

생성형 AI 기반 도서관 운영평가 자동화 가능성에 대한 기초연구

An Exploratory Study on the Feasibility of Using Generative AI to Automate Library Performance Evaluation

나민경 (Min Kyung Na)*

오지은 (Ji Eun Oh)**

이지연 (Jee Yeon Lee)***

초 록

본 연구는 공공도서관 운영평가에 생성형 AI 기반 자동평가를 적용해 보고, 그 도입 가능성을 검토하는 기초연구이다. 이를 위해 현행 공공도서관 운영평가 지표 중 증빙자료를 근거로 사람이 평가하는 4개 영역(공간, 협력, 경영계획, 우수사례)의 문헌정보학 분야 전문가와 생성형 AI 평가 결과를 비교 분석하였으며, 프롬프트 엔지니어링 기법에 따른 신뢰도 변화를 확인하였다. 2024년 서울시 공공도서관 운영평가를 위해 평가 대상 공공도서관 164개 관이 제출한 증빙자료를 대상으로 ChatGPT 5.1을 활용하여 평가를 진행했으며, 그 결과 비교적 평가 내용이 간단하고 평가 척도가 명확한 공간, 협력, 경영계획 영역에서는 전문가와 AI 평가점수 간 일치도가 높은 것으로 나타났다. 반면, 정성평가에 해당하는 우수사례 영역에서는 전문가와 AI 평가 결과 간 차이가 큰 것으로 나타났다. 또한, 평가에 필요한 정보를 구조화하여 입력하는 Task Information(TI)과 예시를 제시하는 Demonstration Information(DI) 프롬프트를 조합하는 조건에서 전문가와 AI 간 신뢰도가 가장 높은 것으로 확인되었으며, 특히 정성평가 영역에서 DI 프롬프트를 추가했을 때 신뢰도가 크게 향상하는 것으로 나타났다.

ABSTRACT

This study is an exploratory research that applies generative AI-based automated assessment to public library performance evaluation and examines its feasibility for adoption. To this end, we compared the evaluation results produced by a human expert in library and information science and by a generative AI system. The comparison focused on four domains of the current evaluation indicators that are scored by humans on the basis of submitted documents: space, collaboration, management planning, and best practices, and examined changes in reliability according to different prompt-engineering techniques. Using ChatGPT 5.1, we conducted automated evaluations on the documents submitted by 164 public libraries in Seoul for the 2024 public library performance evaluation. The results indicated that for domains with relatively simple content and clearly defined rating scales—space, collaboration, and management planning—the agreement between expert and AI scores was high. In contrast, in the best practices domain, which requires qualitative judgment, the discrepancy between expert and AI evaluation results was substantial. Furthermore, the highest level of reliability between expert and AI scores was observed under the condition that combined Task Information (TI) prompts, which provide structured input of the information required for evaluation, with Demonstration Information (DI) prompts, which offer illustrative examples. In particular, in the qualitative assessment domain, reliability improved significantly when DI prompts were added.

키워드: 생성형 AI, 공공도서관 운영평가, 평가 자동화, 인공지능 기반 평가, 프롬프트

Generative AI, library performance evaluation, evaluation automation, AI based assessment, prompt

* 연세대학교 대학도서관발전연구소 전임연구원(keren7337@yonsei.ac.kr) (제1저자)

** 서울도서관 관장(oserome@seoul.go.kr) (공동저자)

*** 연세대학교 문헌정보학과 교수(jlee01@yonsei.ac.kr) (교신저자)

■ 논문접수일자: 2025년 11월 22일 ■ 최초심사일자: 2025년 12월 4일 ■ 게재확정일자: 2025년 12월 8일

■ 정보관리학회지, 42(4), 277-301, 2025. <http://dx.doi.org/10.3743/KOSIM.2025.42.4.277>

※ Copyright © 2025 Korean Society for Information Management

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. 서론

공공도서관 운영평가는 「도서관법」 제37조 및 같은 법 시행령 제29조에 추진 근거를 두고 매년 시행되며 공공도서관의 시설과 인력, 도서관 자료 등 도서관 운영에 관한 사항을 평가한다. 평가의 효과성을 높이고 궁극적으로 공공도서관의 성장을 지원하는 수단으로 기능하기 위해 평가지표에 대한 다양한 논의가 진행됐으며(김홍렬, 2013), 그 과정에서 양적 지표만으로는 평가가 어려운 도서관의 특성을 고려하여 정성평가가 도입되었다. 이후에도 다양한 의견과 사례를 반영하여 평가 항목과 반영 비율 등을 변경하면서 도서관 현장을 파악하고 평가의 궁극적 기능을 실현하기 위한 노력이 계속되고 있다(문화체육관광부, 2022). 이에 현재 공공도서관 운영평가는 정량적, 정성적 평가 항목이 함께 사용되고 있으며, 정성적 평가 항목은 평가 대상 도서관이 증빙자료를 제출하고, 평가자가 직접 자료를 검토하여 점수를 부여하는 방식으로 수행되고 있다. 이러한 평가 과정은 평가자의 전문성을 필요로 하는 동시에 반복적이고 시간이 많이 소요되는 업무로 평가자의 부담을 가중시킬 우려가 있다. 또한, 평가 시기마다 상당한 인력과 비용이 소요되며, 평가의 신뢰성과 공정성을 확보하기 위해 여러 명의 평가자를 투입하고 여러 차례 검토하는 과정이 필요하다는 점에서 운영평가 체계의 효율성 제고는 중요한 과제이다.

한편, 대규모 언어모델(LLM, Large Language Model)을 기반으로 한 생성형 인공지능(Generative AI)의 등장은 그동안 인간이 수행해 온 반복적 업무를 대체하고 자동화할 수 있다는 점에서 그

잠재력이 크다(Liu et al., 2025). 특히, LLM은 방대한 양의 데이터를 학습하여 맥락을 파악하고 자연어 기반의 결과를 생산하는 능력을 토대로 전문가 수준의 판단을 요구하는 일부 업무를 지원하거나 대체하는 방식으로 그 활용 범위를 넓혀가고 있다. 실제 법률, 의학 분야 등에서는 해당 분야의 전문 데이터를 기반으로 새로운 이슈에 대응하거나 조언을 제공하는 형태로 전문가 수준의 과업을 수행하고 있으며 다양한 분야에서 인간이 수행해 왔던 과업의 일부를 자동화하려는 노력이 시도되고 있다(Li et al., 2023; Yue et al., 2023).

현재 사람이 수행하고 있는 공공도서관 운영평가 또한 대량의 텍스트 기반 증빙자료를 검토하고 평가 기준에 따라 객관적 판단을 내려야 한다는 점에서 LLM 기반 자동화에 적합한 영역 중 하나로 볼 수 있다. 자료의 내용과 형식을 일정한 평가 척도에 맞추어 검토하고, 일관된 기준으로 평가하는 것은 규칙성과 반복성이 높다는 점에서 생성형 AI 활용 시 효율성 향상이 기대되는 영역이다.

이에 본 연구는 공공도서관 운영평가에 생성형 AI 기반 자동평가를 적용해 보고, 그 도입 가능성을 검토하고자 한다. 이를 위해 현재 사람이 수행하고 있는 평가를 AI가 얼마나 유사하게 수행하는가, 그리고 AI 평가의 신뢰도와 정확도 향상을 위한 조건을 확인하고자 하였으며 이를 기반으로 향후 도서관 운영평가 자동화를 위한 시사점을 도출하였다.

연구 목적에 따른 연구 질문은 다음과 같다.

- RQ 1. 전문가와 AI 평가 결과는 얼마나 유사한가?
- RQ 2. 공공도서관 운영평가 시, AI 자동평

- 가 도입이 가능한 평가 영역은 무엇인가?
- RQ 3. 프롬프트에 따라 전문가와 AI 간 신뢰도(inter-rater reliability) 차이가 있는가?
 - RQ 4. 생성형 AI 기반 도서관 운영평가 자동화를 위해 고려해야 할 사항은 무엇인가?

본 연구는 현행 공공도서관 운영평가 지표 중 증빙자료를 근거로 사람이 평가하는 4개 영역에 대하여 문헌정보학 분야 전문가와 생성형 AI 평가 결과를 비교하였다. 2024년(23년 실적) 서울시 공공도서관 운영평가 자료를 기준으로 ChatGPT 5.1을 활용하여 평가를 진행했으며, 전문가 평가는 서울시 공공도서관 운영평가를 다년간 수행하여 증빙자료와 평가 기준에 대한 이해도가 높은 문헌정보학과 교수가 진행하였다. 또한, 선행연구를 기반으로 LLM 프롬프트 엔지니어링 기법을 확인하고, 이를 적용하여 프롬프트 입력에 따른 성능 차이를 검토하였다. 이를 토대로 생성형 AI가 인간 전문가의 평가 결과를 얼마나 유사하게 재현하는지 확인하고, 최종적으로 자동평가 도입을 위해 고려해야 할 시사점을 도출하였다.

2. 이론적 배경

2.1 공공도서관 운영평가

2008년 '전국 도서관 운영평가'가 시작된 이후 공공도서관 운영평가의 실효성을 높일 수 있는 평가지표를 제시하고자 하는 다양한 노력이 진행되어 왔다(차미경, 표순희, 2015). 특히,

2022년 운영평가에서 정성평가가 약 80% 비중을 차지하며 평가 체계에 큰 변화가 나타났으며, 이를 통해 정성평가에서 높은 점수를 받은 일부 도서관의 사례가 주목받게 되었다(문화체육관광부, 2022). 정성평가의 확대는 기존 정량평가만으로는 파악하기 어려운 도서관의 다양한 사례를 발굴하고, 기존 양적 지표 중심 평가에서는 좋은 점수를 받기 어려운 규모가 작은 도서관의 참여를 확대하는 데 기여했으며, 이는 운영평가가 도서관의 성장을 유도하는 장치로서 기능할 수 있음을 보여준 사례이다(감미아, 이지연, 2023).

정성평가 확대는 우수사례를 발굴하고 사서의 내적 동기를 유발함으로써 공공도서관의 성장을 유도하였다는 데 그 의의가 있으나 한편으로는 평가의 공정성과 객관성에 대한 문제가 제기되어 왔으며(윤명희, 김기영, 2009), 양적 평가보다 더 많은 인력과 시간이 소요된다는 점 또한 중요하게 고려해야 할 사항이다. 문화체육관광부(2025)의 '2026년(25년 실적) 공공도서관 운영평가 지침'에 따르면 총 7개 항목, 12개 지표, 지자체 상황에 따라 자율적으로 적용하는 자율지표로 구성되어 있다. 일부는 통계 연동에 따른 정량평가가 진행되나 일부는 각 도서관이 제출한 증빙자료를 토대로 사람에게 의한 정성평가가 진행된다. 이와 같이 공공도서관 운영평가에서 정량평가와 정성평가는 병행되어 사용되고 있으며 각 평가 방식의 장단점을 서로 보완할 수 있다는 점에서 둘 다 그 중요성이 크다. 이에 정성평가의 장점은 살려진 평가의 객관성과 효율성을 높이는 방안이 필요하며, 생성형 AI를 활용한 자동평가가 하나의 대안으로 제안될 수 있을 것이다. 실제 교

육 현장에서는 AI를 활용하여 서술형 문항 자동채점 프로그램을 제안하고 그 효과성을 검증한 사례가 다수 존재하며(신병철 외, 2024; 김민경, 2025), 이는 AI를 활용한 정성평가의 가능성을 시사한다.

