 최신의 목록에 해당하면서 도서의 표지와 판권기의 이미지를 획득할 수 있는 문학도서 10권과 과학 도서 10권을 선정하고 그에 해당하는 표지와 판권기 이미지 MARC21 레코드를 수집하였다.

이처럼 수집된 데이터로부터 서지정보를 추출하기 위해, 이미지 멀티모달 처리가 가능한 언어모델이자 이미지 인식 성능이 높은 Gemini 2.5 Pro를 사용해 〈그림 2〉의 (가)와 같이 수집된 도서 이미지를 처리하였다(Comanici et al., 2025; Gemini Team & Google, 2024; 2025).




표지 이미지

판권기 이미지

(가) 도서 이미지

You are an expert librarian specializing in cataloging and creating MARC records. The attached images are key pages from a book, including the front cover and the colophon.
Please Analyze the images and extract all relevant text using OCR. Due to data processing, do not add any comment irrelevant comments in the response.

(나) 서지정보 추출 프롬프트

"### Image 1: Front Cover

**Publisher:** DE GRUYTER
**Series/Audience:** GRADUATE
**Author:** Irena Kostova
**Title:** GENERAL AND INORGANIC CHEMISTRY
**Subtitle:** THEORETICAL FOUNDATIONS, MAIN GROUP ELEMENTS, TRANSITION METALS, LANTHANIDES, AND ACTINIDES
**Publisher Logo:** DE G

***

### Image 2: Colophon

**Author**
Prof. Irena Kostova
Department of Chemistry
Faculty of Pharmacy
Medical University
1000 Sofia
Bulgaria
irenakostova@yahoo.com

**ISBN** 978-3-11-171222-2
**e-ISBN (PDF)** 978-3-11-171224-6
**e-ISBN (EPUB)** 978-3-11-171254-3

**Library of Congress Control Number:** 2025931334

**Bibliographic information published by the Deutsche Nationalbibliothek**
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at [http://dnb.dnb.de](http://dnb.dnb.de).

© 2025 Walter de Gruyter GmbH, Berlin/Boston, Genthiner Straße 13, 10785 Berlin
**Cover image:** MF3d/E+/Getty Images
**Typesetting:** Integra Software Services Pvt. Ltd.
**Printing and binding:** CPI books GmbH, Leck

[www.degruyter.com](https://www.degruyter.com)
Questions about General Product Safety Regulation:
productsafety@degruyterbrill.com"

(다) 추출 결과

〈그림 2〉 서지정보 추출 예시

이때 사용된 프롬프트는 언어 구분과 상관없이 〈그림 2〉의 (나)와 같이 공통된 내용을 입력하였고, 이렇게 데이터와 프롬프트를 동시에 입력하여 추출한 서지정보의 예시는 〈그림 2〉(다)와 같다.

이렇게 추출한 결과는 이후 메타데이터 생성 및 평가 과정에서 구분하기 쉽도록 각각 4개의 데이터 세트인 KL(Korean Literature, 국내 문학도서 서지정보), KS(Korean Science, 국내 과학 도서 서지정보), EL(English Literature, 국외 문학도서 서지정보), ES(English Science, 국외 과학 도서 서지정보)으로 구분하였다.

### 4.3 메타데이터 생성

메타데이터 절차는 〈그림 1〉의 좌측에 제시된 '메타데이터 생성' 부분에 해당한다.

앞서 구축한 4개의 데이터 세트 중, 국내 도서 서지정보(KL, KS)에 대해서는 KORMARC 레코드를, 국외 도서 서지정보(EL, ES)에 대해서는 MARC21 레코드를 생성하도록 설계하였다. 이를 위하여 KORMARC 및 MARC21 생성 전용 프롬프트 각각 작성한 후, 해당 데이터 세트와 함께 대규모 언어모델에 입력하였다.

메타데이터 생성에는 해외 모델 3종과 국내 모델 3종으로, 총 6종의 대규모 언어모델이 활용되었다. 모델 선정에 있어서는 연구 수행 시점인 2025년 8월 기준으로 huggingface에서 Artificial Analysis 社가 제공하는 Artificial Analysis LLM Performance Leaderboard (Artificial Analysis, 2025)에 등재된 상위권 모델 중 연구자가 접근이 가능하거나 완전 공개된 모델을 중심으로 검토하였다. 그 결과, 해외 모델은 OpenAI 社의 GPT-5 Thinking, Google DeepMind 社의 Gemini 2.5 Pro, xAI Holdings 社의 Grok 4를 선정하였다. 국내 모델의 경우, 해당 리더보드에는 LG AI 연구원의 EXAONE (엑사원) 시리즈만이 등재되어 있었으나, 최근 국내 학계 및 산업계에서의 활용도와 연구 동향을 고려하여 Naver의 HyperCLOVAX-SEED-Think-14B, SK텔레콤의 A.X-4.0(에이닷엑스)를 추가로 선정하였다.

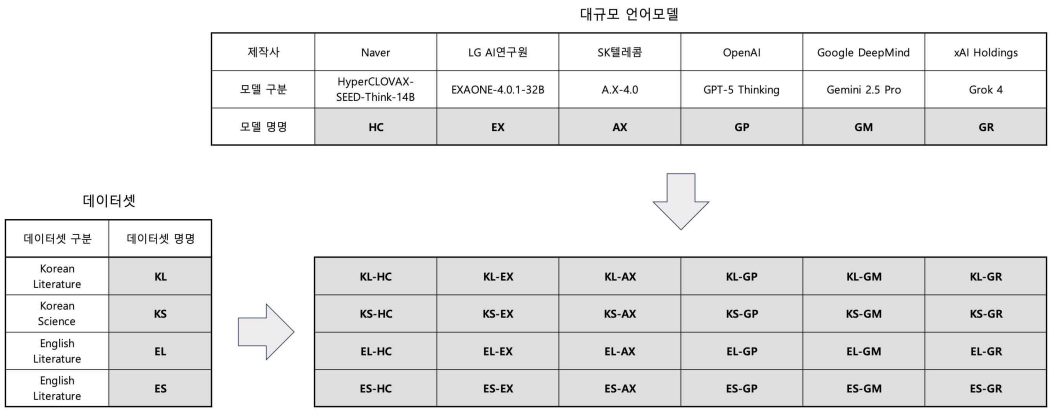
이처럼 4개의 데이터 세트에 대하여 6종의 언어모델을 적용함으로써 총 24개의 메타데이터 생성 결과 세트를 도출하였다. 이 과정에서 모델 설정 파라미터는 별도 조정하지 않았으며, 언어모델이 제공하는 기본값(default)을 그대로 사용하였다. 이렇게 생성된 24개의 결과 세트와 데이터 세트 및 언어모델 간의 대응 관계는 〈그림 3〉과 같이 정리할 수 있다.

### 4.4 평가

본 연구는 이렇게 생성한 결과에 대하여 모델별 성능을 평가하기 위하여 일련의 평가 기준 및 척도를 설정하였다.

한국목록규칙 제4판에 따르면 기술 대상 자료가 단행본인 경우, 그 정보원의 우선순위는 첫째, 표제면, 판권기, 이표제면, 표지, 둘째, 약표제면, 권두, 셋째, 책등, 넷째, 서문, 후기, 본문, 부록 등 그 도서의 나머지 부분, 다섯째, 그 도서 이외의 정보원으로 두고 있다.

한편, 본 연구에서의 수행한 실험은 언어모델에게 표지와 판권기 정보를 제공하고 MARC 레코드를 생성하도록 지시하였다. 이에 따라 표제와 책임표시사항, 판사항, 발행사항, 주기사항,




〈그림 3〉 메타데이터 생성 조합 및 결과

〈표 1〉 제공한 정보원에서 기술 가능한 필드

| 구분 | 필드 |
|---|---|
| 표제와 책임표시사항 | 1XX -, 245, 246 등 |
| 판사항 | 250 |
| 발행, 배포, 간사 사항 | 260 |
| 생산, 발행, 배포, 제작, 저작권 표시 | 264 |
| 주기사항 | 5XX- |
| 표준번호 및 입수조건사항 | 020 |
| 주제명 | 6XX- |
| 부출 | 7XX- |
| 총서사항 | 490, 830 |
| RDA 관련 | 336, 337, 338 |

표준번호 및 입수조건사항에 대한 기술이 가능할 것으로 판단하였다. 이러한 점을 고려하여 위와 같은 사항에서 생성할 수 있는 MARC 주요 필드는 〈표 1〉과 같다.

