학술논문 초록 작성에 대한 대형 언어모델 적용 가능성 탐색*

Exploring the Applicability of Large Language Models for Academic Abstract Writing

김 유 미 (Yumi Kim)**
양 승 원 (Seungwon Yang)***
이 종 욱 (Jongwook Lee)****

목 차

초 록

ChatGPT의 공개 이후 연구자들이 LLM을 적극적으로 활용하면서, 생성형 AI가 학술 글쓰기 과정에서 수행할 수 있는 역할에 대한 관심이 높아지고 있다. 이에 본 연구는 논문의 핵심 내용을 요약하는 초록 작성 단계에서 AI 활용 가능성과 한계를 살펴보고자 하였다. 이를 위해 2022년부터 2024년까지 '한국문헌정보학회지'에 출판된 논문 204편을 대상으로 하여, ChatGPT가 생성한 초록과 저자가 작성한 원초록을 비교·분석하였다. 원초록과 생성초록, 본문, 그리고 프롬프트 유형별 유사도를 정량적으로 비교하고 전문가 인식 조사를 병행하였다. 분석 결과, 원초록과 생성초록은 BERT와 TF-IDF 기준에서 의미적으로 유사했지만, 표현과 어휘 사용에서 차이를 보였다. 또한 한국어 프롬프트로 생성된 초록이 원초록 및 본문과의 유사도가 가장 높게 나타나, 프롬프트 언어가 생성초록의 표현과 내용 반영 정도에 영향을 미친 것으로 나타났다. 전문가들은 LLM을 문체 개선이나 표현 보완에 유용한 보조적 도구로 인식하였다. 본 연구는 이러한 결과를 토대로 인간과 AI가 상호 보완적으로 참여하는 협업적 초록 작성의 가능성을 제시하였다.

ABSTRACT

Since the release of ChatGPT, researchers have shown growing interest in using large language models (LLMs) for academic writing. This study explored the possibilities and limitations of AI use in summarizing the core content of research papers into abstracts. We analyzed 204 papers published in the Journal of the Korean Society for Library and Information Science (2022-2024) by comparing author-written abstracts with those generated by ChatGPT. Quantitative analyses examined similarities among the originals, AI-generated abstracts, full texts, and different prompt types, supplemented by expert perception. Results showed that AI-generated abstracts were semantically close to the originals based on BERT and TF-IDF scores but differed in word choice and expression. Abstracts generated with Korean prompts showed the highest similarity to both the originals and the full texts, indicating that the prompt language affected style and content representation. Experts viewed LLMs as helpful tools for improving clarity and fluency in writing. Overall, the findings suggest the potential of LLMs as collaborative partners in abstract writing.

키워드: 학술 초록, 대형 언어모델, ChatGPT, 프롬프트 엔지니어링, 인간-AI 협업
Research Abstract, Large Language Model (LLM), ChatGPT, Prompt Engineering, Human-AI Collaboration

---

* 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2025S1A5B5A19019132).
** 경북대학교 문헌정보학과 박사과정(yumikim@knu.ac.kr / ISNI 0000 0005 2873 1486) (제1저자)
*** Associate Professor, School of Information Studies, Center for Computation and Technology Louisiana State University (seungwonyang@lsu.edu / ISNI 0000 0005 1448 8630) (공동저자)
**** 경북대학교 문헌정보학과 부교수(jongwook@knu.ac.kr / ISNI 0000 0004 6830 6145) (교신저자)
논문접수일자: 2025년 10월 20일 최초심사일자: 2025년 10월 31일 게재확정일자: 2025년 11월 10일
한국문헌정보학회지, 59(4): 177-198, 2025. http://dx.doi.org/10.4275/KSLIS.2025.59.4.177

1. 서 론

학술논문의 초록은 독자에게 연구의 핵심 내용을 간결하게 전달한다. 연구자들은 초록을 바탕으로 검색된 논문의 적합성을 신속히 판단할 수 있다. 또한 정확하고 명확한 초록은 학술 데이터베이스에서의 논문 검색 효율성을 높여, 연구자들의 정보 탐색 시간을 절약하는 데에도 도움이 된다. 나아가 초록은 연구 분야의 지적 구조를 분석하거나(서선경, 정은경, 2013; 신주은, 김성희, 2021), 연구동향을 파악하는 데 있어 가장 널리 활용되는 데이터 중 하나이다. 예컨대, 초록에서 키워드를 추출하거나 토픽모델링을 활용하여 국내 문헌정보학 연구의 주제 변화를 분석한 연구(박자현, 송민, 2021)나 초록 데이터를 기반으로 한 국내외 통계학 분야 연구동향 분석(양종훈, 곽일엽, 2021)이 그 사례에 해당된다.

나아가 일부 학회는 심사 효율성을 위해 제출된 초록만을 검토하여 심사 대상 여부를 조기 판단하기도 하며(American Chemical Society, 2024), 초록의 품질은 논문 채택 가능성과 심사자의 평가 과정에도 직접적인 영향을 미친다. 명확하고 설득력 있는 초록은 편집자에게 긍정적인 첫인상을 주어 심사 단계로의 진입 가능성을 높이고, 심사자에게 연구의 중요성과 완성도를 효과적으로 전달함으로써 평가 과정 전반에 유리하게 작용한다(Ketcham et al., 2010).

이러한 초록의 중요한 역할에 따라 많은 연구자들은 초록의 품질을 향상시키기 위한 노력을 기울여 왔다. 이 과정에서 초록의 구조와 작성 원칙을 제시하거나, 양질의 초록에 포함되어야 할 핵심 요소를 정의하려는 연구들이 꾸준히 발표되고 있다. 구체적으로 초록 작성의 원칙을 제안한 연구(Ali et al., 2020)를 비롯하여, 초록의 유형을 분류한 연구(Delving et al., 2014)와 내용적 균형을 강조한 연구(Klimova, 2020)가 제시된 바 있다. 또한 이재윤(2024)은 국내 문헌정보학 학술지를 대상으로 구조적 초록 도입 방안을 제시하였다. 그럼에도 불구하고 실제 초록은 작성자 개인의 역량이나 학문적 관습에 따라 편차가 크며, 연구 성격이나 학술지마다 요구하는 기준이 달라 일관성을 확보하기 어렵다는 한계가 존재한다.

이러한 한계점을 보완하기 위하여 초록 자동생성에 대한 다양한 시도가 이루어져왔다. 자동요약 연구는 1950년대부터 꾸준히 진행되어 왔으며, 초기에는 TextRank, LSA, Pointer-Generator Network와 같은 기법들이 주로 활용되었다(Mihalcea & Tarau, 2004; Gong & Liu, 2001). 그러나 이러한 접근 방식은 문맥 이해에 근본적인 제약이 있고, 자연스러운 표현을 구현하기 어렵다는 한계가 있었다. 최근에는 트랜스포머 기반의 대규모 언어모델(Large Language Model, 이하 LLM)의 등장으로 문맥 이해와 자연스러운 텍스트 생성 능력이 크게 향상되면서 자동 요약 연구는 새로운 전환점을 맞이하였다. GPT(Generative Pre-trained Transformer), BART(Bidirectional and AutoRegressive Transformer), T5(Text-to-Text Transfer Transformer) 등 다양한 모델들이 개발되었으며, 특히 ChatGPT는 2022년 말 공개된 이후 빠르게 확산되며 학계와 연구자들로부터 큰 주목을 받았다.

그러나 LLM이 생성하는 요약은 도메인에 따라 품질이 달라질 수 있으며(Widyassari et

al., 2022), 프롬프트 엔지니어링(Leidinger et al., 2023)이나 모델의 특성(Luo et al., 2022)에 크게 의존한다는 점에서 여전히 해결해야 할 과제가 많다. LLM의 급속한 발전과 더불어, ChatGPT의 공개 이후 AI가 생성하거나 수정한 텍스트의 비율이 급격히 증가하였으며, 특히 컴퓨터 과학, 공학, 생물학 등 기술 중심 분야에서 활용이 두드러졌다. 또한 비영어권 국가인 중국, 이탈리아, 인도 등에서 AI 활용 비율이 상대적으로 높게 나타나, 생성형 AI가 학문전반에 걸쳐 널리 사용되고 있음을 보여준다(Cheng et al., 2024). 그럼에도 불구하고 학술논문 요약에 있어 LLM의 적용 가능성이나 성능을 체계적으로 살펴본 연구는 충분하지 않다(Zhao et al., 2023).

이에 본 연구는 LLM 중에서 ChatGPT를 활용하여 학술논문 초록 작성 가능성을 탐색하고자 한다. 구체적으로 2022년부터 2024년까지 '한국문헌정보학회지'에 출판된 논문을 대상으로 원초록과 ChatGPT가 생성한 초록을 정량적 및 정성적으로 비교・분석하여, LLM 기반 초록 생성의 성능을 살펴보고자 하였다. 특히, 논문 초록 생성을 위한 여섯 가지 유형의 프롬프트(Base, Base-Korean, Structure, Two-shot, Role, All combined)를 적용하여 프롬프트 엔지니어링의 영향도 함께 분석하였다. 본 연구 결과는 논문 초록 생성의 LLM 적용 가능성을 제시함으로써 효과적이고 품질 높은 초록을 작성하는 데 기여할 것으로 기대된다.