2.2 전문 영역에서의 LLM 기반 자동화

LLM(Large Language Model, 이하 LLM)의 활용 범위가 일상적 수준을 넘어 의료, 법률, 금융 등 전문성을 요구하는 영역으로 확장되고 있으며, 도메인 특화 에이전트(domain-specific agents)가 발전함에 따라 각 분야에 특화된 에이전트 구성 요소와 특수 모델이 증가하고 있다. 이 과정에서 프롬프트 엔지니어링은 출력의 효과성을 높이고 다양한 도구 간 연결을 가능하게 하는 핵심 매커니즘으로 기능하며(Wang et al., 2024), 특정 영역에 특화된 프롬프트 엔지니어링이 주요 연구 주제로 다루어지고 있다. 특정 분야의 용어, 업무 절차 등을 고려한 최적화된 프롬프트를 사용하는 경우, 일반적인 프롬프트를 사용할 때보다 LLM이 더 정확하고 신뢰할 수 있는 출력을 생성하며(Liu et al., 2025), 이는 LLM이 범용적 언어모델의 특성을 유지하면서도 프롬프트 설계를 통해 전문 영역에 특화된 모델로 기능할 수 있음을 시사한다.

실제로 LLM은 특수분야의 데이터를 기반으로 새로운 이슈에 대응하거나 조언을 제공하는 형태로 활용되고 있으며, 소송절차, 금융 규제 준수, 진료 프로토콜 등 복잡한 워크플로우를 자동화하는 기능을 구현한다. 경영학 분야에서는 데이터 분석 리포트 생성을 위한 AI 프롬프트, 의학 분야에서는 영상의학 특화 프롬

프트를 설계하는 연구가 수행된 바 있다(박예은 외, 2024; 박지우 외, 2025).

이처럼 다양한 전문 영역에서 LLM 기반 업무 자동화가 확산되고 있으며, 이러한 기능을 구현하기 위해서는 각 영역의 맥락을 반영한 특화 프롬프트 엔지니어링이 필수적으로 선행되어야 한다. 전문 영역에 특화된 프롬프트를 설계하기 위해서는 해당 분야의 전문 지식과 고유한 특수성, 결과의 신뢰성, 법적·윤리적 기준 등을 충분히 고려하는 동시에 LLM이 가진 고유의 추론 능력과 맥락 이해 능력을 최대한 발휘할 수 있도록 해야 할 것이다. 이에 본 연구에서는 그동안 사람에 의해 수행되어 온 공공도서관 운영평가 과정을 자동화하는 방안을 모색하고자 하며, 이를 위해 도서관 분야 전문성과 평가의 특수성을 반영한 프롬프트를 설계하고 실제 적용을 통해 그 성능을 검증하고자 한다.

2.3 프롬프트 엔지니어링

LLM에서 프롬프트(prompt)는 단순 질의나 입력이 아닌 모델의 출력 양식과 내용, 추론 전략 등을 통제하는 핵심 요소로 작용한다. Liu et al.(2025)은 프롬프트 엔지니어링을 LLM의 답변에 직접적인 영향을 미치는 중요 기술임을 강조하며, 입력 조정을 통해 LLM의 행동을 체계적으로 유도하는 기술로 정의한다. 이와 같이 프롬프트 엔지니어링은 LLM의 성능을 결정하는 핵심 요인이며, LLM을 기반으로 한 생성형 AI 사용 시에도 원하는 답변을 얻기 위해서는 프롬프트 엔지니어링이 필수적이다(Ekin, 2023). 최근 연구에서는 동일한 모델 내

에서도 프롬프트에 따라 성능 차이가 크게 나타나는 현상이 반복적으로 보고되었으며, 프롬프트 엔지니어링은 단순히 모델의 성능을 높이기 위한 수단을 넘어 LLM 활용을 위한 설계 패러다임으로 이해될 수 있다(Chen et al., 2025).

프롬프트 엔지니어링은 경험적(empirical) 접근의 중요성이 강조되며, 에이전트 시스템의 기능적 구분에 따라 프롬프트 엔지니어링 기법을 분류한다(Liu et al., 2025; Schulhoff et al., 2024). 에이전트 시스템 관련 최근 연구에서는 에이전트의 기능을 Profile, Memory, Planning, Action 4가지 범주로 구분하며(Roy et al., 2024; Wang et al., 2024), 이에 따라 프롬프트 엔지니어링 기법을 프로필 및 지시(Profile and instruction), 지식(Knowledge), 추론 및 계획(Reasoning and Planning), 신뢰도(Reliability) 4개 영역으로 분류한다.

프로필 및 지시 영역에 해당하는 기법은 프롬프트의 기본 구조를 형성하는 전략이다. 성격정보(Personality information)는 LLM에 특정 역할과 나이, 성별, 직업, 심리적인 특성 등을 사전에 부여하고, 해당 역할에 부합하는 방식으로 응답을 생성하도록 유도하는 프롬프트 설계 방식이다. 과업정보(Task information)는 과업 목적과 세부 요구사항, 관련 분야 등에 관한 정보를 제시함으로써 특정 영역에 집중하도록 하여 보다 전문적이며 요구에 부합하는 결과를 도출하는 방식이다. 예시정보(Demonstration information)는 과업 설명에 적절히 설계된 입력-출력 예시(input-output mapping)를 추가하여 LLM이 이미 학습한 지식을 과업에 맞게 호출하고 과업 의도와 요구되는 출력 양식에 따라 일관된 응답을 생성하도록 한다.

지식 영역에 해당하는 기법은 프롬프트 강화 전략으로 기본 RAG 기법(Basic Retrieval-Augmented Generation techniques)은 외부 자원에서 정보를 검색하고 이를 활용하여 최종 프롬프트를 구성하는 방식이며, 고급 RAG 기법(Advanced RAG techniques)은 검색된 지식과 프롬프트 사이의 정합성을 향상시켜 효과적인 프롬프트를 구성하는 것으로 검색 및 대화 이력을 토대로 응답의 정확도를 제고한다.

추론 및 계획 영역에 해당하는 기법 또한 프롬프트 강화를 통해 LLM의 성능을 높이는 전략이다. 목표 분해 추론(Target decomposition reasoning)은 인간과 유사한 단계적 사고과정을 유도함으로써 LLM이 복잡한 과제를 여러 개의 하위 과제 또는 단계로 분해하고 각 단계에서 정교한 추론을 수행하도록 유도하는 프롬프트 설계 방식이다. 도구 및 피드백 강화형 추론(Tools and feedback enhanced reasoning)은 복잡한 계산이나 정보 탐색을 외부 도구에 위임하고, 자기평가 또는 외부 피드백을 반영하여 추론 과정과 출력을 점검·수정함으로써 더 정확하고 신뢰도 높은 응답을 생성하도록 한다.

신뢰도 영역에 해당하는 기법은 다양한 프롬프트 기법을 통합하여 LLM 시스템을 강화하는 전략이다. 내용편향(Content bias)은 LLM의 출력 내용이 과업 요구와 불일치하는 현상을 의미한다. 이를 완화하기 위한 기법으로는 프롬프트의 순서나 표현 등을 변형해 여러 응답을 생성한 뒤, 과업 요구에 가장 부합하는 출력을 선택함으로써 편향을 줄이는 방식이 있다. 가치편향(Value bias)은 LLM이 생성한 내용이 사회적 규범이나 윤리와 충돌하는 현상을

의미하며, 특정 기법이나 강화학습을 통해 유해하거나 부적절한 텍스트를 차단·수정함으로써 인간의 사회적 가치에 부합하는 출력을 생성하는 기법을 활용한다(Liu et al., 2025).

이와 같이 프로필 및 지시 영역에 해당하는 기법은 프롬프트 설계의 초기 단계로 기본 프롬프트를 구축하며, 지식과 추론 및 계획 영역은 프롬프트를 강화하여 LLM의 성능을 향상시키고, 신뢰도 영역의 기법은 최종 단계에서 모델 응답의 안정성을 보장하기 위해 적용된다. 이와 같은 분류체계는 대부분의 프롬프트 엔지니어링 기법을 포괄한다는 점에서 일반성과 확장 가능성이 크다. 따라서 본 연구에서는 해당 분류체계를 기준으로 공공도서관 운영평가 자동화에 적절한 프롬프트 엔지니어링 기법을 검토하였으며, 분류표는 <표 1>과 같다.

본 연구에서는 주어진 증빙자료와 척도만을 사용해서 평가해야 한다는 점을 고려하여 외부 데이터베이스를 활용하는 지식 영역에 해당하는 기법은 제외하였으며, 공공도서관 운영평가 자동화 가능성을 탐색하는 기초연구라는 점, 도서관 운영평가 맥락에서 생성형 AI 적용에 관한 연구가 초기 단계라는 점을 고려하여 프롬

프트의 기본 구조를 설계하는 과업정보(Task Information, 이하 TI) 기법을 사용하였다. 추가로 자동평가의 성능 향상 가능성을 확인하기 위해 추가 예시를 입력하는 예시정보(Demonstration Information, 이하 DI)와 단계적 추론을 유도하는 목표 분해 추론(Target Decomposition Reasoning, 이하 TDR) 기법을 적용하였으며, 신뢰도 영역에 해당하는 기법은 모델이 일정 수준 이상의 성능을 확보한 후에 적용 가능하다는 점에서 제외하였다.

본 연구에서 사용하는 TI는 프롬프트에 명확한 과업 목표와 요구사항을 포함해 생성형 AI의 출력을 표준화하는 프롬프트 기법이며, DI는 프롬프트에 입출력 예시를 추가하는 기법으로 소수의 예시를 제공하는 퓨샷 프롬프트(few-shot prompting)를 적용하였다(Logan et al., 2022). 마지막으로 TDR은 LLM의 추론 능력을 향상시키는 기술로 인간의 추론 과정을 반영하며 복잡한 과업을 여러 개의 하위 목표로 나누어 순서대로 해결하는 방식이다(Austin & Vancouver, 1996). Kojima et al.(2022)은 프롬프트에 “Let’s think step by step”이라는 문장을 입력하면 LLM의 순차적인 사고능력이

<표 1> 생성형 AI 프롬프트 엔지니어링 기법 분류(Liu et al., 2025)

기능적 구분	프롬프트 엔지니어링 기법
프로필 및 지시 (Profile and instruction)	<ul style="list-style-type: none"> • 성격정보(Personality information) • 과업정보(Task information) • 예시정보(Demonstration information)
지식 (Knowledge)	<ul style="list-style-type: none"> • 기본 RAG(Basic RAG techniques) • 고급 RAG(Advanced RAG techniques)
추론 및 계획 (Reasoning and Planning)	<ul style="list-style-type: none"> • 목표 분해 추론(Target decomposition reasoning) • 도구 및 피드백 강화형 추론(Tools and feedback enhanced reasoning)
신뢰도 (Reliability)	<ul style="list-style-type: none"> • 내용편향(Content bias) • 가치편향(Value bias)

구현된다는 것을 입증하였으며, 본 연구에서는 해당 문장을 프롬프트에 입력함으로써 TDR 프롬프트 기법을 구현하였다.