구체적으로 살펴보면 표제와 책임표시사항, 판사항, 발행사항은 표지와 판권기를 통해 기술이 가능하다. 한편 표준번호 및 입수조건사항과 주기사항은 정보원 전반에서 확인할 수 있으며, 실험을 위해 제공한 정보원 범위에서도 각 정보원에 구성에 따라 기술 여부가 결정될 수 있다. 총서사항 역시 정보원 자체 어디에서나 정보 취득이 가능하므로 그 기술 여부는 선택적인 것으로 볼 수 있다.

추가로 주제명 관련 필드인 6XX, 부출 관련한 7XX 필드를 생성할 수 있으며, 기술대상자료의 특성에 따라 총서사항 필드(490/830) 생성도 가능하다. 그리고 RDA 규정을 적용하는 경우, 336, 337, 338 필드의 생성을 추가로 고려할 수 있다.

한편 형태사항은 총서사항과 같이 도서 자체로부터 정보를 획득할 수 있는 항목이지만, 물리적 형태에 대한 직접적 표기가 판권기에 존

재하지 않는 한 형태사항 관련 필드는 작성할 수 없다고 판단된다.

김주용과 신판섭(2021)은 KORMARC 레코드에 기반하여 BIBFRAME으로의 변환 매핑을 위해 노원구립도서관의 서지 데이터를 분석하였다. 그 결과, 단행본에 해당하는 주요 필드는 25개였으며, 이 가운데 모든 KORMARC 레코드에 존재하는 필드는 국제표준도서번호(020), 소장사항(049), 한국십진분류기호(056), 자관 청구기호(090), 표제와 책임표시사항(245), 발행, 배포, 간사 사항(260), 형태사항(300), 로컬 정보-가격(950)인 것으로 확인하였다.

이에 따라 본 연구는 실험을 위해 제공되는 정보원에서 언어모델이 기술 가능한 필드를 〈표 2〉와 같이 나타내었다. 표제 및 책임표시사항과 생산・발행 및 간사 사항에 대한 필드는 제공된 정보원만으로도 필수 생성이 가능할 것으로 판단하였다. 한편 기본표목필드는 MARC21과 KORMARC에 적용되는 기술규칙이 상이하므로, 필수 여부 또한 서로 다르게 나타나며, 정보원에 따라 선택 적용할 수 있다. 〈표 2〉는 이러한 차이를 반영하여 각 필드를 대상으로 기술 여부를 필수(●)와 선택(○)으로 구분하였다. 그리고 본 연구는 언어모델에게 표지, 판권기 정보를 제공하였으며, 이와 같은 한정적인 정보만으로는 물리적 형태를 정확하게 파악하기 어렵다고 판단하였다. 따라서 물리적 형태사항에 관련한 필드생성 및 평가는 분석 대상에서 제외하였다.

김선욱 외(2023)는 ChatGPT가 생성한 더블린코어 메타데이터를 평가하기 위하여, 각 요소의 등장 여부에 따른 완전성과 정보원에 제시된 내용을 해당 요소가 정확하게 기술하고 있는지에 대한 정확성을 확인함으로써 성능을 분석하였으며, 이를 위한 산출식은 다음과 같다.

$$Completeness = \frac{\sum_{i=1}^{N} P(i)}{N} \quad \text{수식 (1)}$$

$$Accuracy = \frac{\sum_{i=1}^{N} H(i)}{N} \quad \text{수식 (2)}$$

〈표 2〉 생성형 AI에서 확인 할 수 있는 기술 필드

| 필드 | 구분 | MARC21(RDA적용) | KORMARC(KCR4) |
|---|---|---|---|
| 1XX | 기본표목 | ○ | - |
| 245 | 표제 및 책임표시사항 | ● | ● |
| 246 | 여러 형태의 표제 | ○ | ○ |
| 250 | 판사항 | ○ | ○ |
| 260 | 발행, 배포, 간사사항 | ● | ● |
| 264 | 생산, 발행, 배포, 제작, 저작권 표시 | ● | ● |
| 490 | 총서사항 | ○ | ○ |
| 8XX | 총서부출 | ○ | ○ |
| 5XX | 주기사항 | ○ | ○ |
| 6XX | 주제명 표목 | ○ | ○ |
| 7XX | 부출표목 | ○ | ○ |
| 020 | 입수사항 | ○ | ○ |

Taniguchi(2024)는 ChatGPT가 생성한 MARC 데이터를 평가하기 위하여, 1차적으로 ChatGPT에게 정답과 비교 평가를 수행하도록 지시하였고, 2차적으로 인적 평가를 진행하였다. 또한, 인적 평가에서는 각 데이터 필드의 적합성과 지시기호의 사용 여부, 의미상으로 적절한 정보의 기입 여부 등을 검토하였으며, 띄어쓰기나 구두점 등의 표기와 구분기호, 식별기호 등은 평가 범위에서 제외하였다. 이를 위한 평가 점수는 〈표 3〉과 같다.

이에 본 연구는 선행연구에서 제시된 검증 방법과 평가 기준 및 척도를 수용하여, 완전성과 정확성 그리고 지시기호 사용에 대한 규칙성을 기반으로 MARC 레코드 생성 결과를 평가하였다.

다만 본 연구에서는 각 정보원으로부터 기술할 수 있는 필드 수를 수식(1)에서의 N으로 정의하였으며, ∑P(i)는 각 언어모델이 정답과 동일하게 생성한 필드의 수의 합을 의미한다. 정확성의 경우, 언어모델이 생성한 필드를 대상으로 정보원에서 실제로 적합한 내용을 기술하였는지를 확인하였고, 이에 따라 수식(2)에서의 N은 언어모델이 정답과 일치하게 생성한 필드의 수로 수식(1)의 ∑P(i)와 동일하다. 이때 ∑H(i)는 각 필드가 정보원에 제시된 내용을 정확하게 기술하였는지 여부에 따라 산출된, '정확한 내용이 입력된 필드 수'의 합을 의미한다.

규칙성 평가와 관련하여, 본 연구는 Taniguchi(2024)의 평가 점수를 역순으로 재구성하여 적용하였다. 또한, 지시기호뿐만 아니라 식별기호 역시 평가 대상에 포함하였으며, 지시기호 및 식별기호 수준에서의 수정 필요 여부에 따라 점수를 부여하였다. 구체적으로, 수정이 불필요한 경우는 1점, 약간 수정이 필요한 경우는 0.5점을 부여하였는데 이때의 범위는 지시기호 단위에서의 오류에 해당한다. 전반적 수정이 필요한 경우는 0점으로 지시기호와 식별기호 모두 오류인 경우이다. 이렇게 설정한 척도를 규칙성으로 정의하였으며, 마찬가지로 정확성이 확보된 필드에 한해서만 규칙성을 산출하였다. 이에 따른 규칙성 산출식은 수식(3)과 같으며, 이때의 N은 정확성 평가에서 확보된 생성 필드의 개수로써 수식(2)의 ∑H(i)와 동일하며, ∑R(i)는 정확한 내용을 기입한 필드가 지닌 규칙성 점수의 합을 의미한다.

$$Rule\ compliance = \frac{\sum_{i=1}^{N} R(i)}{N} \qquad \text{수식 (3)}$$

기술(description)은 원칙적으로 자료 자체에 나타난 정보 그대로를 기입하도록 하며, 기본표목의 경우, 목록규칙에 따라 전거형식으로 작성하게 되어있다. 따라서 제공된 정보에서 작성할 수 없는 내용을 생성한 경우에는 생성형 AI의 특징인 환각(hallucination)으로 판정하고, 평가의 범위에서 제외하였다.