2. 개념적 배경 및 관련 연구

2.1 학술논문 초록의 개념과 유형

학술논문의 초록은 논문의 핵심 내용을 간결하면서도 포괄적으로 제시하는 요약으로, 독자가 논문의 내용을 빠르게 파악하고 자신의 관심사와의 관련성을 판단할 수 있도록 돕는다(Klimova, 2015). Lancaster(2003)는 초록을 문헌의 내용을 간결하면서도 정확하게 표현한 것으로 정의하고 있다. 그는 초록의 유형을 지시적 초록(indicative abstract)과 정보적 초록(informative abstract)으로 구분하고 있는데, 지시적 초록은 문헌이 무엇에 관한 것인지를 기술하여 독자가 해당 논문의 관련성을 판단할 수 있도록 돕는 반면, 정보적 초록은 연구의 목적, 방법론뿐만 아니라 결과, 결론, 권고사항까지 구체적으로 기술하여 일부 목적상 원문 읽기를 대체할 수 있을 정도의 정보를 제공한다. 실제로 두 유형은 명확히 구분되기보다는 혼재되어 나타나며, 두 유형의 특성을 모두 포함하는 경우가 많다. 특히 자연과학 및 의학 분야에서는 연구 결과의 재현성과 검증 가능성을 중시하여 정보적 초록이 표준으로 자리 잡은 반면, 인문학이나 이론적 논의가 중심인 연구에서는 지시적 초록이 더 적합할 수 있다는 견해가 있다(Hartley, 2004).

초록은 또한 정보 전달과 선택을 돕는 복합적인 기능을 가진다. 잘 준비된 초록은 독자가 문헌의 핵심 내용을 빠르게 파악하여 전체 문헌을 읽을지 여부를 결정하는데 도움을 주며, 관심이 없는 논문을 선별하는 데 드는 시간과 노력을 절약하게 한다(Lancaster, 2003). 이와

함께 초록은 논문에 대한 관심을 유도하는 역할을 하며, 경우에 따라 완성도 높은 정보적 초록은 원문 읽기를 일정 부분 대체할 수도 있다(Lancaster, 2003). 나아가 초록은 색인과 정보검색 서비스가 논문을 색인하고 검색할 수 있도록 하며(Klimova, 2015), 연구자들에게 자신의 관심 분야에서 새로 출판된 문헌에 대한 최신 정보를 제공하는 역할도 수행한다(Lancaster, 2003).

2.2 LLM을 활용한 학술 텍스트 요약

최근 GPT 계열로 대표되는 대규모 언어모델은 방대한 텍스트 데이터를 기반으로 문맥 이해와 자연스러운 언어 생성 능력을 획기적으로 향상시켰다. 이러한 기술의 발전은 자동 요약 연구에도 새로운 전환점을 가져왔다. 기존의 추출적 요약 방식에서 벗어나, LLM을 활용하여 문서의 의미 구조와 논리적 흐름을 반영하는 생성적 요약이 가능해졌다(Zhao et al., 2023). 이와 같은 변화는 LLM이 다양한 도메인의 방대한 텍스트를 학습하면서 문맥을 깊이 이해하고 정보를 재구성하는 능력을 갖추게 되었기 때문으로 여러 연구에서 요약 성능 향상이 보고되고 있다(Liu et al., 2022; Zhang et al., 2023). 실제로 GPT-3.5와 같은 LLM이 사람이 작성한 참조 요약(reference summary)보다도 사람들에게 더 자연스럽고 완성도 높은 요약을 만들어낸다는 결과가 제시되고 있다(Liu et al., 2022; Zhang et al., 2023). 이는 LLM이 단순히 문장을 압축하는 수준을 넘어, 사람이 읽기 좋은 방식으로 내용을 재구성할 수 있음을 보여준다. 그러나 동시에 문서 내 세부 정보를 부정확하게 재구성하거나, 존재하지 않는 내용을 만들어내는 환각(hallucination) 문제가 빈번하게 보고되고 있다(Widyassari et al., 2022). 이러한 한계로 인해 최근 연구들은 LLM의 요약 품질을 향상시키기 위한 프롬프트 엔지니어링 기법과 생성 결과의 신뢰성을 체계적으로 평가하는 방안에 주목하고 있다.

이러한 흐름 속에서 최근 연구들은 모델과 이용자 간의 상호작용을 효과적으로 조정하기 위한 입력 설계, 즉 프롬프트의 구성 방식에 주목하고 있다. LLM으로부터 원하는 결과를 얻기 위해서는 프롬프트가 구체적인 맥락 정보를 포함해야 하며, 이러한 정교한 프롬프트를 구성하는 과정과 기술을 '프롬프트 엔지니어링'이라 한다(Marvin et al., 2024). 여러 연구에서 AI와 상호 작용할 때 성능을 향상시키기 위해 프롬프트 엔지니어링의 중요성을 강조하고 있다. 예를 들어, Leidinger et al.(2023)은 프롬프트 구조와 지시문의 세부 표현이 결과물의 품질에 결정적인 영향을 미친다고 보고하였으며, Schulhoff et al.(2024)은 동일한 모델이라도 언어, 예시 제공 여부, 역할(Role) 설정에 따라 요약문의 품질이 상이함을 확인하였다. Luo et al.(2022) 또한 모델 크기와 미세조정의 수준이 생성 요약의 논리적 완결성과 정보 충실도에 영향을 줄 수 있음을 밝혀내었다.

학술 텍스트 요약을 대상으로 한 연구들도 증가하고 있다. Cho et al.(2023)은 인문·사회과학 분야 논문 초록을 대상으로 GPT 모델이 생성한 요약과 원저자 초록을 비교한 결과, LLM이 논리 구조와 핵심 주제는 잘 포착하지만 연구의 맥락적 배경과 강조점은 상대적으로 약하게 반영하는 경향이 있음을 보고하였다. Cheng et al.(2024) 또한 ChatGPT가 생성한 학술 텍

스트를 분석하여, 간결성과 일관성에서는 우수하지만, 세부 내용의 충실성과 내용의 정확성은 상대적으로 낮다고 지적하였다. 그럼에도 두 연구 모두 LLM이 보조적 도구로 활용될 때 학술 초록의 품질을 향상시킬 수 있음을 강조하였다.

LLM 기반 학술 텍스트 요약에서 프롬프트의 역할을 살펴본 연구도 존재한다. Nabata et al.(2025)은 ChatGPT 기반 논문 초록 생성 과정에서 구조화된 프롬프트가 비구조화된 단순 명령문보다 핵심정보의 조화와 논리적 흐름의 완결성을 높인다고 밝혔다. 이들은 Ufnalska & Hartley(2009)가 제시한 초록 평가 기준(이해가능성, 구조, 정보선택, 간결성 등)을 적용하여, AI가 생성한 초록은 언어적 정확성과 문장 흐름에서는 높은 점수를 보였으나 연구 목적의 명확성과 맥락 반영 측면에서는 낮은 점수를 보였다. 이는 LLM의 잠재력과 한계를 동시에 보여주며, 인간의 비판적 검토와 AI의 자동화된 요약 능력을 결합한 인간과 AI 협업이 필요함을 시사한다.

2.3 한국어 문서 요약과 품질 평가 연구

국내에서도 트랜스포머 기반 언어모델을 활용한 한국어 문서 요약 연구가 꾸준히 이루어지고 있다. 초기 연구들은 주로 사전학습 언어모델을 적용하여 기존 통계 기반 요약과의 성능 차이를 검증하는데 초점을 맞추었다. 예를 들여, 김의순과 임희석(2021)은 법률 문서 요약에서 사전학습 언어모델을 적용하여, 기존 TF-IDF 기반 요약보다 문맥 이해 측면에서 우수한 성능을 확인하였다. 송의석과 김남규(2021)는 트랜스포머 구조를 활용한 생성형 요약 모델을 설계하여, 추출적 요약보다 자연스러운 문장 구성을 구현하였다.

이후 연구들은 특정 도메인에 특화된 모델 개발로 확장되었다. 박재언 외(2022)은 BERT 기반 한국어 문서 추출 요약 베이스라인을 설계하여, 한국어 어순과 문법 특성을 반영한 성능 향상을 시도하였으며, 윤수환 외(2021)는 주제 속성(attribute)을 고려한 주제 키워드 기반 문서 요약을 제안하였다. 최근에는 이영재 외(2025)가 대형 언어모델을 활용하여 특허 핵심내용(청구항)을 자동으로 요약·생성하는 연구를 수행하여 한국어 환경에서의 LLM 적용 가능성을 실험적으로 보여주었다. 이들 연구는 대부분 뉴스, 법률, 특허 등 특정 영역에 국한되어 있으며, 학술 텍스트 전반을 대상으로 한 LLM 기반 요약 품질 평가 연구는 아직 드문 편이다.