3. 연구 방법

3.1 평가데이터

본 연구는 '2024년 서울시 공공도서관 운영평가'를 위해 서울시 내 공공도서관 164개 관이 제출한 증빙자료를 활용하여 평가를 진행하였다. 2024년 공공도서관 운영평가 기준, 총 8개 영역의 13개 지표가 제시되었으며 이중 통계로 연동되는 항목을 제외하고 각 공공도서관이 증빙자료를 제출하고 사람이 해당 자료를 평가하는 4개 영역(공간, 협력, 경영계획, 서울시 자율지표: 우수사례)을 본 연구 범위로 설정하였다. 이에 본 연구에서는 서울시 내 공공도서관 164개 관이 공간, 협력, 경영계획, 우수사례 4개 영역에 대해 제출한 증빙자료를 사용하였으며 도서관명은 익명 처리하고 1차 데이터 클리닝을 거친 후 사용하였다. 데이터 형식은 한글, 엑셀, PDF, 이미지 등 다양했으며 한글파일은 ChatGPT 5.1에서 인식할 수 없어 PDF로 변환하여 사용하였다.

4개 평가 영역별 증빙자료 유형 또한 다양했다. 공간 영역에서는 계획서, 인증서, 시설 사진, 협력 영역에서는 MOU 체결 리스트 및 실질적 협력 내용 근거 자료, 경영계획 평가를 위한 자료로는 이용자 만족도 혹은 요구 조사 결과 및 이를 반영한 도서관 계획서 등 관련 근거 자료가 포함되었다. 우수사례 영역은 정해진 양식

에 따라 각 도서관의 실적을 작성하여 제출하였으며, 이는 일관된 항목과 내용으로 구성되었다. 실제 평가자료를 확인한 결과, 동일 평가 영역에 대한 증빙자료임에도 도서관별로 파일 형식 및 유형, 내용 등이 통일되지 않고 다양한 것으로 나타났다.

3.2 평가 척도

공간, 협력, 경영계획, 우수사례, 4개 평가 영역에 대한 평가 척도는 다음과 같다. 본 연구에서는 2023년 실적에 해당하는 증빙자료를 활용하였기에 2024년 공공도서관 운영평가에서 적용된 평가 척도를 사용하였다. 또한, 자율지표의 경우, 지자체 상황에 따라 자율적으로 적용하는 지표이므로 서울시 자율지표인 '우수사례'를 평가하였으며 그 기준은 2024년 서울시 공공도서관 운영평가 지침에 따랐다(〈표 2〉 참조).

3.3 평가 과정

본 연구는 도서관 운영평가 자동화를 위한 생성형 AI 도입 가능성을 확인하고자 하는 탐색적 연구로 자동평가가 가능한 평가 영역을 파악하고, 정확도와 일관성을 향상시킬 수 있는 프롬프트 입력 방식을 확인하고자 하였다. 이를 위해 각 도서관의 증빙자료를 근거로 사람이 평가하는 4개 영역(공간, 협력, 공간계획, 우수사례)에 대하여 ChatGPT 5.1을 활용하여 자동평가를 실시하였다. 또한, 서울시 공공도서관 운영평가를 다년간 수행한 문헌정보학과 교수가 전문가 평가를 수행하였다. 본 연구는 크게 2단계로 진행되었다.

〈표 2〉 영역별 평가 척도

평가 영역	평가 척도	배점
공간 (12)	공유 및 개방 공간	2
	디지털·미디어 공간	2
	복합문화 공간	2
	장애인 서비스 공간 및 시설	2
	유니버설(배리어프리) 디자인	2
	영유아 서비스 공간	2
협력 (10)	1~3개 도서관/유관기관	1
	4~5개 도서관/유관기관	2
	6~7개 도서관/유관기관	3
	8~9개 도서관/유관기관	4
	10개 이상 도서관/유관기관	5
	해당 없음	0
경영계획 (10)	이용자 만족도 혹은 요구 조사를 1년에 1회 이상 실시하고, 이를 문서화 했으며 도서관 계획 및 정책 수립에 반영하였다.	10
	이용자 만족도 혹은 요구 조사를 1년에 1회 이상 실시하고, 이를 문서화 했지만 도서관 계획과 정책 수립에 반영하지 않았다.	8
	이용자 만족도 혹은 요구 조사를 2년 이내에 1회 이상 실시하고, 이를 문서화 했으며 도서관 계획 및 정책 수립에 반영하였다.	6
	이용자 만족도 혹은 요구 조사를 2년 이내에 1회 이상 실시하고, 이를 문서화 했지만 도서관 계획과 정책 수립에 반영하지 않았다.	4
	이용자 만족도 혹은 요구 조사를 3년 이내에 1회 이상 실시하고, 이를 문서화하였다.	2
	이용자 만족도 조사를 실시한 적 없다.	0
우수사례 (20)	혁신성·우수성	5
	상호작용	5
	지속·확산 가능성	5
	논리적 체계성	5

첫 번째 단계에서는 선행연구를 통해 확인한 프롬프트 엔지니어링 기법 중 과업정보(Task Information, 이하 TI)를 적용하여 평가 척도 및 기준 등 평가 관련 기본 정보를 구조화한 프롬프트를 ChatGPT 5.1에 입력하고, 각 도서관의 증빙자료를 업로드하여 평가를 진행하였다. 이때, 평가 로그를 주기적으로 점검하여 평가의 일관성을 유지하고자 했다. 이후 전문가가 동일 증빙자료로 진행한 평가점수와 AI 평가 점수를 4개 영역별로 비교하고 AI 자동평가 도입 가능성을 확인하였다.

두 번째 단계에서는 프롬프트 기법에 따른 차이를 확인하였다. 이를 위해 앞서 사용한 기본 프롬프트인 TI를 중심으로 예시정보(Demonstration Information, 이하 DI), 목표 분해 추론(Target Decomposition Reasoning, 이하 TDR) 기법을 조합하여 기본 프롬프트 외 추가 정보를 입력하는 경우 AI의 평가 기능이 향상되는지 확인하였다. 4개 평가 영역에서 아래의 3가지 조합으로 프롬프트를 입력하고 각각 전문가와 AI 간 신뢰도를 검증하였으며, 그 결과를 토대로 프롬프트 기법 적용에 따른 평가 성능 향상 여

부를 확인했다. 이와 같은 결과를 종합하여 공공도서관 운영평가 자동화를 위한 생성형 AI 도입 가능성과 방안을 모색하였다.

- ① Task Information(TI)
- ② Task Information(TI) + Demonstration Information(DI)
- ③ Task Information(TI) + Target Decomposition Reasoning(TDR)

3.4 프롬프트 입력

AI 자동평가를 위한 프롬프트는 다음과 같이 구성하였다. 프롬프트 엔지니어링 기법 중 프롬프트의 가장 기본 구조를 형성하는 TI를 적용하였으며, 이는 과업의 목표를 규정하는 의도(intent), 세부 요구사항을 명시하는 요구(demand), 과업 관련 정보를 얻을 수 있는 정보원인 도메인(domain)을 프롬프트의 핵심 요소로 포함한다(Liu et al., 2025). 하지만 본 연구에서는 주어진 증빙자료와 척도만을 기준으로 평가해야 하며 외부 자원을 연계할 수 없

는 점을 고려하여 의도와 요구 2개 요소를 반영하였다. 4개 평가 영역 모두 동일한 내용 요소를 포함하도록 구성하였으며, 평가지표에 대한 설명(목적, 내용 및 범위), 평가 척도, 평가 시 검토 사항(증빙자료 인정 및 불인정 기준 등), 평가 결과 제시 방법 4가지 요소로 구성하였다. 이후 기본 TI 프롬프트에 예시자료를 추가로 업로드하거나(DI), LLM의 단계별 추론을 유도하는 문장인 "Let's think step by step"을 입력하는 방식(TDR)으로 프롬프트를 발전시켰다. 다음은 우수사례 평가를 위해 입력한 프롬프트 예시이다(〈표 3〉 참조).

3.5 신뢰도 검증 방법

전문가와 AI 평가 결과 간 상관관계와 일치도를 파악하기 위해 피어슨 상관계수(Pearson Correlation Coefficient, 이하 PCC)와 급내상관계수(Intraclass Correlation Coefficient, 이하 ICC)를 확인하였으며, 이를 통해 인간 전문가와 AI가 동일 자료에 대해 얼마나 유사한 점

〈표 3〉 우수사례 영역 평가를 위한 프롬프트

<p>지금부터 공공도서관 운영평가를 진행합니다. 평가지표는 '도서관 운영 우수사례'이며, 해당 지표에 대한 설명, 평가척도, 평가 시 검토사항, 평가 결과 제시 방법은 다음과 같습니다.</p> <ul style="list-style-type: none"> ① 평가목적: 단위도서관 운영 우수사례를 발굴하여 지역 확산 도모. 양적 지표만으로 평가하기 어려운 도서관 운영 현장의 특성을 평가에 반영 ② 지표의 내용 및 범위: 2023년에 단위도서관에서 지역별 특성을 반영하여 도서관을 운영한 사례. 2023년 이전에 계획을 수립하고, 실제 실행이 2023년에 된 경우도 인정 ③ 평가 척도: 이미지 별도 업로드 ④ 평가 시 검토 사항: 필요시 추진 내용 및 결과를 증빙할 수 있는 보조자료 첨부 ⑤ 평가 결과 제시 방법: 4개 하위 영역별 점수(혁신성·우수성/상호작용/지속·확산 가능성/논리적 체계성)와 총점 제시 <p>지금부터 각 도서관이 제출한 증빙자료를 토대로 점수를 계산해 주세요.</p>
--

수를 부여하는지 검증하였다. 공간, 협력, 경영 계획, 우수사례 4개 영역 내 측정값 간 선형 관계를 통해 전문가 판단에 의한 점수와 AI 자동 평가로 부여되는 점수 간 유사도, 즉 패턴의 일관성을 파악하고자 하였다. 추가로, 두 점수 간 일치도를 확인하기 위하여 평가자 간 신뢰도 분석을 위해 사용되는 지표인 급내상관계수(ICC)를 확인하였다(Koo & Li, 2016). 본 연구의 목적, 평가 조건 등을 고려하여 이차원 변량(two-way random) 모형을 사용하였으며, 평가자 간 얼마나 차이가 있는지 확인하고자 단일 측도(single)를, 전문가와 AI가 동일한 자료를 평가했을 때 같은 점수를 부여하는지 일치도를 확인하기 위해 절대 합치도(Absolute agreement)를 확인하였다(McGraw & Wong, 1996; Shrout & Fleiss, 1979). 통계 패키지 SPSS 28을 사용하여 95% 신뢰도 구간 내에서 ICC(2, 1)를 계산하여 제시하였다.