〈표 3〉 Taniguchi(2024)가 제시한 데이터 필드 평가 점수

| | |
|---|---|
| 수정이 불필요한 경우 | - |
| 약간의 수정이 필요한 경우 | 1 |
| 전반적 수정이 필요한 경우 | 2 |

성능판정을 위한 정답 MARC 레코드의 구성은 데이터 수집 때 각 국립중앙도서관과 미국 의회도서관에서 수집한 MARC 레코드를 활용하여 이를 정답 레코드로 간주하였다. 정답 필드의 결정은 〈표 2〉에서 제시한 필드를 참고함과 동시에, 현장의 목록 전문가가 실제 정보원을 기준으로 생성 가능한 필드를 검증하는 절차를 거쳐 확정하였다.

## 5. 실험 결과

### 5.1 생성 결과

〈표 4〉는 언어모델별 MARC 레코드 생성 결과의 예시로, 연구방법에서 서지정보 추출 예시로 언급하였던 〈그림 2〉의 메타데이터 전체 생성 결과 중 부분을 발췌한 것이다.

데이터 필드에 한정하여 해당 서지정보에서 생성한 레코드의 행 수를 확인한 결과, 가장 행 수를 많이 모델은 HyperCLOVA로 총 37행이었으며, GPT가 12행으로 가장 적은 레코드 수를 생성하였다. 이 서지정보에 대응하는 LC의 MARC 레코드는 총 37행으로 HyperCLOVA와 동일한 행 수를 보였으나, 필드 구성에서는 차이가 존재하였다. HyperCLOVA의 경우, 실제로 존재하지 않는 필드를 생성하여 레코드를 등 MARC 필드의 주요 구분을 정확히 파악하지 못하는 경향을 보였는데, 〈표 4〉에서의 마지막 필드인 990은 존재하지 않는 필드 번호임을 통해 이러한 한계를 확인할 수 있다.

한편, 본 연구에서는 각 언어모델에게 무작위 MARC 레코드를 생성하도록 지시한 것이 아니라, 자동화 목록 생성을 위한 목록규칙과 각 자동화 목록 형식의 지침을 적용하여 MARC 레코드를 생성하도록 프롬프트를 구성하였으며, 그 결과 언어모델들은 필드와 지시기호, 식별기호, 구두점 등을 적용하여 MARC 레코드를 생성한 것으로 확인되었다. 그럼에도 불구하고 HyperCLOVA는 지시기호와 구두점의 존재 자체는 인지하고 있으나, 사용방식이 다소 어색하였으며, 식별기호에 대해서는 적절히 파악하지 못하는 경향을 보였다.

이에 따라 각 언어모델이 생성한 레코드별 필드 수는 〈표 5〉와 같다. HyperCLOVA를 제외한 모든 모델은 정답 레코드보다 적은 레코드 수를 생성한 반면, HyperCLOVA는 앞선 사례에서 나타난 것처럼 필드의 의미나 구성을 충분히 이해하지 못하여, 일부 서지정보에서는 1000번대 이상의 필드 번호를 구성하는 등의 MARC 레코드 생성에 오류를 드러냈다. 이와 같은 오류로 인해 상대적으로 많은 필드의 레코드가 생성되었으며, 그 수는 정답 레코드와 비교하면 약 7행 이상 많은 것으로 나타났다.

### 5.2 전체 성능

〈표 6〉은 언어모델이 생성한 결과에 대한 전체 성능으로, 정답과 일치하는 레코드를 생성한 건에 대하여 완전성, 생성한 레코드에 중에서 정보원에 기반하여 내용으로 정확하게 기술한 건에 대하여 정확성, 그리고 이들 중 지시기호, 식별기호를 적절히 사용한 경우를 기준으로 규칙성을 파악하여 그 생성 성능을 평가하였다.

그 결과, 완전성 측면에서 Gemini가 약 0.88

〈표 4〉 언어모델별 MARC 레코드 생성 결과 예시

| AI type | field | ind | content |
|---|---|---|---|
| HC | 020 | | 978-3-11-171222-2 2025 |
| | 260 | 1 | de gruyter berlin/boston genthiner straße 13 10785 berlin |
| | 500 | | general and inorganic chemistry: theoretical foundations, main group elements, transition metals, lanthanides, and actinides / irena kostova |
| | 650 | 0 | #general and inorganic chemistry #graduate |
| | ... | | |
| | 990 | 0 | #minneapolis minnesota |
| EX | 100 | 1 | $a Kostova, Irena |
| | 245 | 0 | $a General and inorganic chemistry $b Theoretical foundations, main group elements, transition metals, lanthanides, and actinides $i Irena Kostova |
| | 260 | 0 | $a Berlin: Walter de Gruyter GmbH, 2025 |
| | 264 | 1 | $a Berlin: Walter de Gruyter GmbH, 2025 |
| | ... | | |
| | 520 | 0 | $a CPI books GmbH, Leck |
| AX | 020 | | $a978-3-11-171222-2 $q(print) |
| | 100 | 1 | $aKostova, Irena, $d[date] $eauthor. |
| | 245 | 10 | $aGeneral and inorganic chemistry: $btheoretical foundations, main group elements, transition metals, lanthanides, and actinides /$cIrena Kostova. |
| | 260 | | $aBerlin :$aBoston: $bWalter de Gruyter GmbH, $c2025. |
| | ... | | |
| | 700 | 2 | $aWalter de Gruyter GmbH, $epublisher. |
| GP | 020 | | $a 978-3-11-171222-2 |
| | 100 | 1 | $a Kostova, Irena, $e author. |
| | 245 | 10 | $a GENERAL AND INORGANIC CHEMISTRY: $b THEORETICAL FOUNDATIONS, MAIN GROUP ELEMENTS, TRANSITION METALS, LANTHANIDES, AND ACTINIDES / $c Irena Kostova. |
| | 264 | 1 | $a Berlin/Boston: $b Walter de Gruyter GmbH, $c [2025]. |
| | ... | | |
| | 500 | | $a Printing and binding: CPI books GmbH, Leck. |
| GM | 020 | | $a 9783111712222 |
| | 100 | 1 | $a Kostova, Irena, $e author. |
| | 245 | 10 | $a GENERAL AND INORGANIC CHEMISTRY: $b THEORETICAL FOUNDATIONS, MAIN GROUP ELEMENTS, TRANSITION METALS, LANTHANIDES, AND ACTINIDES /$c Irena Kostova. |
| | 264 | 1 | $a Berlin/Boston: $b Walter de Gruyter GmbH, $c [2025] |
| | ... | | |
| | 710 | 2 | $a Integra Software Services Pvt. Ltd., $e typesetter. |
| GR | 020 | | $a 9783111712222 |
| | 100 | 1 | $a Kostova, Irena, $e author. |
| | 245 | 10 | $a General and inorganic chemistry: $b theoretical foundations, main group elements, transition metals, lanthanides, and actinides / $c Irena Kostova. |
| | 264 | 1 | $a Berlin : $a Boston: $b Walter de Gruyter GmbH, $c 2025. |
| | ... | | |
| | 830 | 0 | $a De Gruyter graduate. |

〈표 5〉 언어모델이 생성한 MARC 레코드의 평균 필드 수

| | HC | EX | AX | GP | GM | GR | 정답 |
|---|---|---|---|---|---|---|---|
| 국외문학 | 47.5 | 14.5 | 30.6 | 21 | 17.8 | 21.7 | 33.9 |
| 국외과학 | 39.6 | 12.4 | 28.1 | 19.5 | 18.2 | 18.1 | 45.6 |
| 국내문학 | 26.9 | 12.2 | 24 | 22.9 | 10.7 | 9.9 | 16.5 |
| 국내과학 | 26.8 | 10.1 | 22.1 | 20.8 | 8 | 9.2 | 13.6 |
| 평균 생성 레코드 수 | 35.2 | 12.3 | 26.2 | 21.05 | 13.675 | 14.725 | 27.4 |