요약문의 품질을 정량적으로 평가하려는 연구도 꾸준히 이루어지고 있다. 예를 들어, 고은정과 김남규(2018)는 완전성(completeness)과 간결성(conciseness)을 기준으로, 사람이 작성한 참조 요약문 없이도 자동 평가가 가능한 방법을 제시하였으며, 윤세휘와 신유현(2024)은 다양한 자동 평가 지표(ROUGE, BLEU, BERTScore 등)와 인간 평가 간 상관관계를 분석하여, 한국어 생성 요약 평가에 적합한 지표 선택의 근거를 마련하였다. 또한 강준영 외(2024)는 LLM 기반 대화 생성에서 발생하는 환각 문제를 평가하기 위해 자연스러움, 자기 의인화 배제, 지식 기반 여부의 세 가지 기준을 활용하여 자동 평가 방법을 설계하고, 인간 평가 결과와의 유사성을 검증하였다.

이러한 연구들은 한국어 환경에서의 LLM 성

능을 탐색하고 평가 방법론을 발전시켰다는 점에서 의의가 있지만, 여전히 학술 초록을 대상으로 한 실증적 연구는 부족하다. 또한 대부분 단일 모델과 단일 프롬프트 중심의 실험에 그쳐 다양한 프롬프트 설계나 인간 전문가 평가를 병행한 시도는 찾아보기 힘든 편이다.

3. 연구방법

3.1 데이터셋 개요

본 연구에 사용한 데이터는 '한국문헌정보학회지'에 2022년부터 2024년까지 출판된 논문들로 구성하였다. 한국문헌정보학회지는 한국연구재단의 우수등재지로 등재되어 있으며, 문헌정보학 분야에서 학문적 대표성과 신뢰성을 갖춘 학술지로 실험 데이터셋으로 적합하다고 판단하였다. 한국문헌정보학회 공식 홈페이지(https://kslis.jams.or.kr/)에서 각 논문의 원문 PDF 파일을 수집하였으며, 영어로 작성된 논문 4편을 제외한 총 204편의 국문 논문을 분석 대상으로 선정하였다. 연도별 논문 수는 각각 2024년(제58권 1호~4호) 66편, 2023년(제57권 1호~4호) 68편, 2022년(제56권 1호~4호) 70편이다.

ChatGPT를 활용한 초록 생성 시 원초록의 영향을 받지 않도록 수집된 PDF 파일에서 원초록과 참고문헌을 제거 후, Python의 PyMuPDF 라이브러리를 활용하여 본문 텍스트를 추출하였다. 이 연구에서는 GPT-4o를 LLM AI로 사용하였고, Open API 키를 통해 Python 환경에서 초록을 생성하였다. 총 6개의 서로 다른 유형의 프롬프트를 적용하여 초록을 생성하였다. 이들 프롬프트의 구체적인 설계와 구성 방식은 다음 절에 기술하였다.

3.2 프롬프트 설계 및 유형 구분

최근 다국어 환경에서 영어 프롬프트의 성능이 우수하다는 연구 결과가 다양한 자연어 처리 과제에서 반복적으로 보고되고 있다. Lai et al.(2023)은 다국어 질의응답 과제에서 영어로 제시한 프롬프트가 한국어, 일본어, 스페인어 등 각 목표 언어(즉, 실제 응답이 이루어지는 언어)로 제시한 프롬프트보다 높은 정확도를 보였으며, Kang et al.(2023) 역시 영어 프롬프트가 전반적으로 가장 우수한 성능을 보였다고 보고하였다. 이러한 경향은 질의응답을 넘어 요약, 추론, 번역 등 다양한 과제에 걸쳐 확장되며(Dey et al., 2024), 영어 프롬프트의 효과가 널리 인식됨에 따라 다양한 연구에서 실험 설계 시 영어 지시문을 기본으로 활용하고 있다(Ahuja et al., 2023).

이에 본 연구에서도 초록 생성을 위한 대부분의 프롬프트를 영어로 구성하였다(〈표 1〉 참고). 모든 프롬프트는 '한국문헌정보학회지'의 초록 작성 기준에 따라, 500자 이내로 단락 구분 없이 연속적으로 서술할 것을 요구하였으나, 원초록의 분포가 400자에서 600사이에 분포하고 있음을 발견하고, 이에 따라 프롬프트 조건을 조정하였다.

프롬프트 유형은 (1) 기본 지시만을 포함한 단순 프롬프트(Base), (2) 언어 조건에 따른 효과를 확인하기 위한 한국어 프롬프트(Base-Korean), (3) 연구 목적, 방법, 주요 결과를 포함

〈표 1〉 ChatGPT 기반 초록 생성을 위한 프롬프트 유형

프롬프트 유형	내용
Case 1 (Base)	Please write an abstract for the following academic paper. The abstract must be written in Korean as a single paragraph without any breaks and must be between 400 and 600 characters in length.
Case 2 (Base-Korean)	다음은 한 편의 학술논문 본문입니다. 이 내용을 바탕으로 초록을 한국어로 작성해 주세요. 초록은 400자 이상 600자 이내로, 단락 구분 없이 연속적으로 작성해 주세요.
Case 3 (Structure)	**Base Promprt +** The abstract must include the research purpose, methodology, and main findings.
Case 4 (Few-shot)	**Base Prompt +** Here are two examples of well-written academic abstracts in Korean: Example 1, Example 2[1)] [1)]2021년 발표된 한국문헌정보학회지 논문 중 피인용 횟수 상위 5편을 대상으로, 전문가 검토를 통해 두 편을 선정함
Case 5 (Role)	**Base Prompt +** You are a senior researcher with over 20 years of experience in the field of Library and Information Science.
Case 6 (All combined)	**Base Prompt + Structure + Few Shot + Role**

하라는 구조화 지시 추가 프롬프트(Structure), (4) 고품질 초록 2편의 예시를 제시하는 two-shot 프롬프트(few-shot), (5) 프롬프트 작성자를 문헌정보학 분야 20년 이상 경력을 지닌 전문가로 설정하는 역할 부여 프롬프트(Role), (6) 위의 모든 조건을 통합한 프롬프트(All combined)로 구분하였다. Two-shot 프롬프트의 예시로 사용된 초록은 2021년 '한국문헌정보학회' 학술지에 게재된 논문 중 KCI 피인용 횟수가 높은 상위 5편을 대상으로 하였으며, 이 중 우수하다고 평가된 2편을 연구자 3인과 협의하여 선정하였다.

3.3 유사도 측정

본 연구에서는 생성된 초록의 특성을 분석하기 위해 원초록 및 논문 본문과의 유사도(similarity)를 각각 계산하였다. 유사도는 두 텍스트 간의 내용이 얼마나 의미적으로 겹치거나 일치하는지를 수치화한 지표로 자동 생성된 초록이 저자가 작성한 원초록과 원문 내용을 얼마나 반영하고 있는지를 비교·분석하는 데 활용하였다. 이를 통해 생성초록이 원문을 충실히 요약하면서도 저자 초록과 어떤 차이를 보이는지를 살펴보고자 하였다.

본 연구에서는 유사도 계산 방법으로 TF-IDF 기반의 코사인 유사도(cosine similarity)와 BERTScore를 선택하였다. TF-IDF 기반 코사인 유사도는 텍스트를 단어 단위로 벡터화하여 두 문서의 단어 분포 유사성을 평가하는 전통적인 방법으로 정보검색 분야에서 널리 활용되어 왔다(Singhal, 2001; Salton, 1962). 이 방식은 특정 단어가 얼마나 많이 등장하고, 문서 간 공유되는지를 측정하는 데 강점이 있다. 그

러나 문맥이나 단어의 의미적 유사성까지 반영하지는 못한다는 한계를 가진다. 이러한 한계를 보완하기 위해 본 연구에서는 BERTScore를 함께 활용하였다. BERTScore는 사전 학습된 언어모델(BERT)의 컨텍스트(context) 임베딩을 기반으로 두 초록 간 의미적 일치도를 평가하며, 단어의 순서, 문맥, 의미적 유사성을 반영하는 것이 특징이다(Zhang et al., 2019). 특히 한국어와 같이 형태소 단위의 정보가 중요한 언어에서도 높은 성능을 보이는 점에서 본 연구에 적합하다고 판단하였다. 따라서 본 연구는 표면적 단어기반 유사도(TF-IDF)와 의미 기반 유사도(BERTScore)를 병행하여 활용함으로써, 생성초록의 특성을 다각도로 살펴보고자 하였다.