4. 분석결과

4.1 전문가와 AI 평가 결과

4.1.1 기술통계

4개 평가 영역에 대하여 서울시 164개 공공도서관이 제출한 증빙자료를 기반으로 전문가 평가와 AI 자동평가를 진행하였다. 동일한 평가 척도와 평가 시 고려사항을 적용하였으며, AI 평가의 경우, 프롬프트 구조를 형성하는 TI를 적용하여 평가를 위해 필요한 기본 정보를 입력하였다. 이후 4개 영역별로 각 도서관이 제출한 자료를 업로드하고 평가를 진행하였다. 4

개 영역(공간, 협력, 경영계획, 우수사례)에 대한 전문가 평가와 AI 평가 결과는 <표 4>와 같으며, 영역별 표본 수(N), 최솟값(Min), 최댓값(Max), 평균(Mean), 표준편차(SD)를 제시하였다.

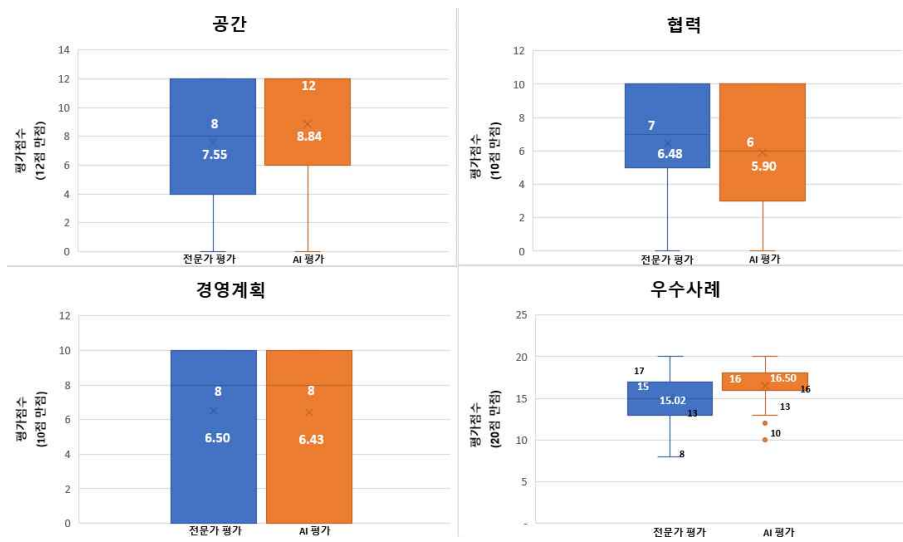
공간 영역의 경우, 평가 척도에 따라 12점 만점으로 평가하였으며 164개 관의 평가 결과는 다음과 같다. 전문가 평가 결과는 평균 7.55점(SD = 4.31), AI 평가는 평균 8.84점(SD = 4.27)으로 AI 점수가 전문가 점수보다 높은 것으로 나타났으며, 두 평가 간 분포는 유사한 수준으로 확인되었다. 협력 영역은 전문가 평가 결과 평균 6.48점(SD = 3.37), AI는 평균 5.9점(SD = 3.58)으로 전문가 평가점수가 더 높게 나타났다. 경영계획 영역은 파일 형식 문제로 ChatGPT 5.1에 업로드 불가능한 경우 15건을 제외하고, 149개 관에 대한 평가를 진행했다. 그 결과, 전문가는 평균 6.5점(SD = 4.09), AI는 평균 6.43점(SD = 4.20)으로 두 평가 간 평균이 거의 동일하고, 분포도 유사한 것으로 나타났다. AI 자동평가와 전문가 평가가 유사한 수준의 일관성을 보였다. 정성평가 영역인 우수사례의 경우, 총 20점 만점으로 평가하였으며 평가자료 미제출 도서관을 제외하고 131개 관에 대한 평가를 진행했다. 해당 영역의 전문가 평가 결과는 평균 15.02점(SD = 2.57), AI는 평균 16.5점(SD = 1.90)으로 AI가 전문가보다 높은 평균을 보였으며 표준편차는 더 낮은 것으로 나타났다.

4.1.2 평가 영역별 점수 분포

<그림 1>은 박스플롯을 활용하여 4개 평가 영역의 전문가와 AI 평가점수 분포를 제시하

〈표 4〉 4개 영역별 평가점수에 대한 기술통계

평가 영역	전문가 평가					AI 평가				
	N	Min	Max	Mean	SD	N	Min	Max	Mean	SD
공간(12)	164	0	12	7.55	4.31	164	0	12	8.84	4.27
협력(10)	164	0	10	6.48	3.37	164	0	10	5.90	3.58
경영계획(10)	149	0	10	6.50	4.09	149	0	10	6.43	4.20
우수사례(20)	131	8	20	15.02	2.57	131	10	20	16.50	1.90



〈그림 1〉 전문가 - AI 평가 간 점수 분포

고 있다. 평가 영역별 박스플롯은 중앙값, 사분위 범위, 전체 범위 및 이상치를 보여주며 두 평가 주체의 점수 분포를 시각적으로 보여준다.

공간 영역에 해당하는 그림을 살펴보면, 평균값이 각각 7.55점, 8.84점(12점 만점)으로 AI 자동평가 점수가 평균적으로 약 1.3점 높아 타 영역보다 점수 차이가 큰 것으로 나타났다. 두 평가 주체 모두 0에서 12점까지 척도 전체를 활용하고 있지만, 전문가 점수의 중앙값은 8점이고 주로 4~12점 사이에 위치해 비교적 점수가 고르게 분포하는 것을 확인할 수 있다. 반면, AI 평가에서는 중앙값이 12점(만점)으로 나타

났으며 이는 점수 분포가 위쪽으로 강하게 몰린 형태로 AI가 높은 점수를 집중적으로 부여함을 알 수 있다. 즉, AI 점수가 11~12점 사이 상위 점수대에 집중되어 있다는 점은 AI가 동일한 수행을 더 높게 평가하는 경향이 있으며, 이는 AI가 전문가보다 중간 이하 수준의 사례를 정확하게 구분하지 못하고, 상대적으로 높은 점수를 부여하는 것으로 해석할 수 있다. 정리하면, 공간 영역에서는 AI가 평균적으로 더 높은 점수와 1사분위 수를 보여 전문가보다 높은 점수를 부여하는 경향이 있음을 보여준다.

협력 영역에서는 AI 평가 결과의 중앙값과

평균이 낮게 나타나 전반적으로 AI가 전문가보다 낮은 점수를 부여하는 경향이 확인되었다. 전문가 점수는 5~10점, AI는 3~10점 구간에 있어 AI 평가점수는 중하위 점수대(3~6점)에 많이 분포하고, 전문가 점수는 5점 이상에 집중된 것으로 나타났다. 점수 분포를 살펴보면 두 평가 모두 0~10점까지 평가 척도 전 영역을 사용하고 있으나, AI 평가점수를 나타내는 박스가 아래쪽에 더 치우친 것으로 확인된다. 종합하면, 협력 영역에서는 AI가 전문가보다 낮은 점수를 부여하며, 두 평가 주체 모두 우수한 사례에 대해서는 높은 점수를 부여하나 협력 수준이 중간 이하인 경우는 AI가 더 보수적으로 평가하여 두 평가 주체 간 편차가 비교적 크게 나타나는 것으로 확인되었다.

경영계획 영역에서는 전문가와 AI 평가점수 수준과 분포가 거의 유사하게 나타났다. 두 주체의 중앙값은 8점으로 동일했고, 평균 또한 전문가 6.5점, AI 6.43점으로 거의 같게 나타나 해당 영역에서는 두 주체의 평가 결과가 매우 유사함을 알 수 있다. 점수 범위, 중앙값, 평균, 변동성 등에서 전문가와 AI 평가 결과 간 차이가 거의 없고, 편향성 또한 나타나지 않아 4개 영역 중 전문가 평가와 AI 자동평가의 평가 양상이 가장 일치하는 영역으로 볼 수 있다. 즉, AI가 제시된 평가 척도와 인간 전문가의 판단 기준을 비교적 안정적으로 재현하는 것으로 해석할 수 있다.

마지막으로 우수사례는 전문가 평가와 AI 자동평가 점수 분포가 가장 뚜렷하게 차이 나는 영역이다. 전문가 평가점수는 중앙값과 평균값 모두 약 15점(20점 만점), AI 평가점수는 중앙값 17점, 평균 16.5점으로 전문가 점수보다

약 1.5~2점 높게 나타났다. 전문가 점수의 경우, 주로 13~17점 사이에 분포하며, 최솟값 8점부터 최댓값 20점까지 점수 폭이 큰 것으로 확인되었다. 반면, AI 평가점수는 16~18점 사이에 주로 분포하여 상자가 좁은 구간에 집중되어 있고 10점, 12점 등 하위 이상치가 나타났다. 이는 AI가 대부분 자료에 대해 다소 높은 점수를 부여하면서도 소수 사례에 대해서는 예외적으로 낮은 점수를 부여함을 보여준다. 이처럼 우수사례 영역에서는 AI가 전문가보다 높은 점수를 부여하는 경향이 뚜렷하게 나타나지만 점수 분포에 있어서는 차이를 보인다. 전문가 평가점수는 비교적 넓게 분포하는 반면, AI 평가점수는 16~18점에 밀집해 있어 변동성이 작고 우수사례 간 차이를 구분하는 변별력이 다소 낮은 것으로 해석된다. 또한, AI 평가에서 몇 개의 이상치가 나타났는데 이는 대부분 사례에 점수를 높게 부여하는 패턴이 뚜렷하면서도 특정 사례에 대해서는 매우 낮은 점수를 부여하여 평가의 일관성과 신뢰도에 대한 검증이 필요함을 시사한다.

4개 평가 영역별 기술통계와 점수 분포를 살펴본 결과, 경영계획 평가에서는 전문가와 AI 평가의 점수 분포, 평균, 중앙값 등에서 차이가 크지 않아 평가 경향이 매우 유사한 것으로 나타났다. 반면, 우수사례 평가에서는 전문가와 AI 간 차이가 뚜렷하게 났으며 평균과 중앙값, 점수 분포 모두 두 점수 간 차이가 가장 크게 나타났다. 특히 공간과 우수사례 영역에서 AI 평가점수가 특정 구간에 밀집하는 것으로 나타났다. 이는 AI가 사람이 평가하는 것보다 점수를 더 높게 또는 더 낮게 부여하는 경향이 있고, 평가의 변별력이 상대적으로 낮음을 의미

한다. 이와 같이 4개 영역별로 전문가 평가와 AI 자동평가를 진행하고 점수를 비교한 결과 영역별 점수 패턴에 차이가 있음을 확인하였으며, 이는 평가 내용과 척도에 따라 AI 자동평가의 정확도와 성능이 다를 수 있음을 시사한다.