〈표 6〉 언어모델이 생성한 MARC 레코드에 대한 전체 성능

| 척도 \ AI 유형 | HC | EX | AX | GP | GM | GR |
|---|---|---|---|---|---|---|
| 완전성 | 0.3754 | 0.3754 | 0.6540 | 0.8768 | 0.8827 | 0.8123 |
| 정확성 | 0.7891 | 0.8594 | 0.8318 | 0.9197 | 0.9136 | 0.9711 |
| 규칙성 | 0.4951 | 0.5273 | 0.8922 | 0.8800 | 0.8818 | 0.9410 |

로 가장 높은 성능을 보였으며, 정확성과 규칙성 측면에서는 Grok이 각각 약 0.97과 0.94로 가장 우수한 성능을 나타냈다. 즉, 정답에 근접한 필드로 레코드를 구성하는 능력은 Gemini가 가장 뛰어났고, Grok은 완전성에서는 3순위지만, 생성한 레코드의 내용 정확성과 그 필드 작성방식 측면에서는 타 언어모델보다 상대적으로 우수하게 학습되어있다고 판단할 수 있다.

한편, HyperCLOVA는 정답과 일치하는 필드를 생성하는 완전성에서 0.4 미만의 낮은 성능을 보였으나, 생성된 필드에 올바른 내용을 추출하여 입력하는 정확성은 약 0.79로 나타났다. 반면, 지시기호, 식별기호 등의 적용방식과 관련된 규칙성은 0.5 미만으로 전체 언어모델 중 각 기준마다 가장 낮은 성능을 보였다.

EXAONE의 경우 완전성 측면에서는 HyperCLOVA와 유사한 수준을 보였으나, 내용 입력과 규칙 적용에서는 HyperCLOVA보다 상위 수준의 성능을 확인할 수 있었다. 국내 언어모델 가운데 A.X는 세 기준 모두에서 가장 안정적인 성능을 보였으며, 완전성에서 약 0.65로 정답 필드의 절반 이상을 일치하게 생성하였다. 규칙성 측면에서도 GPT나 Gemini보다 약 0.01 정도 우수한 성능을 보여, 전반적으로 필드 입력방식을 비교적 충실히 준수하고 있음을 알 수 있다.

이에 따라, 〈표 7〉은 MARC21과 KORMARC 레코드로 구분한 성능 결과이며, 전체 성능과의 비교는 〈그림 4〉와 같다.

국내외 도서를 구분하지 않고 살펴보면, 모든 평가 기준에서 0.7 이상의 성능을 보인 모델은 GPT였다. 이에 GPT는 국내 도서 기반의 MARC 레코드를 생성할 때, 완전성 측면에서 0.97 이상의 값을 보여, 정답과 매우 유사한 필드 구성을 가진 것으로 파악되었다. 다만 규칙성에 있어 국외 도서를 대상으로 한 MARC21 레코드를 생성할 때보다 약 0.2 정도 낮은 성능

〈표 7〉 언어모델이 생성한 MARC 레코드를 국내외로 구분한 성능

| 척도 \ AI 유형 | | HC | EX | AX | GP | GM | GR |
|---|---|---|---|---|---|---|---|
| MARC21 | 완전성 | 0.2692 | 0.3889 | 0.6838 | 0.8761 | 0.8419 | 0.9316 |
| | 정확성 | 0.8889 | 0.9890 | 0.8844 | 0.9610 | 0.9645 | 0.9908 |
| | 규칙성 | 0.5179 | 0.5333 | 0.9293 | 0.9416 | 0.9474 | 0.9630 |
| KOR MARC | 완전성 | 0.6075 | 0.3458 | 0.5888 | 0.9720 | 0.8785 | 0.5514 |
| | 정확성 | 0.6923 | 0.5405 | 0.6825 | 0.8173 | 0.8191 | 0.9830 |
| | 규칙성 | 0.4667 | 0.5000 | 0.7674 | 0.7353 | 0.6883 | 0.7759 |


〈그림 4〉 전체 성능과의 비교: MARC21 레코드 생성(위), KORMARC 레코드 생성(아래)

을 보여, 한국목록규칙과 KORMARC 규칙 적용에 있어 다소 미흡한 점을 나타냈다. 이러한 경향은 국외 언어모델인 Gemini와 Grok에서도 유사하게 나타났는데, Grok은 MARC21 레코드 생성에서 0.93 이상의 완전성 보인 반면, KORMARC 레코드의 필드생성에서는 0.55 수준에 그쳐 국내 도서에 대해서는 상대적으로 더 취약한 성능을 보이는 것으로 나타났다.

한편, 전체 성능의 모든 측면에서 가장 낮은 값을 보인 HyperCLOVA는 KORMARC 레코드 생성의 경우, MARC21 레코드 생성에 비해 정답과 유사한 필드를 상대적으로 더 많이 생성하는 것으로 보이나, 정확성과 규칙성에서는 MARC21 레코드 생성 때 보다 낮은 성능을 나타내었다. 이는 HyperCLOVA가 필드의 의미나 구성을 충분히 고려하지 않고 굉장히 많은 필드를 생성하는 특성에 기인하는 것으로 추정된다. 그 결과 정답과 일치하는 필드가 생성된 사례 자체는 많았으나, 필드의 적절한 사용이라는 관점에서는 타 모델에 비하여 부족한 양상을 보였다. 그럼에도 불구하고 국외 도서에 대한 정확성이 비교적 높게 나타난 것은 국내 데이터보다 국외 데이터가 학습 데이터로 더 풍부하게 제공되어 사례 기반 학습이 가능했기 때문으로 해석할 수 있다.

EXAONE이나 A.X의 경우, 모두 전체 성능과 비교하면 MARC21 레코드 생성 성능이 더 높게 나타났으며, KORMARC 레코드 생성 성능은 상대적으로 낮은 경향을 보였다. 이는 HyperCLOVA와 마찬가지로, 학습 데이터가 국내보다 국외 데이터에 편중되어 있어 국내 데이터생성을 위한 사례학습이 상대적으로 부족하므로 나타난 현상일 것으로 여겨진다.

### 5.3 필드별 성능

〈표 8〉은 앞서 평가에서 언급한 〈표 2〉의 사용 가능 필드 중 주요 필드별 성능을 살펴본 것이다.

표제 및 책임표시사항에 해당하는 245 필드는 HyperCLOVA를 제외한 모든 언어모델이 생성하였으며, 언어모델 대부분이 해당 필드에 정확한 정보를 입력한 것으로 확인되었다. 다만 A.X는 타 언어모델과 마찬가지로 총 40건의 필드를 생성하였으나, 이 중 1건에서 저자 기입의 오류가 발생하여 정확성은 0.975로 나타났다. 규칙성 측면에서 Gemini와 Grok, A.X이 지시기호나 식별기호 작성 규칙을 비교적 충실히 준수한 반면, GPT는 이들 보다 약 0.02 정도 성능이 낮았으며, EXAONE의 경우에는 전체 사례 중 대략 절반 정도만 규칙에 부합하는 방식으로 필드를 생성한 것으로 나타났다.