TF-IDF 유사도 계산을 위한 전처리에는 단어 수준의 일관성 확보를 위해 형태소 분석기(Okt)를 사용하였으며, 명사, 동사, 형용사만을 추출한 뒤 불용어 및 한 글자 이하의 단어를 제거하였다. 이때, 한글, 한자, 영어 고유명사 등은 유지하여 정보 손실을 최소화하였다. 최종적으로 전처리된 텍스트를 기반으로 TF-IDF 벡터화를 수행하고, 문서 간 코사인 유사도를 계산하였다. BERTScore의 경우에는 문맥 정보를 기반으로 의미적 유사도를 평가하므로, 별도의 전처리는 수행하지 않았다. xlm-roberta-base (Conneau et al., 2019) 모델은 다국어를 지원하는 사전학습 모델로, 입력 텍스트에 대해 내부적으로 토크나이징(tokenizing)과 정규화를 자동으로 수행한다. 본 연구에서는 해당 모델을 활용하여, 한국어로 작성된 논문 초록과 같은 비교적 정형화된 텍스트에 대해 별도의 추가 전처리 없이 실험을 진행하였다.

3.4 전문가 인식 조사

본 연구에서는 원초록과 LLM 기반 생성초록 간의 비교 분석 결과를 보다 심층적으로 이해하고자 전문가 인식 조사를 추가적으로 실시하였다. 이를 위해 문헌정보학 분야의 전문가 10인(현직 교수 5인, 학술지 게재 경험이 있는 박사과정생 5인)을 대상으로 설문조사를 수행하였다. 이들에게 원초록과 생성초록 간의 TF-IDF 코사인 유사도 기준 상위 5편과 하위 5편에 해당하는 논문 중 무작위로 선정된 2편의 논문 초록쌍(논문당 원초록과 생성초록 한 쌍씩)을 배정하여 초록의 질적 완성도와 원문 반영 정도에 대한 평가를 요청하였다. 초록 제시 순서는 편향을 방지하기 위해 무작위로 조정되었다.

특히, 평가 과정에서 충분한 맥락을 고려할 수 있도록 원초록을 제외한 후 해당 논문의 본문도 제공하였다. 평가 항목은 선행연구(Ufnalska & Hartley, 2009; Nabata et al., 2025)를 참고하여 '이해가능성', '구조', '정보의 선택', '문법・맞춤법', '간결성'으로 구성하였으며, 모든 항목은 1점(전혀 그렇지 않다)에서 5점(매우 그렇다)까지의 5점 리커트 척도로 측정하였다. 평가자들은 각 항목에 근거하여 원초록과 생성초록을 비교한 뒤, 두 초록 중 어느 초록이 AI가 생성한 것이라고 판단하는지 선택하고, 그 이유를 서술하게 하였다. 아울러 전문가(교수 5인)에게는 서면 면담지를 활용하여 추가적으로 초록 생성을 위한 생성형 AI 활용 경험과 인식에 관한 조사를 진행하였다. 면담 문항은 총 4개 주요 영역 즉, (1) ChatGPT나 Gemini 등 생성형 AI 활용 경험과 방식, (2) 초록 작성 시 AI 활용의 장점과 한계, (3) 초록 및 원고

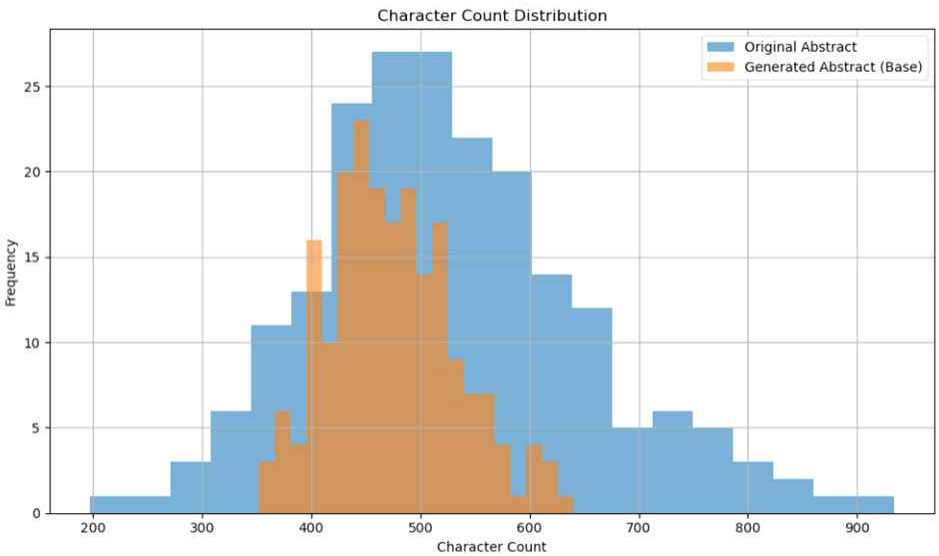
작성에서 AI 활용이 허용될 수 있는 범위, (4) AI 활용 사실의 명시 여부 및 표기 방식에 대한 견해 등으로 구성하였다.

4. 연구 결과

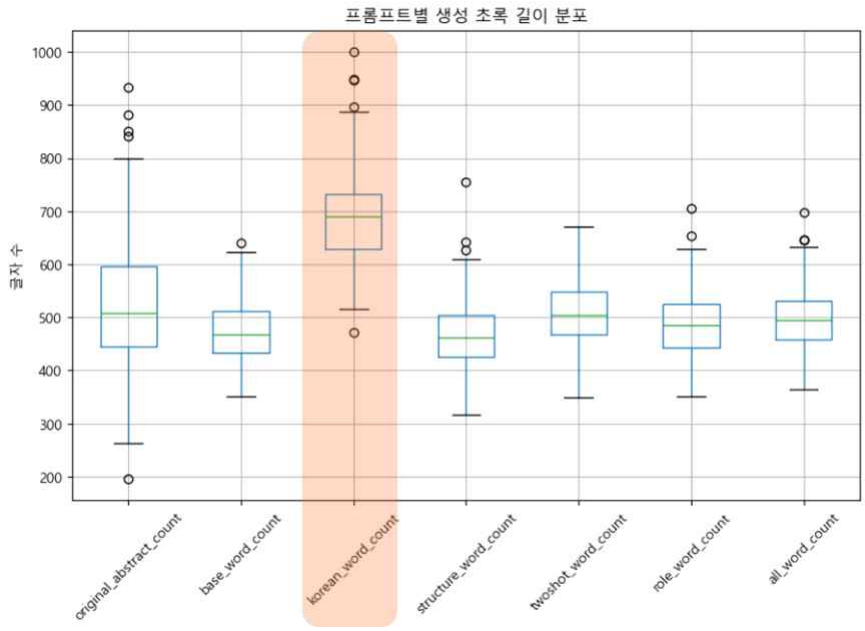
4.1 초록의 글자 수 비교

먼저 분석 대상인 204편의 논문에 대해 원초록과 생성초록 간의 글자 수 분포를 조사하였다. 〈그림 1〉에 제시된 바와 같이, 원초록의 평균 글자 수는 526자이며 대체로 400~600자 사이에 분포하였다. 다만 일부는 900자 이상 길게 작성된 경우도 있었다. 반면 Base 프롬프트로 생성된 초록의 평균 글자 수는 473자였으며, 주로 400~500자 구간에 집중되어 상대적으로 좁은 분포를 보였다. 이는 프롬프트에서 제시한 400~600자 범위의 조건이 영향을 미친 결과로 볼 수 있다. 다만 일부 사례에서는 이러한 조건을 초과하거나 그보다 짧게 생성된 경우도 확인되었다. 연도별로는 초록 글자 수 분포에서 뚜렷한 차이가 발견되지 않았다.

프롬프트 유형에 따른 글자 수를 비교한 결과, 한국어 프롬프트는 상대적으로 더 긴 초록을 생성하는 경향을 보였으며, 영어 프롬프트는 400~600자 범위를 보다 엄격히 준수하는 것으로 나타났다. 이는 한국어 프롬프트에서 사용된 "~ 이상", "~ 이내"와 같은 완곡한 표현이 모델에 의해 엄격한 조건으로 해석되지 않았을 가능성을 시사한다. 이러한 차이는 〈그림 2〉의 박스 플롯으로 확인할 수 있다. 또한 프롬프트별로 글자 수가 가장 많은 생성 초록에 해당하는 논문을 검토한 결과, 논문 간 중복은 없었으며, 초록의 길이에 영향을 줄 만한 고유명사나 특수한 표현도 발견되지 않았다.



〈그림 1〉 원초록과 생성초록(Base 프롬프트) 글자 수 분포



〈그림 2〉 프롬프트 유형별 생성초록 글자 수 분포

4.2 유사도 분석

원초록과 생성초록 간의 의미적 유사도를 나타내는 BERTScore는 모든 프롬프트 유형에서 평균 0.89 이상으로 매우 높게 나타났으며, 프롬프트 간 차이 또한 0.006 이내로 미미하였다. 이는 LLM이 생성한 초록이 전반적으로 원초록과 의미적으로 높은 일치도를 보이는 것을 나타내며, BERTScore 기준에서는 프롬프트 유형에 따른 성능 차이를 뚜렷하게 구분하기 어려운 것으로 해석된다.

반면, 〈표 2〉에서 확인할 수 있듯이, TF-IDF 유사도 값이 BERTScore보다 낮게 나타났다. 이는 TF-IDF가 단어 출현 빈도와 분포 패턴에 의존하는 통계적 방법으로, 의미적 유사성보다는 어휘적 일치 정도를 반영하기 때문이다(Singhal, 2001).