4.2 전문가-AI 간 신뢰도

4.2.1 평가 영역별 신뢰도

4개 평가 영역별 전문가 평가와 AI 평가 간 유사도와 일치도를 확인하기 위하여 피어슨 상관관계수(PCC)와 급내상관계수(ICC)를 확인하였다. 피어슨 상관관계수의 크기는 0.10 미만이면 무시 가능한 효과, 0.10 이상 0.30 미만은 작은 효과, 0.30 이상 0.50 미만은 중간 효과, 0.50 이상은 큰 효과로 해석하였다(Cohen, 2013). ICC의 경우, 추정값 95% 신뢰구간을 기준으로 0.50 미만은 낮은 신뢰도(poor reliability), 0.50 이상 0.75 미만의 값은 보통 수준의 신뢰도(moderate reliability), 0.75 이상 0.90 미만의 값은 좋은 신뢰도(good reliability), 0.90을 초과하는 값은 매우 높은 신뢰도(excellent reliability)를 나타낸다(Koo & Li, 2016).

동일한 평가 척도를 적용하여 전문가가 평가한 경우와 TI 프롬프트 기법을 적용하여 AI 평가를 진행한 경우, 두 평가 결과 간 유사도와 일치도를 확인한 결과는 다음과 같다. 4개 영역 중 공간, 협력, 경영계획 3개 영역의 피어슨 상관관계수는 0.7 이상, ICC는 0.8 이상으로 전문가 점수와 AI 점수 간 유사도와 일치도가 매우 높은 것으로 나타났다. 반면 우수사례는 PCC = 0.182, ICC = 0.252로 낮은 수준의 일치도를 보였다.

4개 영역 중 공간 영역은 PCC = 0.868, ICC = 0.908, 95% CI [.791, .950]으로 가장 높게 나타났고, 상한값뿐만 아니라 하한값도 기준치인 0.75 이상으로 전문가와 AI 간 신뢰도가 매우 높은 수준으로 나타난다. 협력 영역에서는 ICC = 0.835, 95% CI [.772, .880], 경영계획은 ICC = 0.854, 95% CI [.799, .895]로 두 영역 모두 0.75 이상의 높은 신뢰도를 보였으며 피어슨 상관관계수 또한 각각 0.726, 0.745로 전문가와 AI 자동평가 점수 간 일치도가 높은 것으로 확인되었다. 하지만 정성평가 영역에 해당하는 우수사례는 PCC = 0.182, ICC = 0.252, 95% CI [-.024, .458]로 기준치인 0.50 미만으로 낮은 신뢰도를 보였다. 정성평가의 경우, 전문가와 AI 평가 결과 간 신뢰도가 낮으며 두 평가점수 간 유사한 경향이 발견되지 않았다. 이는 전문가와 AI가 평가 시 적용하는 기준이나 평가자료를 해석하는 방식에 차이가 있을 가능성을 시사한다.

이와 같은 내용을 고려하였을 때, 공간, 협력, 경영계획 영역에서는 전문가와 AI 평가점수 간 높은 수준의 일치도가 확보되어 AI 평가가 전문가에 의한 평가를 상당 부분 대체하거나 1차 스크리닝 도구로 활용될 수 있음을 시사한다. 반면, 정성평가가 이루어지는 우수사례의 경우 신뢰도가 낮아 여전히 전문가의 정성적 판단이 필요한 영역으로 확인되었다. 이와 같은 차이는 평가 척도의 구체성에 따라 그 결과가 달라질 수 있음을 보여준다. 신뢰도가 높게 나타난 3개 영역의 경우, 평가자료를 토대로 유형별 공간 수, 협력 기관 및 건수, 만족도 조사 시행 여부 및 계획서 내 반영 여부와 같이 특정 내용의 존재 여부와 수치를 기준으로 비교적 명확한 평가 척도 제시가 가능한 반면, 우수사례의 경

우 활동의 독창성 및 참신성과 같이 평가자의 주관적 판단이 필요한 영역이기 때문에 상대적으로 명확하고 구체적인 평가 척도 제시가 어렵다는 문제점이 있다. 따라서 정성평가 영역에서 AI를 활용한 자동평가의 정확도를 높이기 위해서는 더욱 구체적인 척도 제시, 예시 제공 등 추가 조정이 필요하며, 현 수준에서는 AI 평가가 전문가의 평가를 완전히 대체하기보다는 제한적인 보조 도구로 활용하는 것이 적절함을 시사한다(〈표 5〉 참조).

4.2.2 하위 척도별 신뢰도

4개 평가 영역 중 우수사례의 경우, 정성평가가 이루어지는 영역으로 특정 공간 설치 여부, 협력 건수, 만족도 조사 시행 여부 및 계획 여부 등을 평가하는 타 영역보다 평가의 전문성과 일관성이 더욱 강조된다. 위의 평가 영역별 전문가와 AI 간 신뢰도를 확인한 결과, 정성평가에 해당하는 우수사례는 다른 3개 영역과 비교하였을 때 두 점수 간 유사도와 일관성 모두 낮게 나타났다. 이는 AI를 활용한 평가 자동화를 위해서는 정성평가 영역의 신뢰도를 높이는 것이 중요 사항임을 시사하며 이를 위해서는 명확한 평가 척도 설정이 특히 중요할 것이다. 본 연구에서는 우수사례 영역 평가 시 총점뿐만 아니라 각 척도 수준에서도 사람과 AI의

평가가 유사하게 나타나는지, 어떤 척도에서 상대적으로 신뢰도가 높게 또는 낮게 나타나는지 비교하였으며 추후 도서관 운영평가 자동화를 위한 시사점을 도출하고자 하였다.

우수사례 영역은 혁신성·우수성, 상호작용, 지속·확산 가능성, 논리적 체계성 4개의 하위 영역으로 구성되며 영역별 전문가와 AI 간 신뢰도 결과는 다음과 같다. 우수사례 하위 영역의 전문가와 AI 평가점수 간 상관은 혁신성·우수성(PCC = 0.22), 상호작용(PCC = 0.18), 지속·확산 가능성(PCC = 0.09), 논리적 체계성(PCC = 0.07) 모두 낮은 수준에 머물렀고 ICC 또한 0.105~0.279 사이로 나타나 기준치인 0.50에 미치지 못하는 것으로 나타났다. 우수사례의 하위 평가 영역에서 전문가 평가점수와 AI 평가에 의한 점수 간 상관관계와 일치도가 모두 낮으며 영역별 차이가 크지 않은 것으로 나타났다. 이는 자동평가 도입을 위해서는 하위 척도의 조정 및 재설정이 필요함을 시사한다(〈표 6〉 참조).

4.2.3 프롬프트 기법별 신뢰도

프롬프트 기법에 따라 전문가와 AI의 평가 점수 간 차이가 있는지 확인하기 위해 TI를 적용한 경우, TI에 추가로 예시를 제시하는 경우(DI), 생성형 AI의 단계적 추론을 유도하는

〈표 5〉 전문가 - AI 간 신뢰도(Inter-rater Reliability)

평가 영역	ICC(2,1)	95%CI		F Test With True Value 0			
		하한값	상한값	값	df1	df2	유의확률
공간	.908	.791	.950	14.177	163	163	.000
협력	.835	.772	.880	6.277	163	163	.000
경영계획	.854	.799	.895	6.825	148	148	.000
우수사례	.252	-.024	.458	1.422	130	130	.023

〈표 6〉 우수사례 영역 전문가 - AI 간 신뢰도(Inter-rater Reliability)

하위 평가 영역	ICC(2,1)	95%CI		F Test With True Value 0			
		하한값	상한값	값	df1	df2	유의확률
혁신성·우수성	.243	-.074	.469	1.540	130	130	.007
상호작용	.279	.005	.481	1.433	130	130	.021
지속·확산 가능성	.166	-.179	.409	1.198	130	130	.152
논리적 체계성	.105	-.217	.348	1.131	130	130	.242

“Let’s think step by step” 프롬프트를 추가하는 경우(TDR), 3개 상황에서 신뢰도를 확인하고 이를 비교 분석하였다. TI 프롬프트 기법을 적용하여 영역별 평가를 위한 기본 사항인 평가 척도와 평가 시 고려사항 등을 프롬프트로 입력하였으며, 이에 추가로 전문가 평가에서 높은 점수를 받은 ‘평가자료-평가점수’ 쌍으로 구성된 예시 또는 LLM의 논리적 추론을 유도하는 프롬프트를 입력하는 방식으로 진행하였다. 이를 통해 기본 프롬프트를 제시하는 경우보다 추가 예시를 제시하거나 추론을 유도했을 때 신뢰도가 향상되는지 확인하고자 하였으며, 예시를 제시하는 경우와 논리적 추론을 유도하여 단계별로 과제를 수행하도록 하는 경우 중 어떤 프롬프트가 신뢰도 상승에 더 많이 기여하는지 확인하였다.

〈표 8〉은 TI 프롬프트만 제시했을 때, TI에 예시를 제공하는 DI 기법을 추가 적용했을 때, TI에 순차적 사고를 유도하는 TDR을 추가했을 때, 3가지 조건에서 전문가와 AI 간 신뢰도가 어떻게 달라지는지 보여준다. TI에 DI를 추가한 결과는 다음과 같다. 공간 영역에서는 TI 프롬프트만으로도 ICC = 0.908로 매우 높은 수준의 신뢰도를 보였기에 프롬프트 추가에 따른 신뢰도 상승이 제한적이거나 DI 프롬프트 추가 이후 ICC가 소폭 상승한 것으로 확인되었

다. 협력과 경영계획 영역 또한 TI만으로도 높은 신뢰도를 보였으나 예시를 추가한 이후 ICC가 소폭 상승하고 신뢰구간 하한도 올라가는 것으로 나타났다. 이와 같이 공간과 협력, 경영계획 영역에서는 TI 프롬프트만으로도 전문가와 AI 간 신뢰도가 높게 나타나 예시 추가 입력에 따른 신뢰도 상승 차이가 크지는 않았으나, 예시를 추가로 제공했을 때 신뢰도 개선 효과가 있는 것으로 확인되었다.