국내의 경우 발행, 배포, 간사사항에 있어 260 필드에, 국외의 경우에는 생산, 발행, 배포, 제작, 저작권 표시인 264 필드에 기술한다. 이에 따라 결과를 살펴보면, 국내 언어모델인 HyperCLOVA, EXAONE, A.X는 모두 264 필드에 비해 260필들 생성 성능이 더 높게 나타났다. 특히 HyperCLOVA는 264필들을 전혀 생성하지 못하여, 국외 도서를 대상으로 한 MARC21 레코드에서 발행 관련 사항 정보를 전혀 기술하지 못한 것으로 확인되었다. 한편, GPT는 260과 264필드를 모두 생성하였으며, 완전성과 정확성 측면에서 두 필드 모두 관련 정보를 비교적 정확하게 입력한 것으로 파악되었다. Gemini는 국외 도서에서 저작권 사항을 누락한 사례가 있어 완전성 측면에서 GPT보다 낮은 성능을 보

〈표 8〉 주요 필드별 성능

| 필드 | AI유형 / 척도 | HC | EX | AX | GP | GM | GR |
|---|---|---|---|---|---|---|---|
| 020 | 완전성 | 0.5000 | 0 | 0.6571 | 0.9857 | 0.7714 | 0.9857 |
| | 정확성 | 0.8857 | 0 | 0.9130 | 0.9565 | 0.9259 | 0.9855 |
| | 규칙성 | 0.4839 | 0 | 1 | 0.8333 | 0.7900 | 0.9853 |
| 100 | 완전성 | 0.0556 | 0.9444 | 0.9444 | 0.8889 | 0.9444 | 0.9444 |
| | 정확성 | 0 | 0.9412 | 0.9412 | 1 | 1 | 1 |
| | 규칙성 | 0 | 0.5938 | 0.8750 | 1 | 1 | 1 |
| 245 | 완전성 | 0 | 1 | 1 | 1 | 1 | 1 |
| | 정확성 | 0 | 1 | 0.9750 | 1 | 1 | 1 |
| | 규칙성 | 0 | 0.5125 | 0.7179 | 0.6875 | 0.700 | 0.7250 |
| 260 | 완전성 | 0.9500 | 0.8500 | 0.7500 | 1 | 1 | 0.0500 |
| | 정확성 | 0 | 0 | 0.9333 | 1 | 1 | 1 |
| | 규칙성 | 0 | 0 | 1 | 0.9750 | 0.9750 | 1 |
| 264 | 완전성 | 0 | 0.6538 | 0.0385 | 1 | 0.9231 | 0.8462 |
| | 정확성 | 0 | 1 | 1 | 1 | 1 | 1 |
| | 규칙성 | 0 | 0.5 | 1 | 0.9808 | 0.9792 | 1 |
| 490 | 완전성 | 0.0625 | 0 | 0.0625 | 0.8125 | 0.9375 | 0.9375 |
| | 정확성 | 0 | 0 | 0.5000 | 1 | 1 | 1 |
| | 규칙성 | 0 | 0 | 0 | 0.6923 | 0.7000 | 0.5333 |
| 830 | 완전성 | 0.0833 | 0 | 0 | 0 | 1 | 0.3333 |
| | 정확성 | 0 | 0 | 0 | 0 | 0.8333 | 1 |
| | 규칙성 | 0 | 0 | 0 | 0 | 0.9000 | 1 |

였으나, 생성한 필드에 한해서는 정보 추출과 내용 입력이 정확하게 이루어진 것으로 나타났다. Grok의 경우 KORMARC 레코드에서 260 필드를 단 1건만 생성하여, 0.05에 그쳤으나, 해당 사례의 정확성과 규칙성은 적절한 것으로 확인되었다. 그러나 Grok이 발행 관련 필드를 전혀 생성하지 못한 것은 아니며, 국내 도서에서도 264필들을 사용하여 발행 관련 사항을 기술한 사례가 존재하므로, 이러한 필드 선택 경향이 위와 같은 성능 수치에 반영된 것으로 판단할 수 있다.

생성을 위해 제공한 도서에서 총서사항을 기술할 수 있는 사례는 국외 11건, 국내 5건이었으며, 이에 따른 성능을 살펴보았다. 우선 국외 도서의 정답 MARC21 레코드에는 총서사항 490 필드의 제1 지시기호를 0으로 기입하여 총서를 부출하지 않고 있어, 830 필드가 존재하지 않는 경우가 있었다. 반면 국내 도서의 정답인 KORMARC 레코드에는 490 필드의 제1 지시기호를 1로 설정하여 총서를 부출하고 있으며, 이에 따라 830 필드가 함께 작성되어 있다.

이러한 기준에 따라 성능을 분석한 결과, EXAONE은 총서 관련 필드 군을 전혀 생성하지 못하였고, HyperCLOVA는 해당 필드 군을 생성하기는 하였으나, 정확한 내용을 기술하지는 못하였다. 앞서 언급한 바와 같이 HyperCLOVA는

필드의 의미와 무관하게 세 자리 숫자를 나열하는 방식으로 MARC 레코드를 생성하는 경향을 보였고, 총서 관련 필드 역시 이러한 맥락에서 형성된 것으로, 실제 생성 결과는 1건에 불과하였다.

반면 AX도 역시 490 필드에서 1건을 생성하였으나, HyperCLOVA와는 달리 총서명 자체는 기입하였으나, 지시기호나 식별기호 사용은 적절하지 않았다. 나머지 총서사항 관련 사례들은 모두 440 필드에 기술하는 양상을 보여, 필드 선택과 사용방식에서 성능 차이가 존재하는 것으로 확인되었다.

이처럼 국내 언어모델들은 총서사항 관련 필드 생성에서 전반적으로 0에 가까운 수준의 성능을 보였으며, 총서사항을 입력하는 방식을 제대로 학습하지 못했거나, 더 이상 사용하지 않는 필드를 사용하는 등의 오기입 사례를 보여, 관련 규칙 학습의 부족한 것으로 판단하였다.

한편, GPT는 830 필드를 전혀 생성하지 않아서 이 필드생성에 대한 성능은 0인 것으로 나타났으나, 490 필드에 대한 완전성은 0.8로, 2건을 제외한 모든 총서사항을 추출하고 그 내용을 정확하게 생성하였다. 다만 규칙성에서는 0.7 미만의 성능을 보였는데, 실제로 GPT는 정답 레코드와는 달리 490 필드의 제1 지시기호로 1로 기입해 총서를 부출하고, 830 필드를 생성하여 총서를 부출하는 방식을 취하였다. 이러한 경향은 총서사항을 추출하지 못한 2건을 제외한 모든 국외 도서에서 공통으로 나타났으며, 이로 인해 정답과의 차이로 규칙성 점수가 낮게 산출되었으나, GPT가 MARC21에서의 총서사항 기입 규칙 자체는 인지하고 있음을 보여주는 결과로 해석할 수 있다. 그러나 KORMARC 레코드에서는 490 필드와 그 내용을 정확하게 추출하여 생성하였으나, 지시기호를 모두 누락하였다. 이를 통해 KORMARC 레코드와 MARC21 레코드를 구분하여 처리하고 있음은 알 수 있지만, KORMARC 레코드의 작성방식 측면에서는 MARC21에 비하여 다소 미흡한 것을 보였다.

이러한 경향은 Gemini나 Grok에서도 확인할 수 있었는데, 정확성에 비하여 규칙성 성능이 낮게 나타난 것은, 490 필드의 지시기호에 정답과 다른 값을 기입하고 부가적으로 830필드를 생성하여 부출 하는 방식 등을 사용하였기 때문이다. 특히 Gemini는 KORMARC 레코드 생성에 있어서도 제1 지시기호에 1로 기입하고, 830 필드를 생성하는 등 해당 용례를 비교적 정확히 활용하는 양상을 보였으며, Grok은 GPT와 마찬가지로 KORMARC 레코드 생성에서는 다소 오류를 보였다. 그럼에도 이 3종의 모델은 모두 총서사항에서 지시기호의 존재를 인지하고, 이를 MARC 레코드 생성에 활용하고 있다는 점에서 생성형 AI가 지닌 생성의 역량의 한 부분을 보여주는 결과라고 판단할 수 있다.