특히, 한국어 프롬프트는 TF-IDF 유사도 값이 0.6436으로, 다른 프롬프트 유형보다 다소 높게 나타나 본문 내용을 가장 충실히 반영한 것으로 해석된다. 이러한 경향은 생성초록과 본문 간의 유사도 분석 결과에서도 동일하게 나타났다. 즉, 본 연구에서는 한국어 프롬프트가 영문 프롬프트보다 원초록 및 원문과 생성초록 간의 TF-IDF 유사도에서 더 높은 평균값을 보였다. 이는 영어 프롬프트가 다국어 환경에서 상대적으로 우수한 성능을 보인다고 보고한 선행연구(Kang et al., 2023)의 결과와 상반된다. 비록 수치 차이는 크지 않았지만, 한국어 프롬프트의 유사도가 일관되게 높게 나타난 점은 주목할 만하다.

연도별 유사도 추이를 분석한 결과, 〈표 3〉과 같이, 2022년부터 2024년까지 TF-IDF 및 BERTScore 유사도가 모두 점진적으로 상승

〈표 2〉 프롬프트 유형별 원초록과 생성초록 간의 유사도 평균 비교

No	프롬프트 유형	원초록 vs. 생성초록		본문과의 비교(TF-IDF 유사도)	
		TF-IDF 유사도	BERTScore	본문 vs. 생성초록	본문 vs. 원초록
1	Base	0.6263	0.8962	0.7087	0.7146
2	Base-Korean	0.6436	0.8922	0.7412	
3	Structure	0.6377	0.8981	0.6982	
4	Few-Shot	0.6329	0.8952	0.7131	
5	Role	0.6331	0.8957	0.7158	
6	All combined	0.6371	0.8973	0.7058	

〈표 3〉 연도별 TF-IDF 및 BERTScore 기반 유사도 평균(Base Prompt 기준)

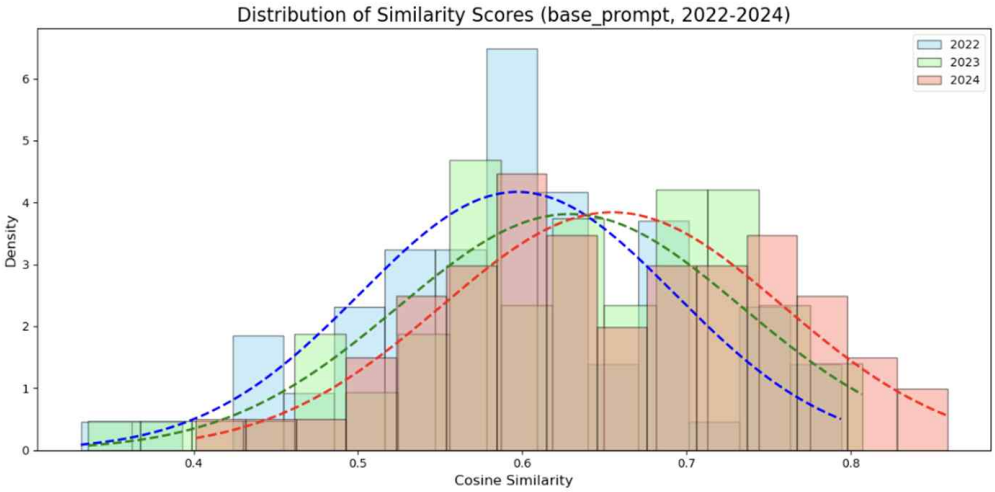
연도	원초록 vs. 생성초록		본문과의 비교(TF-IDF 유사도)	
	TF-IDF 유사도	BERTScore	본문 vs. 생성초록	본문 vs. 원초록
2022	0.597	0.893	0.690	0.697
2023	0.629	0.896	0.715	0.717
2024	**0.655**	**0.900**	**0.722**	**0.731**

하는 추세를 보였다. 특히 2024년에는 TF-IDF 유사도 평균이 0.655, BERTScore가 0.900으로 각각 가장 높은 수치를 기록하였다. 초록과 본문 간의 유사도를 비교한 결과, 모든 기간에서 원초록이 생성초록보다 소폭 더 높은 값을 보였다. 이는 연구자가 작성한 초록 역시 LLM이 생성한 초록만큼이나 본문 내용을 충실히 반영하고 있음을 의미한다. 동시에 원초록과 생성초록 모두 시간이 지남에 따라 본문과의 유사도가 점차 높아지는 경향을 보였다. 이러한 흐름은 LLM의 발전과 더불어 학술 글쓰기 과정에서 AI 활용이 확대된 결과일 가능성도 있다(Cho et al., 2023; Cheng et al., 2024).

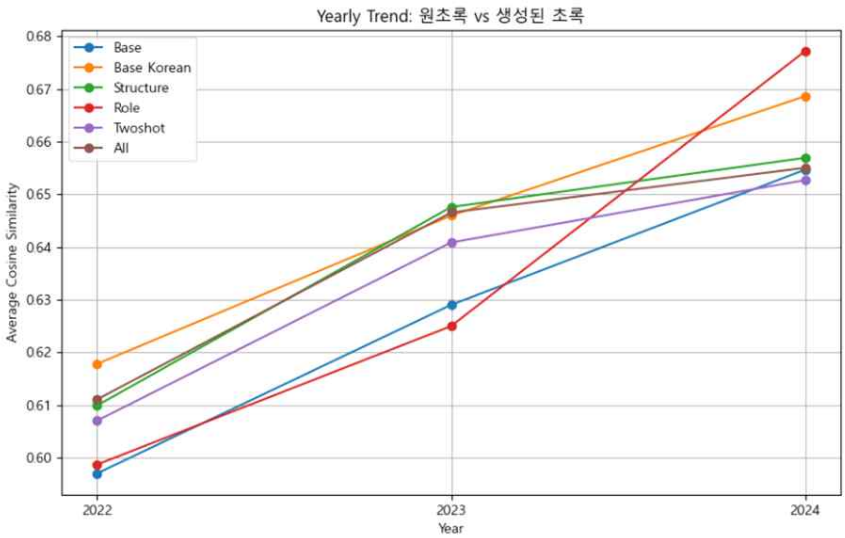
유사도 분포를 나타내는 〈그림 3〉에서도 2024년으로 갈수록 전반적으로 유사도가 오른쪽으로 이동하며 높은 유사도 구간에 집중되는 경향이 확인되었다. 또한 프롬프트 유형별 추이 〈그림 4〉를 살펴보면, 모든 프롬프트에서 유사도가 꾸준히 증가하고 있으며, 특히 역할 프롬프트가 2024년에 가장 큰 폭의 향상을 보였다.

4.3 전문가 인식 조사

먼저 참가자들이 생성초록을 구분할 수 있는지를 살펴보는 판별 정확도를 측정하였다. 그 결과, 정확도는 60%로, 총 20건 중 12건의 초록쌍에서 AI가 생성한 초록을 올바르게 구분한 것으로 나타났다. 이는 전반적으로 AI가 작성한 초록을 식별하기가 쉽지 않음을 보여준다. 참가자 유형별로 살펴보면, 박사과정생의 판별 정확도는 50%(10건 중 5건), 교수진의 판별 정확도는 70%(10건 중 7건)으로 나타났다. 즉 연구 경험이 풍부한 교수진이 비교적 더 높은 정확도로 생성초록을 식별하였으나, 두 집단 모두 완벽하게 판별하지는 못하였다. 또한 TF-IDF 유사도 상위 논문과 하위 논문 간 정확도에는



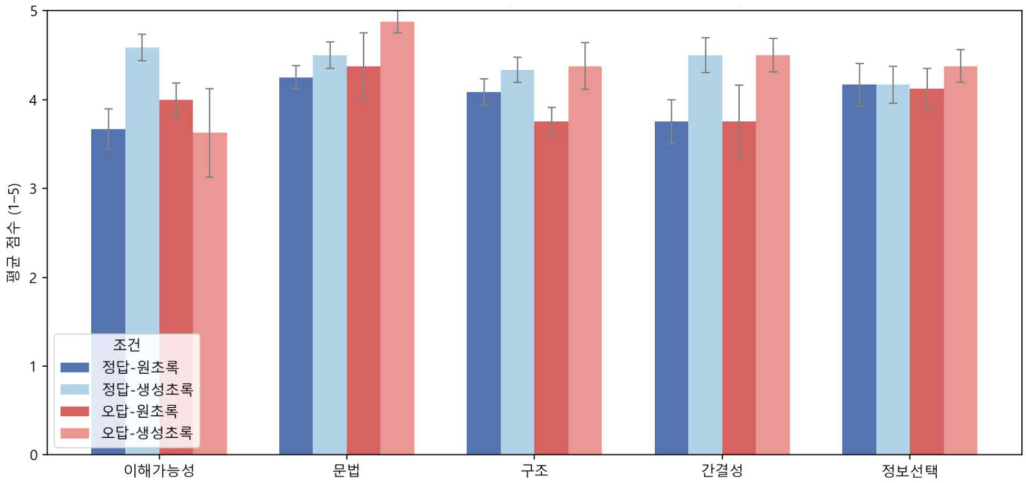
〈그림 3〉 연도별 원초록과 생성초록 간 유사도 분포



〈그림 4〉 프롬프트 유형에 따른 원초록과 생성초록 간 유사도 연도별 추이

차이가 나타나지 않았다.