우수사례 영역에서는 DI 프롬프트 추가에 따른 신뢰도 상승 폭이 매우 큰 것으로 확인되었다. TI 프롬프트만 입력했을 때는 ICC = 0.252, 신뢰구간이 0 이하까지 내려가 전문가와 AI 간 일치도가 거의 없는 것으로 나타났으나 예시를 추가한 DI 프롬프트 입력 이후 ICC = 0.485로 2배 가까이 상승하여 보통 수준에 가까운 신뢰도를 보였으며, 신뢰구간 또한 전체적으로 위로 이동하는 것을 확인할 수 있다. 이는 상대적으로 평가 과정이 복잡하고 평가자료에 대한 질적 판단이 필요한 정성평가 영역에서 예시를 제시하면 자동평가의 신뢰도를 높일 수 있으며 전문가 판단과 유사한 평가를 유도할 수 있다는 점을 시사한다. 정성평가에 해당하는 우수사례의 경우, 추가로 예시 입력 방식을 검증하였다. 전문가 평가에서 20점 만점 또는 19점을 받은 우수도서관의 자료 총 13건의 예시를 사

용했으며 모든 사례를 한꺼번에 입력하지 않고, 총 4번에 나누어 순차적으로 제시하면서 예시가 누적될수록 평가의 정확도가 향상되는지 확인하였다. ICC를 확인한 결과, 0.535, 0.440, 0.559, 0.423 순으로 신뢰도 값이 나타났으며, 예시 입력 사례 수 증가와 신뢰도는 크게 관련이 없는 것으로 나타났다. 신뢰도가 약간 상승하는 경우가 있으나 그 경향이 일정하지 않으며 자동평가 기능 향상은 크지 않은 것으로 보인다. 즉, 도서관 운영평가라는 맥락 내에서 생성형 AI를 활용하기 위해서는 다수의 예시를 제공하는 것도 중요하나 양질의 예시를 선별하여 입력하는 것이 자동평가의 정확도를 높이는 방법임을 시사한다.

다음으로 TI에 생성형 AI의 순차적 추론을 유도하는 TDR 프롬프트를 추가하였다. 앞선 TI 프롬프트에 단계적 추론을 유도하는 “Let’s think step by step” 프롬프트를 입력한 결과, 평가척도 정리, 심사영역별 세부 기준(루브릭 설정), 평가 절차, 평가 결과 제시 방법 순으로 평가 과정을 제시하였으며, 이후 순서에 따라 평가를 진행하였다. <표 7>은 우수사례 영역 평가를 위한 TDR 적용 시 AI의 답변이며 TDR 프롬프트 추가에 따른 결과는 다음과 같다. 공간 영역에서는 이미 1차 프롬프트 입력만으로도 매우 높은 수준의 신뢰도가 확보되었기에(ICC = 0.908), 추가 프롬프트 입력으로 결과가 크게 달라지지 않았으나 TDR 프롬프트를 추가했을 때 신뢰도가 소폭 하락(ICC = 0.888)하는 것으로 나타났다. 하지만 두 조건 모두 0.8 이상의 높은 신뢰도를 보인다는 점에서 프롬프트 입력에 따른 차이는 크지 않은 것으로 보인다. 협력 또한 공간과 유사하게 TDR 프롬프트

추가 입력 후 신뢰도가 감소하였고, 감소 폭은 더 큰 것으로 확인되었다. ICC가 0.835에서 0.775로 떨어지고, 신뢰구간 하한이 0.77에서 0.65로 낮아져 여전히 높은 수준의 신뢰도 범주에 속하긴 하지만 TDR 프롬프트 추가 시 전문가와 AI 간 일치도가 오히려 감소하는 것을 알 수 있다. 경영계획 영역에서도 추가 프롬프트 입력 후 신뢰도가 소폭 감소하는 경향이 나타났으나 그 폭이 크지 않았으며 추가 입력 후에도 ICC = 0.830으로 높은 수준의 신뢰도를 유지하였다. 우수사례 영역에서는 TDR 프롬프트 추가에 따른 효과가 거의 없었으며, 여전히 낮은 수준의 신뢰도를 보였다. 이는 예시를 추가하는 DI 프롬프트를 입력하는 경우와 반대되는 결과로 예시 추가 입력 시에는 신뢰도가 매우 크게 상승했으나 논리적 추론을 유도하는 방식은 자동평가의 성능 향상에 크게 기여하지 않음을 시사한다. 특히 복합적인 요소를 종합적으로 판단해야 하는 우수사례의 경우, 추가적인 목표 분해, 추론 요구 등이 오히려 전문가와 AI의 판단 기준을 어긋나게 만드는 것으로 해석할 수 있다.

프롬프트 추가 입력에 따른 AI 자동평가 성능 향상 여부를 정리하면 다음과 같다. 기본 TI 프롬프트에 DI 프롬프트를 추가로 입력한 경우, 정도의 차이는 있으나 4개 영역 모두 신뢰도가 상승하는 것으로 나타났으며 특히 정성평가 영역에서 신뢰도가 크게 상승하는 것으로 나타났다. 이는 평가 난이도가 높은 정성평가 영역에서 단순히 평가 척도와 기준만 제시하는 것보다 구체적인 예시를 함께 제공하는 프롬프트가 적절함을 의미한다. 이를 통해 전문가 평가와의 유사도 및 일치도를 높일 수 있으며 AI 자동평가의 성능 향상이 가능할 것이다. 반면,

<표 7> TDR 프롬프트 입력 시(우수사례 영역) AI 답변

1. 기본 평가척도 정리

모든 문항은 다음 5점 척도로 봅니다.

1점: 매우 아니다 / 2점: 약간 아니다 / 3점: 보통이다 / 4점: 약간 그렇다 / 5점: 매우 그렇다

각 심사영역(혁신성·우수성 / 상호작용 / 지속·확산 가능성 / 논리적 체계성)마다 3개 세부 문항에 대해 각각 1~5점 부여 → 평균 내서 영역 점수 산출(반올림) 하겠습니다.

2. 심사영역별 세부 기준(루브릭) 설정

① 혁신성·우수성

[문항]

- 제시한 활동이 독창적이고 새로운가?
- 기존에 타 도서관에서 시행되었으나, 제출 도서관에서 참신한 아이디어를 보강하였는가?
- 타 도서관에서 적극 도입하도록 추천할 만한 내용과 성과를 갖춘 사례인가?

[평가 시 확인할 핵심 요소]

- 지역·이용자 특성을 반영한 '새로운' 기획인지
- 기존 사업을 단순 모방이 아니라, 구조·내용·운영 방식에서 변형/심화했는지
- 정량/정성 성과(참여자 수, 만족도, 언론·SNS 반응, 타 기관 문의 등)가 있는지

[점수 해석 예시]

- 5점: 지역·국내에서 보기 힘든 수준의 새로운 모델이며, 성과가 뚜렷하고 다른 도서관이 벤치마킹하고 싶은 수준
 - 3점: 어느 정도 차별성은 있으나, 기존 프로그램의 변형 수준, 성과도 보통
 - 1점: 유사 사례가 흔하고, 차별성과 성과 근거가 거의 없음
- ※ 상호작용, 지속·확산가능성, 논리적 체계성 영역도 동일한 형식으로 루브릭 제시

3. 실제 평가 절차

각 도서관 자료가 올라오면, 도서관별로 다음 순서로 진행하겠습니다.

[Step 1. 사례 개요 파악]

- 사업명, 대상, 기간, 추진 배경·목적, 주요 내용, 성과 부분을 먼저 읽습니다.

[Step 2. 심사영역별 증빙 정리]

- 혁신성·우수성 관련 근거(새로움, 차별성, 성과)
- 상호작용 관련 근거(수요조사, 내부 논의, 외부협력)
- 지속·확산 가능성 근거(정례화·재원·매뉴얼·확산 구조)
- 논리적 체계성 근거(추진배경-추진 내용-실적 구조, 성과 증빙, 구체성)

[Step 3. 각 영역의 3개 문항에 대해 1~5점 부여]

- 위 루브릭에 따라 문항별 점수와 짧은 평가 코멘트 작성

[Step 4. 영역 점수 산출]

- (세 문항 점수의 산술평균) → 반올림하여 1~5점 결정
- 필요하면 "평균은 4.3점이지만, 전반적 수준을 고려해 4점으로 판단"처럼 간단한 보정 근거도 함께 제시

[Step 5. 최종 결과 정리 형식]

- 도서관별·영역별 결과는 아래와 같은 형식으로 드리겠습니다.

[도서관 명] - 혁신성·우수성

문항1: 활동의 독창성/새로움 - ○점 (근거 요약)

문항2: 기존 사업의 보강·차별화 - ○점 (근거 요약)

문항3: 타 도서관 추천 가능성 - ○점 (근거 요약)

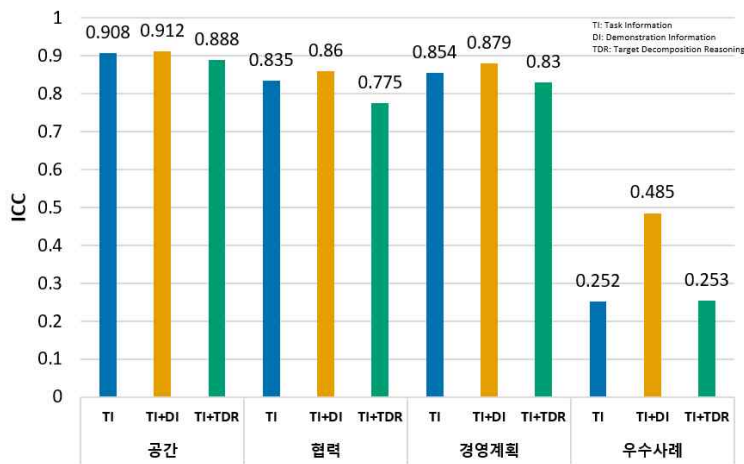
→ 영역 점수: ○점

다른 영역(상호작용, 지속·확산 가능성, 논리적 체계성)도 같은 형식으로 정리하겠습니다.

이제 이 기준은 고정해 두고, 앞으로 증빙자료를 올려 주시면 위 단계에 따라 바로 분석하고, 영역별 3개 문항 점수 + 영역 점수를 도서관별로 정리해 드리겠습니다.

〈표 8〉 프롬프트 기법별 전문가 - AI 간 신뢰도(Inter-rater Reliability)

평가 영역	프롬프트 기법	TI				TI+DI				TI+TDR			
		PCC	ICC(2.1)	95%CI		PCC	ICC(2.1)	95%CI		PCC	ICC(2.1)	95%CI	
				하한값	상한값			하한값	상한값			하한값	상한값
공간		.868	.908	.791	.950	.871	.912	.802	.954	.846	.888	.722	.944
협력		.726	.835	.772	.880	.760	.860	.783	.910	.648	.775	.649	.855
경영계획		.745	.854	.799	.895	.720	.879	.808	.924	.789	.830	.729	.893
우수사례		.182	.252	-.024	.458	.334	.485	.173	.681	.179	.253	-.134	.521



〈그림 2〉 프롬프트 입력에 따른 전문가-AI 신뢰도(ICC)

1차 프롬프트 입력 후 단계적 추론을 유도하는 TDR 프롬프트를 추가하는 방식은 신뢰도 향상에 도움이 되지 않았으며 오히려 감소하는 경향이 나타났다. 이를 통해 목표 분해, 추론을 추가로 지시하는 프롬프트는 공공도서관 운영 평가 맥락에서 AI 평가 기능 향상에 크게 도움이 되지 않으며, 이는 상대적으로 복잡한 판단을 요구하는 정성평가 영역에서도 유사하게 나타났다. 즉, 도서관 운영평가를 위한 프롬프트 입력 시, 단순히 평가 척도와 내용 및 조건 등을 제시하는 정보 제시형 프롬프트보다 예시를 추가로 제공하는 프롬프트가 효과적이며, 특히 정성평가 시 그 효과는 더욱 크게 나타나는 것

으로 확인되었다.