그러나 기본표목인 100 필드의 처리에서는 세 모델 모두 MARC21과 KORMARC 레코드 작성방식을 명확히 구분하지 못하는 양상을 보였다. 100 필드를 생성한 경우 해당 필드의 정확성과 규칙성 자체는 1에 가까운 비싼 값을 기록하였으나, 이는 KORMARC 레코드에서도 MARC21의 저자 주기입방식을 그대로 적용한 결과였다. 실제로 모델들은 KORMARC 레코드를 생성할 때에도 100 필드를 기본표목으로 설정하고 245 필드의 제1 지시기호를 '1'로 부출지시하는 패턴을 보였으며, 그 결과 국내 도

서데이터 20건 중 저자가 9명인 1건을 제외한 19건에서 모두 100 필드를 생성하였다.

한편, 입수사항에 해당하는 020 필드는 제공된 정보원에서 추출할 내용이 상대적으로 적으며, 추출 대상이 숫자 중심이라는 점 등을 고려할 때로 언어모델들에게 유리하게 작용할 것으로 예상하였으나, 020 필드가 반복사용이 가능하다는 특성과 부가기호 작성을 위한 지시기호의 오기입 등이 복합적으로 영향을 미치면서, 정확성과 규칙성 모두에서 다소 낮은 성능을 나타내었다. 이 중에서도 EXAONE은 020 필드를 단 1건도 생성하지 못하여, 해당 필드의 성능값이 0으로 나타났다.

지시기호와 식별기호의 대표적인 오류사례는 〈표 9〉와 같으며, 지시기호만을 오기입한 오류, 식별기호만 오기입한 오류, 지시기호와 식별기호를 모두 오기입한 경우로 구분할 수 있다. 특히 표제 및 책임표시사항에 해당하는 245 필드에서는 지시기호를 포함하여 식별기호를 함께 오기입하는 오류가 다수 확인되었는데, 이는 타 필드 보다 많은 식별기호와 지시기호가 사용된다는 점이 영향을 미친 결과로 추정된다.

따라서 이러한 결과를 종합하면 상용적으로 사용 중인 GPT, Gemini, Grok 모델은 MARC 레코드 생성에 있어 그 필드 사용의 의미와 작성방식을 비교적 구체적으로 인지하고 생성에 적용하고 있는 것으로 보인다. 다만 MARC21과 KORMARC의 구분은 다소 부정확하여, KORMARC를 생성할 때에도 MARC21 작성방식을 그대로 따르는 경향을 보였다. 그럼에도 주어진 정보에 대응하는 필드를 생성하고 내용을 정확하게 기입하며, 연계된 필드를 함께 생성하는 양상은 MARC21과 KORMARC

〈표 9〉 규칙 오류의 대표 사례

| | 생성 MARC21 | | 정답 MARC21 |
|---|---|---|---|
| | 필드 | 내용 | 내용 |
| 지시기호 오류 | 245 | 14$aScratching the surface: $bexploring Earth's layers / $cKate Allen Fox ; illustrated by Erin Brown. | 10$a Scratching the surface: $b exploring earth's layers / $c by Kate Allen Fox ; illustrated by Erin Brown. |
| 식별기호 오류 | 100 | 1␣$a Falligant, Erin, author. | 1␣$a Falligant, Erin, $e author. |
| 지시기호 식별기호 오류 | 245 | 00$a Energy reactions in the kitchen$i Ann McCallum Staats. | 10$a Energy reactions in the kitchen / $c Ann McCallum Staats, M.Ed. |

| | 생성 KORMARC | | 정답 KORMARC |
|---|---|---|---|
| | 필드 | 내용 | 내용 |
| 지시기호 오류 | 490<br>490 | 1␣ $a 그냥 홈치면 되지<br>␣␣ $a 그냥 홈치면 되지 | 10▼a위픽 =▼aWefic :▼v89 |
| 식별기호 오류 | 020<br>020 | ␣␣ $a 9791192549484$c 03370<br>␣␣ $a 978-89-6090-938-6 $d 03810 | ␣␣▼a9791192549484▼g03370:▼c\17500<br>␣␣▼a9788960909380▼g03810:▼c\17000 |
| 지시기호 식별기호 오류 | 245<br>245<br>260 | 10 $a 바르셀로나의 유서 /$c 백세희.<br>41$a앞으로 무엇을 더 이렇게 :$v91<br>00$a김학원$a(주)휴머니스트출판그룹 | 00▼a바르셀로나의 유서 /▼d지은이: 백세희<br>00▼a고백의 시대 /▼d지은이: 이현석<br>␣␣▼a서울:▼bH(휴머니스트출판그룹),▼c2025 |

필드 가운데 공통으로 용법이 유사하게 사용되는 경우 해당 필드를 높은 확률로 생성할 수 있음을 시사한다.

한편 국내 모델들은 앞선 3가지의 모델에 비해 모든 측면에서 현저히 낮은 성능을 보였으며, 필드 의미나 구성, 작성방식 측면에서도 상대적으로 미흡한 양상을 나타냈다. 이 가운데 HyperCLOVA는 020 필드를 제외한 모든 필드에서 정확성과 규칙성이 0으로 나타나, MARC 레코드에 대한 학습 데이터가 거의 없거나 관련 학습이 사실상 이루어지지 않은 것으로 추정된다.

## 6. 결 론

본 연구는 생성형 AI 모델 6종에 기반하여, 국내외 도서 40권을 대상으로 MARC 레코드를 생성하였으며 이를 통해 각 모델의 생성 현황과 그 성능을 비교, 분석하는 것을 목적으로 하였다. 이에 따른 결과는 다음과 같다.

첫째, GPT, Gemini, Grok 등의 언어모델 3종은 세 가지 평가 척도인 완전성, 정확성, 규칙성 전반에서 국내 소버린 AI 모델(HyperCLOVA, EXAONE, A.X)보다 일관되게 높은 성능을 보였다. 특히 이들은 레코드 대부분에서 정답과 유사한 필드 구성을 유지하면서도 지시기호, 식별기호 등 형식적 요소를 비교적 안정적으로 처리한, 반면, 국내 모델은 필드를 누락, 필드 번호의 오류, 지시기호 및 식별기호 사용 규칙의 위반이 빈번하게 나타나 서지 메타데이터 생성 측면에서 기술 성숙도가 GPT, Gemini, Grok에 비해 상대적으로 낮은 것으로 확인되었다.

둘째, GPT, Gemini, Grok은 MARC21 레코드 기술에서 높은 완전성과 규칙성을 보였으나, 동일 모델이라도 서지 기술 대상이 국내 도서로 전환되면, 필드 구성, 지시기호, 책임표시 방식 등에서 오류가 증가하여 성능이 상대적으로 저하되는 경향을 보였다. 한편, 국내 소버린 AI 모델은 전반적인 성능 수준이 낮게 나타났으며, 이는 국내도서를 대상으로 할지라도 성능 향상은 기대하기 어려운 것으로 분석되었다.

셋째, 필드별로 살펴보면, 표제와 책임표시사항(245)처럼 모델 대부분이 안정적으로 생성할 수 있는 필드가 있는 반면, 발행, 배포, 간사 사항 혹은 생산, 발행, 배포, 제작, 저작권 표시(260/264), 총서사항(490/830) 등 규칙 의존도가 높은 필드에서는 모델 간 성능 차이와 오류 유형이 크게 확대되었다. HyperCLOVA와 같은 일부 모델은 존재하지 않는 가공의 필드 번호를 생성하였으며, 이외 타 모델들은 MARC21의 총서 처리 방식을 KORMARC에 그대로 적용하는 등, 서지 규칙에 대한 구조적 이해 부족을 드러냈으며, 특히 총서・부출・표목 관련 필드에서 지시기호, 식별기호, 책임표시 구성이 맞지 않는 사례가 등장하였다.

이러한 결과는 현시점에서 생성형 AI를 도서관 메타데이터 업무에 도입할 때, 이를 전면적인 자동목록 도구로 활용하기보다는 서지 레코드 초안 생성과 오류 탐지・보완을 지원하는 보조 도구로 활용하는 것이 타당함을 시사한다.