생성초록을 올바르게 판별한 참가자들은 AI가 생성한 초록에 원초록보다 전반적으로 더 높은 점수를 부여하는 경향이 있었다(〈그림 5〉 참고). 특히, 이해가능성(M=4.58), 문법(M=4.50), 간결성(M=4.50) 항목에서 생성초록의 평균 점수가 원초록(각각 M=3.67, M=4.25, M=3.75)보다 더 높은 것으로 나타났다. 이는 생성초록이 문체적으로 명료하고 문법적으로 정제되어 있다는 인상을 주었음을 알 수 있다. 반면, 정보의 선택이나 구조적 측면에서는 두 초록 간 큰 차이가 발견되지 않았다. 이에 반해 생



〈그림 5〉 정답자 및 오답자 집단의 평가 항목별 평균 점수 비교

성초록을 올바르게 구분하지 못한 참가자들의 경우, 이해가능성 항목에서 원초록(생성초록으로 오인한 경우, M=4.00)이 생성초록(원초록으로 오인한 경우, M=3.62)보다 높게 평가되었다. 그러나 문법(M=4.88), 구조(M=4.38), 간결성(M=4.50)에서는 정답자 그룹과 마찬가지로 생성초록(원초록으로 오인한 경우)에 더 높은 점수를 부여하였다. 이는 두 그룹 모두 생성초록의 문장을 문법적으로 매끄럽고 정제된 표현으로 인식했음을 보여준다.

참가자들에게 자신이 선택한 초록을 AI 생성초록으로 판단한 이유를 서술하도록 한 결과, 생성초록을 정확히 구분한 참가자들은 문체적 완결성과 표현의 일반화 정도를 주요 단서로 언급하였다. 이들은 AI가 생성한 초록이 "목적-현황-문제-해결-기대효과"와 같은 정형화된 구조를 따르고, 문장이 지나치게 매끄럽고 간결하며 포괄적 표현을 사용한다는 점을 근거로 들었다. 반면, 저자 초록은 구체적인 연구 범위나 방법, 복잡하거나 덜 정제된 문장 구조에서 사람의 서술 흔적이 드러난다고 언급하였다. 다시 말해, 참가자들은 AI 초록의 과도한 정제성과 인간 초록의 구체성과 불균질성을 대조적으로 인식하며 이를 판별의 기준으로 삼는 것으로 나타났다.

반면 생성초록을 올바르게 구분하지 못한 참가자들은 AI와 저자 초록을 명확한 기준없이 판단하는 경향을 보였다. 일부는 문장이 매끄럽지 않거나 형식이 어색한 초록을 AI가 작성한 것으로 보았고, 다른 일부는 반대로 간결하고 구조화된 초록을 기계적으로 요약된 글로 인식하였다. 또한 내용의 구체성이나 분량, 연구 결과의 서술 정도를 근거로 삼았지만, 그 판단 기준은 참가자마다 상이하였다. 이러한 응답은 참가자들이 문체적 완성도나 서술 방식만으로는 AI 초록을 일관되게 구분하기 어려웠음을 보여준다.

추가적으로 전문가 인식조사를 보완하기 위해 참가자 5명을 대상으로 후속 서면 면담을 실시하였다. 먼저 초록 작성 과정에서 AI를 활용

한 경험을 묻는 질문에 대해 대부분의 참여자는 초록을 직접 작성한 뒤 문법 오류를 점검하거나 어휘를 다듬는 수준에서 주로 AI를 활용했다고 응답하였다. 일부는 글자 수를 조정하거나 언어 표현을 다듬고 번역을 돕는 기능을 통해 시간을 절약하고 초록의 완성도를 높일 수 있었다고 평가하였다. 특히 AI와의 상호적인 피드백 과정을 통해 문장을 점진적으로 수정·보완하면서 보다 자연스럽고 일관된 표현을 완성할 수 있었다는 점도 긍정적으로 언급되었다. 한편, 일부 참가자는 AI를 활용한 경험이 없다고 응답했으며, 그 이유로는 AI가 생성한 글이 모호하거나 구체적인 표현이 부족해 신뢰하기 어렵다고 판단했기 때문이라고 답했다. 즉 이들은 생성형 AI를 논문 초록을 자동으로 작성하는 도구라기보다 이미 작성된 초록의 문법과 표현을 보완하는 보조의 역할로 인식하는 경향이 있었다.

한편, 생성형 AI가 사용하는 표현이 학문적 문체와 완전히 일치하지 않아 연구자의 의도나 분야 특유의 어휘가 충분히 반영되지 못한다는 한계도 지적되었다. 특히 AI가 논문 전반의 내용을 균등하게 다루려는 경향이 있어, 저자가 강조하고자 하는 핵심 결과나 기여점을 부각하기 어렵다는 의견이 제시되었다. 또한 생성초록이 지나치게 서술적으로 장황해지는 경우가 많으며, 학계에서 일반적으로 사용되지 않는 어휘를 사용하여 오히려 추가 수정이 필요하다는 지적도 있었다. AI가 제안하는 문장들은 대체로 의미상 오류는 없으나 표현이 추상적이고 세부적인 분석이 결여되어 있다는 점이 공통적으로 언급되었다.

끝으로 전문가들은 AI의 효율성과 편의성을 인정하면서도, 연구자가 작성하는 글에 대해 전적인 책임을 지고 AI의 사용 목적과 범위를 명확히 밝히는 것이 필요하다는 점에 공감하였다. 이러한 인식은 초록 작성에서 AI가 '생성자'보다는 '언어적 보조 도구'로 제한적으로 수용되고 있음을 보여준다.

5. 논의 및 결론

본 연구는 학술논문 초록 작성에 대한 LLM의 적용 가능성을 탐색적으로 분석하였다. ChatGPT의 공개 이후 전 세계 연구자들은 학술 글쓰기의 다양한 단계에서 LLM을 적극적으로 활용하고 있다(Cheng et al., 2024). 이러한 변화는 생성형 AI가 더 이상 단순한 보조 도구가 아니라, 글쓰기 과정의 실질적 협력자로 자리매김하고 있음을 보여준다. 이에 본 연구는 학술논문의 핵심 내용을 요약하는 초록 작성 단계에 AI 활용 가능성을 조사하였다. 이를 위해 LLM이 생성한 초록과 저자가 작성한 원초록을 대상으로 초록 간 및 본문과의 유사도와 프롬프트 유형별 차이를 정량적으로 분석하였다. 또한 전문가 평가와 후속 면담조사를 통해 LLM 활용에 대한 인식과 실제 활용 양상을 종합적으로 탐색하였다. 이처럼 정량적 분석과 정성적 평가를 병행하여 LLM이 학술 초록 작성 과정에서 수행할 수 있는 역할과 한계를 규명하고, 학술 글쓰기의 보조 도구로서 그 적용 가능성을 다각적으로 검토하였다.

연구 결과, BERT 기반 분석에서는 원초록과 생성초록 간 유사도가 전반적으로 높게 나타난 반면, TF-IDF 기준에서는 상대적으로 낮

은 수치를 보였다. 이는 두 초록이 의미적으로는 유사하더라도, 생성초록이 표현이나 어휘 선택에서 원초록과 다소 다른 방식으로 기술되었음을 시사한다. 특히 한국어 프롬프트를 활용한 경우, 생성초록이 원초록 및 본문과의 유사도가 소폭 높게 나타나 프롬프트 언어와 원문의 언어가 일치할 때 유사도가 다소 높게 나타나는 경향을 보였다.

이러한 결과는 프롬프트 언어가 생성되는 출력의 언어와 요약 방식에 일정한 영향을 줄 수 있다는 기존 연구들과도 일치함을 보인다. 대표적으로 Fu et al.(2022)은 LLM이 프롬프트 언어를 출력 언어 및 태스크 수행 방식의 힌트로 해석하는 경향이 있음을 보였으며, 영어로 제시된 프롬프트의 경우 명시적인 번역 지시가 없어도 입력 문서를 영어로 번역하거나, 영어식 문체와 표현 방식으로 결과를 생성하는 경향이 있음을 보고하였다. 이에 더해, Razumovskaia et al.(2022)은 교차언어 이야기 생성(crosslingual story generation) 실험을 통해 프롬프트의 언어와 서술 구조가 생성 결과의 자연스러움과 일관성에 영향을 미친다는 점을 확인하였다. 이러한 선행연구들을 고려할 때, 본 연구에서 한국어 프롬프트가 영어 프롬프트보다 더 높은 유사도 점수를 나타낸 결과는 프롬프트의 언어가 모델의 이해와 표현 방식에 영향을 미친 결과로 볼 수 있다. 따라서 후속 연구에서는 프롬프트 언어와 본문 언어의 일치 여부가 초록 생성에 미치는 영향을 보다 체계적으로 분석할 필요가 있다. 또한 한국어 프롬프트의 효과를 확인하기 위하여 한-영 혼합 프롬프트나 다양한 모델 간 비교 실험을 추가로 수행하는 것도 의미 있을 것이다.