5. 결론 및 제언

본 연구는 생성형 AI 기반 공공도서관 운영 평가 자동화 가능성을 탐색하고자 하였으며 이를 위해 증빙자료 기반으로 사람이 평가하는 4개 평가 영역에 대해 ChatGPT 5.1을 활용한 자동평가를 진행하였다. 이후 전문가 평가 결과와 AI 평가 결과를 분석하고, 두 점수 간 신뢰도를 분석하였다. 이를 통해 두 평가 주체의 특성과 AI 자동평가 도입 가능성, 도입 가능 영

역 등을 확인하였다. 또한, LLM을 기반으로 한 생성형 AI에서는 프롬프트의 역할이 매우 크고 이에 따라 출력 결과가 달라질 수 있다는 점을 고려하여 자동평가의 성능을 최대화할 수 있는 프롬프트 기법을 탐색하였다. 선행연구를 토대로 확인한 프롬프트 엔지니어링 기법 중 본 연구 목적과 맥락에 적합한 3가지 기법을 적용하였다. 프롬프트의 구조를 형성하는 기본 프롬프트인 '과업정보(Task Information, TI)', 추가로 예시를 제시하는 '예시정보(Demonstration Information, DI)', 단계적 추론을 유도하는 '목표 분해 추론(Target Decomposition Reasoning, TDR)' 3가지 프롬프트 기법을 조합하여 검증하였다.

연구 결과를 정리하면 다음과 같다. 전문가 평가점수와 AI 평가점수 간 일치도는 평가 영역별로 다르게 나타났다. 특정 유형의 공간 수, 협력 기관 수 등과 같이 정량평가에 가까운 평가 척도를 사용하거나 증빙자료 내 내용 존재 여부 등을 확인하는 수준으로 비교적 평가 내용이 간단한 영역에서는 전문가와 AI 간 신뢰도가 높게 나타났으며 AI 자동평가가 전문가에 의한 평가를 상당 부분 대체할 수 있을 것으로 예상된다. 하지만 AI 평가에서 점수가 매우 좁은 영역에 분포하여 상위점수 또는 하위점수를 집중적으로 부여하는 경향이 발견되었으며, 이는 AI 평가의 변별력을 검증할 필요가 있음을 의미한다. 반면, 비교적 평가 과정이 복잡하고 평가자의 판단이 필요한 정성평가의 경우, 전문가와 AI 평가점수 간 차이가 큰 것으로 나타났다. 마지막으로 도서관 운영평가라는 특수한 상황에서 적절한 프롬프트 입력 방법을 검증하고 이를 통해 AI 기반 운영평가 자동화의

성능을 높이는 방안을 모색하고자 하였다. 이를 위해 TI, TI+DI, TI+TDR 3가지 프롬프트 조합을 검증하였으며, 그 결과 평가에 관한 기본 정보를 구조화하여 입력하는 TI와 추가 예시를 제시하는 DI를 조합하는 조건에서 전문가와 AI 점수 간 일치도가 가장 높은 것으로 나타났다. 특히 신뢰도가 낮게 나타난 정성평가 영역에서 예시를 추가로 제시했을 때 그 성능이 2배 가까이 상승하여 복잡한 판단이 필요한 평가 영역에서 그 효과가 더욱 크다는 점을 알 수 있다.

이와 같은 결과를 토대로 향후 도서관 운영평가에 생성형 AI 기반 자동화 시스템을 도입하기 위해 고려해야 할 시사점을 도출하였다.

첫째, 평가 기준의 구체성과 명확성이다. 공간, 협력, 경영계획 3개 영역은 각 공공도서관이 제출한 자료를 평가 척도와 실적 인정 및 불인정 기준에 따라 평가한다. 예를 들어, 경영계획(10점 만점) 평가 시, 전문가와 AI 점수 차이가 6점 이상인 경우가 16건 확인되었으며, 그 원인을 분석한 결과 증빙자료 인정 기준 적용 여부에 따라 점수 차이가 나타났다. 해당 영역의 평가 기준에서는 '도서관 운영 전반(대출, 프로그램 등)에 대한 만족도 조사만 인정하며, 개별 프로그램 만족도 혹은 어느 한 영역에 대한 만족도 조사는 불인정(예. 상반기 프로그램 만족도 조사 등)'하는 것으로 제시되었으나 '도서관 운영 전반'의 범위가 불명확하며, '어느 한 영역'의 구분 또한 모호하다. 이러한 문제로 동일한 자료에 대해 전문가는 만점인 10점을 부여하였으나, AI는 0점을 부여하는 경우 또는 그 반대의 경우가 나타났다. 실제 증빙자료를 확인한 결과, 전문가의 점수가 맞는 경우가 9건, AI가 맞는

경우가 7건으로 반드시 전문가 또는 AI의 평가 결과가 일괄적으로 잘못된 것은 아니었으며, 모호한 평가 기준에 의한 사람과 AI의 판단 차이가 있는 것으로 예상된다. 따라서 추후 운영평가에 자동화를 도입하기 위해서는 보다 구체적이고 명확한 평가 기준이 요구되며, 특히 도서관 운영평가 특성상 평가 영역별 성격이 다르고 각 도서관이 제출하는 자료의 내용과 형식 또한 다양하므로 이를 실적으로 인정할 것인가 인정하지 않을 것인가에 대한 보다 명료한 기준이 요구된다. 현행 평가에서 제시되는 기준 또한 과거 평가 사례를 토대로 비교적 구체적인 기준을 제시하고는 있으나, 여전히 해석에 따라 판단이 모호한 경우가 존재하는 것이 사실이다. 따라서 평가 기준을 인정 사례, 불인정 사례로 구분하고 항목별 예시를 제공하는 방식으로 제시한다면 평가자의 임의적인 판단을 최소화할 수 있고, 자동평가의 정확도를 또한 높일 수 있을 것이다.

둘째, 평가의 근거가 되는 증빙자료의 파일 형식 통일이다. 4개 영역 평가를 위해 각 공공도서관에서 제출한 증빙자료의 양은 방대했으며 한글, PDF, 이미지, 엑셀 파일 등 그 형식 또한 다양했다. 실제 ChatGPT 5.1을 활용하여 평가 진행 시, PDF 파일의 인식률이 가장 높은 반면 한글과 엑셀 파일은 인식률이 낮고 평가 불가능한 경우도 존재했다. 하지만 적절한 파일 형식을 제공하는 경우 생성형 AI의 내용 인식률은 매우 높은 것으로 확인되었으며 평가에 필요한 요소를 추출하는 데 문제가 없었다. 실제 협력 영역 평가를 진행하면서 전문가와 AI의 점수가 3점 이상 차이 나는 경우가 존재해 그 원인을 분석한 결과, AI가 평가자료를 제대

로 인식하지 못하고 협력 사례를 불인정한 경우가 대부분이었다. 이처럼 증빙자료의 파일 형태는 특히 AI를 활용한 자동평가 상황에서 평가 결과에 영향을 미칠 수 있는 부분이기 때문에 일관성이 요구된다. 또한, 각 도서관이 제출하는 증빙자료의 양이 방대하고 이를 정리·변환하는 과정에서 많은 시간이 소요된다는 점을 고려했을 때 효율적인 자동평가 시스템을 도입하기 위해서는 증빙자료의 파일 형식을 통일할 필요성이 있다. 실제 우수사례 영역의 경우, 각 도서관의 운영 사례를 정해진 양식에 따라 작성하여 제출하도록 사전에 안내하였으며 파일 형식 및 내용 구조가 통일되어 있어 AI 자동평가 시 내용 인식을 또한 높은 것으로 나타났다.

셋째, 생성형 AI 답변의 안정화이다. ChatGPT 5.1을 활용하여 도서관 운영평가를 진행하는 과정에서 처음에 프롬프트로 제시한 평가 척도가 중간에 변경되는 사례가 종종 나타났다. 평가 시작 시, 생성형 AI에 평가 척도와 기준을 입력하고 이에 따라 평가를 진행하도록 프롬프트를 입력하였으나 평가를 진행하면서 척도가 임의로 변경되는 현상이 나타났으며, 특히 평가 초반부에는 이러한 상황이 더욱 빈번하게 발생하여 프롬프트를 다시 입력해야 하는 상황이 발생했다. 이와 같은 현상은, 사전 학습 이후에도 사용자가 입력한 프롬프트 내용과 예시를 활용해 새로운 과제에 적응하는 대규모 언어모델의 문맥 내 학습(in-context learning) 능력과 관련이 있으며, 사용자가 제공한 예시와 선호 데이터를 기반으로 추가 학습이 이루어질 때 동일한 질의에 대한 AI의 응답이 달라지는 현상과도 연관된다(Brown et al., 2020; Ouyang et al., 2022). 즉 생성형 AI는 이전 프롬프트를

통해 입력받은 대량의 증빙자료와 텍스트를 학습하여 다음 도서관 평가에 반영하며, 이 과정에서 평가 척도를 조정하는 것으로 해석된다. 따라서 생성형 AI를 기반으로 한 도서관 운영평가 자동화를 위해서는 AI의 학습 능력을 고려한 평가의 신뢰도와 안정성 확보가 필수적이다. 이를 위해 평가 척도를 변경 불가능한 고정값으로 명시하여 평가 과정에서 임의의 조정이 일어나지 않도록 하고, 동일 사례에 대한 반복 평가를 통해 AI 응답의 내적 일관성을 점검하며, 표본 일부를 사람과 AI가 이중 평가하여 일치도를 확인하는 방식 등으로 검증 절차를 체계화할 필요가 있다. 또한, AI의 평가 로그를 주기적으로 분석해 평가 척도 적용 방식 변동 여부를 모니터링함으로써 평가 전·중·후 전 단계에 걸친 다층적 검증 장치를 구축해야 할 것이다.

넷째, 도서관 운영평가에 특화된(domain-specific) 프롬프트 설계이다. 본 연구에서 3개의 프롬프트를 조합하여 적용했을 때, 전문가와 AI 간 신뢰도 차이가 나타나는 것을 확인하였으며, 특히 정성평가 영역에서 그 효과가 매우 크게 나타났다. 4개 영역에서 모두 기본 정보를 구조화하여 제공하는 프롬프트와 예시를 제공하는 프롬프트(TI+DI) 조합이 가장 안정적인 것으로 확인되었다. ChatGPT 5.1이 평가 중간에 임의로 평가 기준을 변경하는 상황이 현저히 적게 나타나 안정적인 평가가 가능했고, 전문가와 AI 간 신뢰도 분석 결과에서도 TI+DI 조합의 신뢰도가 가장 높은 것으로 나타났다. 이는 현재 평가 체계에서는 TI와 DI 두 가지 프롬프트 기법을 조합하여 사용하는 방식이 평가의 안정성과 정확도를 높이는 방법임을 시사

한다. 이처럼 적절한 프롬프트 적용에 따라 자동평가의 성능 차이가 존재함을 고려하여 이후 도서관 운영평가 맥락과 평가 척도 등 분야의 특수성을 충분히 반영한 특화된 프롬프트가 설계되어야 할 것이며 이는 궁극적으로 자동평가 성능 향상의 핵심 요소로 작용할 것이다.