GPT, Gemini, Grok 등 글로벌 생성형 AI는 MARC21 중심의 규칙을 상당 부분 학습하여 내면화하고 있음에도, KORMARC 및 한국목록규칙을 완전하게 구분하거나 적용하지 못하는 한계를 보였다. 또한, 국내 소버린 AI는 전

반적인 성능과 규칙 준수 면에서 아직은 준비 단계에 머물러 있는 것으로 나타났다. 따라서 현재 생성형 AI가 다양한 산업 분야에 적용 되고 있는 현시점에도, 정보조직 업무만큼은 생성형 AI가 제안한 결과를 사서가 검증하고 최종적으로 판단하는 절차가 필수적으로 수반되어야 하며, 나아가 이와 같은 검증 절차를 전제로, 고도의 추론 능력을 갖춘 생성형 AI와 사서가 협력하는 업무 모델이 가장 현실적인 활용 방향으로 판단된다.

더불어 국내 소버린 AI의 안정적인 서지 메타데이터 생성 성능을 확보하기 위해서는 서지 데이터를 기반으로 한 체계적인 학습이 선행되어야 하며, 도서관형 소버린AI 구현을 위해서는 학습 데이터 선별이 주요한 과제로 여겨진다.

본 연구는 일부 주제에 한정된 국내외 도서 40종을 대상으로 실험하였기에, 보다 다양한 주제와 정보자원으로 생성 성능을 일반화하기에는 한계가 있다. 또한 제공 정보원을 표지와 판권기로 한정하였으므로, 추가적인 정보원 제공에 따른 성능은 파악하지 못하였다. 더불어 본 연구에서는 주요 필드 군을 중심으로 성능을 평가하였으므로, 보다 세분화된 필드 전반에 대한 구체적인 성능 분석에는 제약이 있었음도 한계로 지적할 수 있다.

그럼에도 불구하고 본 연구는 여러 종류의 생성형 AI 모델을 대상으로 MARC 레코드 생성 방식과 필드별 성능을 완전성, 정확성, 규칙성이라는 세 가지 척도에 따라 체계적으로 비교 및 분석했다는 점에서 의의를 지닌다. 특히 모델별 경향과 한계를 파악하고 도출함으로써, 향후 도서관이 특정 업무 영역에서 어떠한 유형의 모델을 우선 적용할지 판단하는데 근거를 제공할 수 있을 것이며, 나아가 도서관 정책에 있어, 다양한 유형의 자원을 대상으로 생성형 AI를 혼합 사용하거나 선택적 활용 전략을 구축할 때 또한 도서관 맞춤형 소버린 AI 전략을 설계하는 과정에서의 기초자료로 활용될 수 있을 것으로 기대한다.

## 참 고 문 헌

과학기술정보통신부 (2024. 5. 13.). 대한민국 인공지능 3대 강국 도약을 위해, 대한민국 대표 인공지능 연구거점 구축 추진. 출처: https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&bbsSeqNo=94&nttSeqNo=3184486

과학기술정보통신부 (2025. 9. 9.). 「독자 인공 지능 기초 모형(AI 파운데이션 모델)」 사업(프로젝트) 착수식 개최. 출처: https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&bbsSeqNo=94&nttSeqNo=3186227

김선욱, 이혜경, 이용구 (2023). ChatGPT가 자동 생성한 더블린 코어 메타데이터의 품질 평가: 국내 도서를 대상으로. 정보관리학회지, 40(2), 183-209.

https://doi.org/10.3743/KOSIM.2023.40.2.183

김주용, 신판섭 (2021). KORMARC로 표현된 공공도서관 서지 데이터의 BIBFRAME 변환 연구. 한국컴퓨터정보학회논문지, 26(11), 139-147. https://doi.org/10.9708/jksci.2021.26.11.139

이미화 (2015). 국내 하이브리드 서지레코드 생성 방안에 관한 연구. 한국문헌정보학회지, 49(4), 203-220. https://doi.org/10.4275/KSLIS.2015.49.4.203

이용구 (2023). BERT 모형을 이용한 주제명 자동 분류 연구. 한국문헌정보학회지, 57(2), 435-452. https://doi.org/10.4275/KSLIS.2023.57.2.435

정준화 (2025. 5. 29.). 인공지능(AI) 3대 강국 도약을 위한 AI 정부조직 과제 (이슈와 논점 제2414호). 대한민국. 국회. 국회입법조사처. 출처: https://www.nars.go.kr/report/view.do?cmsCode=CM0018&brdSeq=48189

한국은행 (2025. 8. 18.). AI의 빠른 확산과 생산성 효과: 가계조사를 바탕으로 (BOK 이슈노트 제2025-22호). 출처: https://www.bok.or.kr/portal/bbs/P0002353/view.do?nttId=10093071&searchCnd=1&searchKwd=&depth=201150&pageUnit=10&pageIndex=1&programType=newsData&menuNo=200433&oldMenuNo=201150

Artificial Analysis (2025). Artificial Analysis LLM Performance Leaderboard. Available: https://huggingface.co/spaces/ArtificialAnalysis/LLM-Performance-Leaderboard

Bommasani, R. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258

Brador, I. (2024, November 5). Could artificial intelligence help catalog thousands of digital library books? An interview with Abigail Potter and Caroline Saccucci. The Signal. Available: https://blogs.loc.gov/thesignal/2024/11/could-artificial-intelligence-help-catalog-thousands-of-digital-library-books-an-interview-with-abigail-potter-and-Caroline-saccucci/

Brown, R. (2025, March 17). AI that can match humans at any task will be here in five to 10 years, Google DeepMind CEO says. CNBC. Available: https://www.cnbc.com/2025/03/17/human-level-ai-will-be-here-in-5-to-10-years-deepmind-ceo-says.html

Chen, H., Jiao, F., Li, X., Qin, C., Ravaut, M., Zhao, R., Xiong, C., & Joty, S. R. (2023). ChatGPT's One-year anniversary: are open-source large language models catching up? ArXiv, arXiv:2311.16989. https://doi.org/10.48550/arXiv.2311.16989

Chiang, W., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., & Xing, E. P. (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. lmsys.org. Available:

https://lmsys.org/blog/2023-03-30-vicuna/

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N., …, Ramabhadran, B. (2025). Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities[Preprint]. arXiv. https://arxiv.org/abs/2507.06261

D'Souza, J., Sadruddin, S., Israel, H., Begoin, M., & Slawig, D. (2025). SemEval-2025 task 5: LLMs4Subjects -- LLM-based automated subject tagging for a national technical library's open-access catalog. ArXiv, arXiv:2504.07199. https://doi.org/10.48550/arXiv.2504.07199

Dale, R. (2025). Sovereign AI in 2025. Natural Language Processing, 31(5), 1312-1321. https://doi.org/10.1017/nlp.2025.10007

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: efficient finetuning of quantized LLMs. ArXiv, arXiv:2305.14314. https://doi.org/10.48550/arXiv.2305.14314

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., & Hashimoto, T. (2023). AlpacaFarm: a simulation framework for methods that learn from human feedback. ArXiv, arXiv:2305.14387. https://doi.org/10.48550/arXiv.2305.14387

Fedus, W., Zoph, B., & Shazeer, N. M. (2021). Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. ArXiv, arXiv:2101.03961. https://doi.org/10.48550/arXiv.2101.03961

Gemini Team, Google (2024). Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. ArXiv, abs/2403.05530. https://doi.org/10.48550/arXiv.2403.05530

Gemini Team, Google (2025). Gemini: a family of highly capable multimodal models(Version 5). ArXiv, abs/2312.11805 arXiv. https://doi.org/10.48550/arXiv.2312.11805

Golub, K. (2019). Automatic subject indexing of text. Knowledge Organization, 46(2), 104-121. https://doi.org/10.5771/0943-7444-2019-2-104

Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., Zhang, W., Chen, G., Bi, X., Wu, Y., Li, Y. K., Luo, F., Xiong, Y., & Liang, W. (2024). DeepSeek-coder: when the large language model meets programming - the rise of code intelligence. ArXiv, arXiv:2401.14196. https://doi.org/10.48550/arXiv.2401.14196