전문가 인식조사 결과, 참가자들은 전반적으로 AI가 생성한 초록과 연구자가 작성한 원초록을 명확히 구분하기 어려워했으며, 두 집단 모두 완전히 구분하지는 못했다. 이들 대부분은 LLM이 생성한 초록은 문법적으로 정확하고 문장이 매끄러워 읽기에는 자연스럽지만, 세부 내용의 구체성과 연구 의도의 강조는 부족하다고 인식하였다. 한편 Yurchenko와 Nalyvaiko(2025)가 지적하듯, 연구 과제 자체가 AI 탐지를 전제로 이루어질 경우, 참가자들은 실제 텍스트의 차이보다 'AI적 특징'을 찾아내려는 인식적 편향(cognitive bias)의 영향을 받을 수 있다. 이러한 점에서 본 연구의 참가자들 또한 문체적 단서에 과도하게 의존하거나, AI가 작성했을 가능성을 지나치게 추론하려는 경향을 보였을 가능성이 있다.

추가 면담조사 결과, 이러한 판단 경향은 전문가들이 AI의 글쓰기 능력을 일정 부분 인정하면서도 그 활용 범위를 신중하게 바라보는 태도와 맞닿아 있었다. 본 연구에 참여한 전문가들은 AI가 글을 대신 써주는 도구가 아니라 연구자가 주도적으로 통제하고 책임을 지는 범위 내에서만 보조적으로 활용되어야 한다는 인식이 우세했다. 일부 전문가들은 AI가 생성한 문장이 두리뭉실하거나 구체적이지 않다고 지적하며, 학문적 문체와 분석적 깊이의 부족을 한계로 언급하였다. 동시에 작성된 글에 대한 저자의 책임과 AI 사용에 대한 투명한 기술의 필요성에 대한 의견이 공통적으로 제시되어 AI 활용의 윤리성과 저자로서의 책임의 중요성이 강조되었다.

본 연구는 LLM을 활용하여 학술논문 초록 작성에 대한 적용 가능성을 탐색했다는 점에서

학문적·실천적 의의를 지닌다. 먼저, 학문적 시사점으로는 LLM이 생성한 초록이 연구의 핵심 내용을 얼마나 정확하고 충실하게 반영하는지를 정량적 및 정성적 방법을 결합하여 분석하였다. 이를 통해 학술 글쓰기 과정에서 LLM의 잠재력과 한계를 균형 있게 이해할 수 있는 근거를 제시하였다. 다음으로 실천적 시사점으로는 연구자가 초록을 작성할 때 AI가 생성한 초록을 참고 자료로 활용하여 누락되거나 과소평가된 부분을 보완함으로써, 인간의 비판적 사고와 AI의 언어적 정확성이 결합된 협업적 초록 작성의 가능성을 제시한다는 데 있다. 아울러 학회나 학술지 차원에서도 초록 작성 단계에서의 AI 활용 원칙과 사사(acknowledgement) 표기 방식 등 명확한 가이드라인을 마련할 필요성을 제기하였다.

나아가 본 연구의 결과는 LLM이 단순한 초록 생성 보조 도구를 넘어 학술 커뮤니케이션의 여러 단계에서 실질적 활용 가능성을 지님을 시사한다. 예를 들어, 초록 작성 교육 및 훈련 도구로서 AI가 생성한 초록과 전문가 초록의 차이를 비교함으로써 학습 자료로 활용하거나, 초록 품질 평가 및 피드백 시스템으로서 LLM을 1차 검토 도구로 활용하여, 초록 내 필수 요소 즉, 목적, 방법, 결과, 결론의 충실도를 점검하고 투고 전 품질 자가 진단을 지원할 수 있다. 더 나아가 언어적·문화적 맥락이 다른 학문 공동체 간의 격차를 완화하여 비영어권 연구자의 언어 장벽 완화에 기여할 수 있을 것이다. 이러한 논의는 향후 LLM을 초록 작성의 보조자이자 학술 소통의 매개체로 활용하기 위한 구체적 프레임워크로 발전될 수 있을 것이다.

본 연구는 일부 한계도 지닌다. 우선, 분석 대상이 2022년부터 2024년까지의 '한국문헌정보학회지' 논문으로 한정되어 있어서, 결과를 다른 학문 분야에 일반화하기에는 한계가 있다. 따라서 후속 연구에서는 사회과학, 자연과학, 공학 등 다양한 분야로 대상을 확대하여 LLM의 초록 생성 및 활용 가능성을 폭넓게 검증할 필요가 있다. 또한 이 연구에서는 전문가 10인의 인식을 조사하였는데, 이러한 소규모 전문가 평가가 여러 선행연구(강준영 외, 2024; 윤세휘, 신유현, 2024)에서도 공통적으로 나타나는 경향이지만 그 결과를 일반화하기에는 무리가 있다. 따라서 향후 연구에서는 이러한 한계를 보완하기 위해 실제 연구자와 대학원생을 포함한 폭넓은 설문조사를 병행하여, LLM 활용에 대한 인식과 수용 정도를 보다 체계적으로 파악할 예정이다. 이러한 후속 연구를 통해 분야와 언어의 차이를 고려한 LLM의 효과적인 활용 방식을 구체화하고, 보다 수준 높은 초록 작성의 방향을 제시할 수 있을 것이다.

참 고 문 헌

강준영, 김성은, 김자경, 이은송, 오동석 (2024). 대규모 언어모델을 활용한 지식기반 대화 환각 현상 자동 평가. 언어와 인간 기술, 566-569.

고은정, 김남규 (2018). 완전성과 간결성을 고려한 텍스트 요약 품질의 자동 평가기법. 지능정보연구, 24(2), 125-148. https://doi.org/10.13088/jiis.2018.24.2.125

김의순, 임희석 (2021). 사전학습 기반의 법률문서 요약 방법 비교연구. 한국정보처리학회 학술대회, 2021, 614-617. https://doi.org/10.3745/PKIPS.Y2021M11A.614

박자현, 송민 (2013). 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석. 정보관리학회지, 30(1), 7-32. https://doi.org/10.3743/KOSIM.2013.30.1.007

박재언, 김지호, 이홍철 (2022). BERT 기반의 사전 학습 언어 모형을 이용한 한국어 문서 추출 요약 베이스라인 설계. 한국정보기술학회논문지, 20(6), 19-32. https://doi.org/10.14801/jkiit.2022.20.6.19

서선경, 정은경 (2013). 동시출현단어 분석 기반 오픈 액세스 분야 지적구조에 관한 연구. 한국비블리아학회지, 24(1), 207-228.

송의석, 김남규 (2021). 사전학습 언어 모델을 활용한 트랜스포머 기반 텍스트 요약. 경영과 정보연구, 40(4), 31-47. https://doi.org/10.29214/damis.2021.40.4.002

신주은, 김성희 (2021). 국내 오픈액세스 분야의 지적구조 분석에 관한 연구. 한국문헌정보학회지, 55(2), 147-178. https://doi.org/10.4275/KSLIS.2021.55.2.147

양종훈, 곽일엽 (2021). 초록데이터를 활용한 국내외 통계학 분야 연구동향. 응용통계연구, 34(2), 267-278.

윤세휘, 신유현 (2024). 한국어 생성 요약 성능 평가 지표 분석 연구. 정보처리학회 논문지, 13(12), 691-699. https://doi.org/10.3745/TKIPS.2024.13.12.691

윤수환, 김아영, 박성배 (2021). 주제 어트리뷰트 모델을 이용한 주제 키워드 기반 한국어 문서 요약. 정보과학회논문지, 48(6), 688-695. https://doi.org/10.5626/JOK.2021.48.6.688

이영재, 이홍철, 김지호 (2025). 대형 언어모델을 활용한 특허 청구항 생성요약에 관한 연구. 한국지능시스템학회 논문지, 35(3), 205-211.