다섯째, 도서관 운영평가 자동화 과정에서의 데이터 보안이다. 생성형 AI를 활용하여 평가 수행 시 각 도서관의 운영 관련 데이터를 처리하고 저장하는 단계에서 발생할 수 있는 보안 문제는 중요한 고려 요소이다. 특정 도서관을 식별할 수 있는 예산, 인력, 서비스 수준 등의 정보가 AI에 입력되거나 재학습에 활용되는 경우, 외부에 민감 정보가 노출되지 않도록 입력 데이터 최소화, 익명처리 및 코드화, 기술적 보호조치 등에 관한 사항을 명확히 규정할 필요가 있다. 또한, 평가 대상 도서관에 관한 정보뿐만 아니라 평가 전 과정에서 생성되는 로그 또한 안전하게 관리할 수 있는 체계를 구축해야 할 것이다.

이와 같이 본 연구는 도서관 운영평가에 자동화 과정을 도입하기 위한 기초연구로서 생성형 AI 도입 가능성과 고려해야 할 시사점을 도출하였다는 데 그 의의가 있다. 하지만 본 연구는 AI 자동평가 결과의 신뢰도 검증을 위해 전문가 1인의 평가 결과만을 기준으로 했기 때문에 해당 전문가의 판단이 결과에 직접적으로 반영되었으며, 이에 따른 주관성을 완전히 배제하기 어렵다. 이와 더불어 여러 명의 평가자 간 신뢰도와 생성형 AI 평가 결과의 내적 일관성을 검증하지 못했다는 점에서 본 연구에서 제시한 AI 평가 결과를 일반화하는 데에는 신중한 해석이 요구된다. 또한, 일부 상황에서는

AI 평가가 전문가보다 더 적절할 가능성이 존재한다는 점에서 본 연구 결과는 전문가 평가를 기준으로 한 상대적·탐색적 비교로 이해되어야 한다. 이와 같은 한계를 보완하기 위하여 후속 연구에서는 여러 전문가 집단의 평가 결과를 기준으로 생성형 AI 평가의 신뢰도를 검증하고, 반복 평가를 통해 AI 응답의 일관성을 확인하는 과정이 필요할 것이다. 또한, 본

연구는 프롬프트 기법에 따른 자동평가의 성능 변화를 검증하는 데 초점을 두고 있으므로 향후 이를 기반으로 공공도서관 운영평가 맥락을 충분히 반영한 특화된 프롬프트 프레임워크를 제안하는 연구가 진행될 수 있을 것이며, 이를 통해 생성형 AI 기반 자동평가의 가능성을 보다 구체적이고 정교하게 검증할 수 있을 것이다.

참 고 문 헌

- 감미아, 이지연 (2023). 공공도서관 평가의 최근 변화 분석 및 향후 방향에 대한 제안. *한국비블리아학회지*, 34(1), 189-207. <http://doi.org/10.14699/kbiblia.2023.34.1.189>
- 김민경 (2025). Building an automated scoring system for EFL learners' paraphrases via a customized GPT. *영어어문교육*, 31(1), 27-57. <http://doi.org/10.35828/etak.2025.31.1.27>
- 김홍렬 (2013). 공공도서관 운영평가지표의 개선 내용에 관한 연구. *한국도서관·정보학회지*, 44(2), 77-95. <https://doi.org/10.16981/kliss.44.2.201306.77>
- 문화체육관광부 (2022). 2022년도('21년 실적) 전국도서관 운영평가 결과보고서.
- 문화체육관광부 (2025). 2026년('25년 실적) 공공도서관 운영평가 지침.
- 박예은, 김동훈, 유시현, 배운진, 김경희 (2024). 비즈니스 데이터 분석 리포트 생성을 위한 생성형 AI 프롬프트 엔지니어링 연구. *한국정보기술학회논문지*, 22(11), 167-180. <https://doi.org/10.14801/jkiit.2024.22.11.167>
- 박지우, 이지현, 윤민아, 김동현, 정준영, 이영한 (2025). 근골격 영상의학에서 대형 언어모델과 생성형 인공지능의 임상 적용, 과제 및 한계, 그리고 향후 활용을 위한 권고사항. *Journal of the Korean Society of Radiology*, 86(5), 655-670. <https://doi.org/10.3348/jksr.2025.0018>
- 신병철, 이준수, 유연주 (2024). 프롬프트 엔지니어링을 통한 GPT-4 모델의 수학 서술형 평가 자동 채점 탐색: 순열과 조합을 중심으로. *수학교육*, 63(2), 187-207. <https://doi.org/10.7468/mathedu.2024.63.2.187>
- 윤명희, 김기영 (2009). 공공도서관 평가지표 유용성에 대한 사서들의 인식에 관한 연구: 2008년 전국도서관 운영평가 중 공공도서관 평가지표를 중심으로. *한국문헌정보학회지*, 43(3), 221-245. <https://doi.org/10.4275/KSLIS.2009.43.3.221>

- 차미경, 표순희 (2015). 전국 공공도서관 운영평가의 성과에 관한 연구: 2010년~2013년도를 중심으로. *한국비블리아학회지*, 26(2), 241-268. <https://doi.org/10.14699/kbiblia.2015.26.2.241>
- Austin, J. T. & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, 120(3), 338-375. <https://doi.org/10.1037/0033-2909.120.3.338>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, D., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. <https://doi.org/10.48550/arXiv.2005.14165>
- Chen, B., Zhang, Z., Langrene, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6), 101260. <https://doi.org/10.1016/j.patter.2025.101260>
- Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. New York: Routledge.
- Ekin, S. (2023). *Prompt Engineering for ChatGPT: A Quick Guide to Techniques, Tips, and Best Practices*. New Jersey: Authorea Preprints.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199-22213. <https://doi.org/10.48550/arXiv.2205.11916>
- Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (LLaMA) using medical domain knowledge. *Cureus*, 15(6). <https://doi.org/10.7759/cureus.40895>
- Liu, Y. Y., Zheng, Z., Zhang, F., Feng, J. C., Fu, Y. Y., Zhai, J. D., He, B. S., Zhang, X., & Du, X. Y. (2025). A comprehensive taxonomy of prompt engineering techniques for large language models. *Frontiers of Computer Science*, 20(3), 2003601. <https://doi.org/10.1007/s11704-025-50058-z>
- Logan IV, R., Balažević, I., Wallace, E., Petroni, F., Singh, S., & Riedel, S. (2022, May). Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics 2022*, 2824-2835.

<https://doi.org/10.18653/v1/2022.findings-acl.222>

McGraw, K. O. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1996(1), 30-46.

<https://doi.org/10.1037/1082-989X.1.1.30>

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744. <https://doi.org/10.48550/arXiv.2203.02155>

Roy, D., Zhang, X., Bhave, R., Bansal, C., Las-Casas, P., Fonseca, R., & Rajmohan, S. (2024, July). Exploring llm-based agents for root cause analysis. *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 208-219. <https://doi.org/10.1145/3663529.3663841>

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., Da Costa, H., Gupta, S., Rogers, M. L., Goncarenco, I., Sarli, G., Galyunker, I., Peskoff, D., Carpuat, M., White, J., Anadkat, S., Hoyle, A., & Resnik, P. (2024). The prompt report: A systematic survey of prompt engineering techniques. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2406.06608>

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428. <https://doi.org/10.1037/0033-2909.86.2.420>

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 186345. <https://doi.org/10.1007/s11704-024-40231-1>

Yue, S., Chen, W., Wang, S., Li, B., Shen, C., Liu, S., Zhou, Y., Xiao, Y., Yun, S., Huang, X., & Wei, Z. (2023). Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2309.11325>

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Cha, Mikyeong & Pyo, Soon Hee (2015). A study on the results of the national evaluation on

- public library management during 2010~2013. *Journal of the Korean Biblia Society for Library and Information Science*, 26(2), 241-268.
<https://doi.org/10.14699/kbiblia.2015.26.2.241>
- Kam, Miah, & Lee, Jeeyeon (2023). Analysis of recent changes in public library evaluation and recommendations for future directions. *Journal of the Korean Biblia Society for Library and Information Science*, 34(1), 189-207. <http://doi.org/10.14699/kbiblia.2023.34.1.189>
- Kim, Hong Ryul (2013). A study on the improvement in evaluation indicators of public libraries. *Journal of Korean Library and Information Science Society*, 44(2), 77-95.
<https://doi.org/10.16981/kliss.44.2.201306.77>
- Kim, Minkyung (2025). Building an automated scoring system for EFL learners' paraphrases via a customized GPT. *English Language & Literature Teaching*, 31(1), 27-57.
<http://doi.org/10.35828/etak.2025.31.1.27>
- Ministry of Culture, Sports and Tourism (2022). Report on the Results of the 2022 National Library Operation Evaluation(Performance in 2021).
- Ministry of Culture, Sports and Tourism (2025). Guidelines for the 2026 Public Library Operation Evaluation(Performance in 2025).
- Park, Jiwoo, Lee, Ji Hyun, Yoon, Min A, Kim, Dong Hyun, Jung, Joon Yong, & Lee, Young Han (2025). Clinical applications, challenges & pitfalls, and recommendations for large language model and generative AI in musculoskeletal imaging. *Journal of the Korean Society of Radiology*, 86(5), 655-670. <https://doi.org/10.3348/jksr.2025.0018>
- Park, Yeeun, Kim, Donghun, Yoo, Sihyeon, Bae, Yunjin, & Kim, Keungoui (2024). Research on prompt engineering for generative AI to create business data analysis reports. *The Journal of Korean Institute of Information Technology*, 22(11), 167-180.
<https://doi.org/10.14801/jkiit.2024.22.11.167>
- Shin, Byoungchul, Lee, Junsu, & Yoo, Yunjoo (2024). Exploring automatic scoring of mathematical descriptive assessment using prompt engineering with the GPT-4 model: Focused on permutations and combinations. *The Korean Society of Mathematical Education*, 63(2), 187-207. <https://doi.org/10.7468/mathedu.2024.63.2.187>
- Yoon, Myung Hee & Kim, Giyeong (2009). A study on the perception of librarians on the usefulness of measures in public library operation assessment: In the 2008 Korean library operation assessment. *Journal of the Korean Society for Library and Information Science*, 43(3), 221-245. <https://doi.org/10.4275/KSLIS.2009.43.3.221>