Huo, W., Feng, X., Huang, Y., Fu, C., Li, B., Ye, Y., Zhang, Z., Tu, D., Tang, D., Lu, Y., Wang, H., & Qin, B. (2025). Enhancing Non-English capabilities of English-Centric large language models through deep supervision fine-tuning. Proceedings of the AAAI Conference on Artificial Intelligence, 39(23), 24185-24193.

https://doi.org/10.1609/aaai.v39i23.34594

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2024). Mixtral of experts. ArXiv, arXiv:2401.04088. https://doi.org/10.48550/arXiv.2401.04088

Kaplan, J., McCandlish, S., Henighan, T. J., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. ArXiv, arXiv:2001.08361. https://doi.org/10.48550/arXiv.2001.08361

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: a scoping literature review. Computers and Education: Artificial Intelligence, 3, 100101. https://doi.org/10.1016/j.caeai.2022.100101

LG AI Research (2024). EXAONE 3.0 7.8B instruction tuned language model. ArXiv, arXiv: 2408.03541. https://doi.org/10.48550/arXiv.2408.03541

LG AI Research (2025). EXAONE 4.0: Unified large language models integrating non-reasoning and reasoning modes. ArXiv, arXiv:2507.11407. https://doi.org/10.48550/arXiv.2507.11407

Li, Y., Bubeck, S., Eldan, R., Giorno, A. D., Gunasekar, S., & Lee, Y. T. (2023). Textbooks Are All You Need II: phi-1.5 technical report. ArXiv, arXiv:2309.05463. https://doi.org/10.48550/arXiv.2309.05463

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics. ArXiv, arXiv:2210.10341. https://doi.org/10.1093/bib/bbac409

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M. A., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: A Survey. ArXiv, arXiv:2402.06196. https://doi.org/10.48550/arXiv.2402.06196

Mumuni, A. & Mumuni, F. (2025). Large language models for artificial general intelligence (AGI): a survey of foundational principles and approaches. ArXiv, arXiv:2501.03151. https://doi.org/10.48550/arXiv.2501.03151

NAVER Cloud HyperCLOVA X Team (2025). HyperCLOVA X THINK technical report. ArXiv, arXiv:2506.22403. https://doi.org/10.48550/arXiv.2506.22403

Nellis, S. (2024, March 2). Nvidia CEO says AI could pass human tests in five years. Reuters. https://www.reuters.com/technology/nvidia-ceo-says-ai-could-pass-human-tests-five-years-2024-03-01/

NVIDIA (2025, October 30). NVIDIA, South Korea government and industrial giants build AI infrastructure and ecosystem to fuel Korea innovation, industries and jobs. NVIDIA Newsroom. https://nvidianews.nvidia.com/news/south-korea-ai-infrastructure

Pappu, A. (2024). 3 new ways to stay productive with Gemini for Google Workspace. Google Blog. Available: https://blog.google/products/workspace/google-gemini-workspace-may-2024-updates/

Poley, C., Uhlmann, S., Busse, F., Jacobs, J.-H., Kähler, M., Nagelschmidt, M., & Schumacher, M. (2025). Automatic subject cataloguing at the German National Library. LIBER Quarterly: The Journal of the Association of European Research Libraries, 35(1), 1-29. https://doi.org/10.53377/lq.19422

Reuters (2025, November 7). US to block Nvidia's sale of scaled-down AI chips to China, The Information reports. Available: https://www.reuters.com/world/china/us-block-nvidias-sale-scaled-back-ai-chips-china-information-says-2025-11-07/

Shao, Z., Dai, D., Guo, D., Liu, B., Wang, Z., & Xin, H. (2024). DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model. ArXiv, arXiv:2405.04434. https://doi.org/10.48550/arXiv.2405.04434

South Korea's brain drain: Why top talent is leaving (2025, July 7). The Straits Times. Available: https://www.straitstimes.com/asia/east-asia/south-koreas-brain-drain-why-top-talent-is-leaving

Southworth, J., Migliaccio, K., Glover, J., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for AI across the curriculum: transforming the higher education landscape via innovation in AI literacy. Computers and Education: Artificial Intelligence, 4, 100127. https://doi.org/10.1016/j.caeai.2023.100127

Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms.LIBER Quarterly: The Journal of the Association of European Research Libraries, 29(1), 1-25. https://doi.org/10.18352/lq.10285

Taniguchi, S. (2024). Creating and evaluating MARC 21 bibliographic records using ChatGPT. Cataloging & Classification Quarterly, 62(5), 527-546. https://doi.org/10.1080/01639374.2024.2394513

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Roziere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: open and efficient foundation language models. ArXiv, arXiv:2302.13971.

https://doi.org/10.48550/arXiv.2302.13971

Visser, T. P. d., Klein, H., & Le Fichant, E. (2023, August 22). Utopia, threat or opportunity first? artificial intelligence and machine learning for cataloguing. 88th ILFA World Library and Information Congress, Rotterdam, Netherlands.

What does a leaked Google memo reveal about the future of AI. (2023, May 11). The Economist. Available: https://www.economist.com/leaders/2023/05/11/what-does-a-leaked-google-memo-reveal-about-the-future-of-ai

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: a large language model for finance. ArXiv, arXiv:2303.17564. https://doi.org/10.48550/arXiv.2303.17564

Yan Tao, Olga Viberg, Ryan S Baker, & René F Kizilcec (2024). Cultural bias and cultural alignment of large language models, PNAS Nexus, 3(9), 346. https://doi.org/10.1093/pnasnexus/pgae346

**• 국문 참고자료의 영어 표기**
(English translation / romanization of references originally written in Korean)

Bank of Korea (2025, August 18). Rapid spread of AI and productivity effects: Based on household surveys (BOK Issue Note, No. 2025-22). Available: https://www.bok.or.kr/portal/bbs/P0002353/view.do?nttId=10093071&searchCnd=1&searchKwd=&depth=201150&pageUnit=10&pageIndex=1&programType=newsData&menuNo=200433&oldMenuNo=201150

Jeong, Junhwa (2025, September 29). Tasks for AI government organizations to leap forward as one of the top 3 AI powers(Issues and Points, No. 2414). Republic of Korea, National Assembly, National Assembly Research Service. Available: https://www.nars.go.kr/report/view.do?cmsCode=CM0018&brdSeq=48189

Kim, JooYong & Shin, PanSeop (2021). A study on converting bibliographic data of public libraries expressed in KORMARC into BIBFARME. Journal of the Korea Society of Computer and Information, 26(11), 139-147. https://doi.org/10.9708/jksci.2021.26.11.139

Kim, SeonWook, Lee, Hyekyung, & Lee, Yong-Gu (2023). Quality evaluation of automatically generated metadata Using ChatGPT: focusing on Dublin Core for korean monographs. Journal of the Korean Society for Information Management, 40(2), 183-209.

https://doi.org/10.3743/KOSIM.2023.40.2.183

Lee, Mihwa (2015). A study on the creation of hybrid bibliographic records. Journal of the Korean Society for Library and Information Science, 49(4), 203-220. https://doi.org/10.4275/KSLIS.2015.49.4.203

Lee, Yong-Gu (2023). A study on automatic classification of subject headings using BERT Model. Journal of the Korean Society for Library and Information Science, 57(2), 435-452. https://doi.org/10.4275/KSLIS.2023.57.2.435

Ministry of Science and ICT (2024, May 13). Promoting the establishment of a representative AI research hub in Korea to leap forward as one of the top 3 AI powers. Available: https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&bbsSeqNo=94&nttSeqNo=3184486

Ministry of Science and ICT (2025, September 9). Launch ceremony held for the “Independent Artificial Intelligence Basic Model (AI Foundation Model)” project. Available: https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&bbsSeqNo=94&nttSeqNo=3186227