이재윤 (2024). 국내 문헌정보학 분야 학술지에 구조적 초록을 도입하기 위한 예비 연구. 한국문헌정보학회지, 58(2), 121-150. https://doi.org/10.4275/KSLIS.2024.58.2.121

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., Bali, K., & Sitaram, S. (2023). Mega: multilingual evaluation of generative ai. arXiv preprint arXiv:2303.12528. https://doi.org/10.48550/arxiv.2303.12528

Ali, S., Szmuda, T., & Wszołek, Z. (2020). How to write a scientific paper? lessons from a distinguished scientist and editor. European Journal of Translational and Clinical Medicine, 3(1), 74-78. https://doi.org/10.31373/ejtcm/118954

American Chemical Society (2024). ACS meeting abstract submission guidelines. Organic Division of ACS. https://www.organicdivision.org/meetingsupport/acsmeetings/abstractsubmissio

nguidelines/

Cheng, H. Z., Sheng, B., Lee, A., Chaudhary, V., Atanasov, A. G., Liu, N., Qiu, Y., Wong, T. Y., Tham, Y. C., & Zheng, Y. F. (2024). Have AI-generated texts from LLM infiltrated the realm of scientific writing? A large-scale analysis of preprint platforms. bioRxiv, 2024-03. https://doi.org/10.1101/2024.03.25.586710

Cho, W. I., Cho, E., & Cho, K. (2023). PaperCard for reporting machine assistance in academic writing. arXiv preprint arXiv:2310.04824. https://doi.org/10.48550/arxiv.2310.04824

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116. https://doi.org/10.48550/arxiv.1911.02116

Delving, E., Pillay, T. S., & Newman, A. (2014). How to write a scientific paper: practical guidelines. Ejifcc, 25(3), 259.

Dey, K., Tarannum, P., Hasan, M. A., Razzak, I., & Naseem, U. (2024). Better to Ask in English: evaluation of large language models on english, low-resource and cross-lingual settings. arXiv preprint arXiv:2410.13153. https://doi.org/10.48550/arxiv.2410.13153

Fu, J., Ng, S. K., & Liu, P. (2022). Polyglot prompt: multilingual multitask promptraining. arXiv preprint arXiv:2204.14264. https://doi.org/10.48550/arxiv.2204.14264

Gong, Y. & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. Association for Computing Machinery, 19-25. https://doi.org/10.1145/383952.383955

Hartley, J. (2004). Current findings from research on structured abstracts. Journal of the Medical Library Association, 92(3), 368-371.

Kang, H., Blevins, T., & Zettlemoyer, L. (2023). Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models. arXiv preprint arXiv:2304.13803. https://doi.org/10.48550/arxiv.2304.13803

Ketcham, C. M., Hardy, R. W., Rubin, B., & Siegal, G. P. (2010). What editors want in an abstract. Laboratory Investigation, 90(1), 4-5. https://doi.org/10.1038/labinvest.2009.122

Klimova, B. F. (2015). Teaching English abstract writing effectively. Procedia-Social and Behavioral Sciences, 186, 908-912. https://doi.org/10.1016/j.sbspro.2015.04.113

Klimova, B. F. (2020). An off-line scaffolding tool for writing abstracts of qualification papers. Procedia Computer Science, 176, 1271-1278. https://doi.org/10.1016/j.procs.2020.09.136

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H.

(2023). Chatgpt beyond english: towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:2304.05613. https://doi.org/10.18653/v1/2023.findings-emnlp.878

Lancaster, F. W. (2003). Indexing and Abstracting in Theory and Practice (3rd ed.). London, UK: Facet Publishing.

Leidinger, A., Van Rooij, R., & Shutova, E. (2023). The language of prompting: what linguistic properties make a prompt successful?. arXiv preprint arXiv:2311.01967. https://doi.org/10.48550/arxiv.2311.01967

Liu, Y., Fabbri, A.R., Liu, P., Zhao, Y., Nan, L., Han, R., Han, S., Joty, S. R., Wu, C., Xiong, C., & Radev, D. R. (2022). Revisiting the gold standard: grounding summarization evaluation with robust human evaluation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 4140-4170. https://doi.org/10.48550/arxiv.2212.07981

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics, 23(6), bbac409. https://doi.org/10.1093/bib/bbac409

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2024). Prompt engineering in large language models. In: Jacob, I. J., Piramuthu, S., Falkowski-Gilski, P. eds. Data Intelligence and Cognitive Informatics: ICDICI 2023. Algorithms for Intelligent Systems. Singapore: Springer. https://doi.org/10.1007/978-981-99-7962-2_30

Mihalcea, R. & Tarau, P. (2004). Textrank: binging order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 404-411.

Nabata, K. J., Alshehri, Y., Mashat, A., & Wiseman, S. M. (2025). Evaluating human ability to distinguish between ChatGPT-generated and original scientific abstracts. Updates in Surgery, 1-7. https://doi.org/10.1007/s13304-025-02106-3

Razumovskaia, E., Maynez, J., Louis, A., Lapata, M., & Narayan, S. (2022). Little red riding hood goes around the globe: crosslingual story planning and generation with large language models. arXiv preprint arXiv:2212.10471. https://doi.org/10.48550/arXiv.2212.10471

Salton, G. (1962). Some experiments in the generation of word and document associations. In Proceedings of Fall Joint Computer Conference (AFIPS '62 (Fall)), 234-250. https://doi.org/10.1145/1461518.1461544

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., Goncearenco, I., Sarli, G., Galynker, 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 arXiv:2406.06608. https://doi.org/10.48550/arXiv.2406.06608

Singhal, A. (2001). Modern information retrieval: a brief overview. IEEE Data Eng. Bull., 24(4), 35-43.

Ufnalska, S. & Hartley, J. (2009). How can we evaluate the quality of abstracts. European Science Editing, 35(3), 69-72.

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., & Affandy, A. (2022). Review of automatic text summarization techniques & methods. Journal of King Saud University-Computer and Information Sciences, 34(4), 1029-1046. https://doi.org/10.1016/j.jksuci.2020.05.006

Yurchenko, V. & Nalyvaiko, O. (2025). How ChatGPT shapes a new reality of writing: Is there a place for humans in an artificial world? Educational Challenges, 30(1), 138-155. https://doi.org/10.34142/2709-7986.2025.30.1.09

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: evaluating text generation with bert. arXiv preprint arXiv:1904.09675. https://doi.org/10.48550/arxiv.1904.09675

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. (2023). Benchmarking large language models for news summarization. Transactions of the Association for Computational Linguistics, 12, 39-57. https://doi.org/10.1162/tacl_a_00632

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J. Y., & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223, 1(2).

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

Go, Eunjung & Kim, Namgyu (2018). Automatic quality evaluation with completeness and succinctness for text summarization. Journal of Intelligence and Information Systems, 24(2), 125-148. https://doi.org/10.13088/jiis.2018.24.2.125

Kang, Junyoung, Kim, Sungeun, Kim, Jakyoung, Lee, Eunsong, & Oh, Dongsuk (2024). Automatic

evaluation of hallucination effects in knowledge-based dialogue using large language model. in annual conference on human and language technology. Human and Language Technology, 566-569.

Kim, Euisoon & Lim, Heuiseok (2021). Comparative study of legal document summary method based on pre-trained model. Annual Conference of KIPS, 614-617. https://doi.org/10.3745/PKIPS.Y2021M11A.614

Lee, Jae Yun (2024). A preliminary study on the adoption of structured abstracts by korean library and information science journals. Journal of the Korean Society for Library and Information Science, 58(2), 121-150. https://doi.org/10.4275/KSLIS.2024.58.2.121

Lee, Young Jae, Lee, Hongchul, & Kim, Jiho (2025). A study on patent claim abstractive summarization using large language model (LLM). Journal of Korean Institute of Intelligent Systems, 35(3), 205-211.

Park, Jae Eon, Kim, Jiho, & Lee, Hongchul (2022). Designing baseline for korean document summarization using BERT-based pre-trained encoder. Journal of Korean Institute of Information Technology, 20(6), 19-32. https://doi.org/10.14801/jkiit.2022.20.6.19

Park, Jahyun & Song, Min (2013). A study on the research trends in library & information science in Korea using topic modeling. Journal of the Korean Society for Information Management, 30(1), 7-32. https://doi.org/10.3743/KOSIM.2013.30.1.007

Seo, Sun Kyung & Chung, Eunkyung (2013). Domain analysis on the field of open access by co-word analysis. Journal of the Korean Biblia Society for Library and Information Science, 24(1), 207-228.

Shin, Jueun & Kim, Seonghee (2021). A study on the intellectual structure of domestic open access area. Journal of the Korean Society for Library and Information Science, 55(2), 147-178. https://doi.org/10.4275/KSLIS.2021.55.2.147

Song, Euiseok & Kim, Namgyu (2021). Transformer-based text summarization using pre-trained language model. Management & Information Systems Review, 40(4), 31-47. https://doi.org/10.29214/damis.2021.40.4.002

Yang, Jong-Hoon & Kwak, Il-Youp (2021). Research trends in statistics for domestic and international journal using paper abstract data. The Korean Journal of applied Statistics, 34(2), 267-278.

Yoon, Sehwi & Shin, Youhyun (2024). A study on automatic metrics for korean text abstractive summarization. The Transactions of the Korea Information Processing Society, 13(12), 691-699. https://doi.org/10.3745/TKIPS.2024.13.12.691

Yoon, Su-Hwan, Kim, A-Yeong, & Park, Seong-Bae (2021). Topic centric korean text summarization using attribute model. Journal of KIISE, 48(6), 688-695. https://doi.org/10.5626/JOK.2021.48.6.688