Journal of Korean Library and Information Science Society
Vol. 57, No. 1

문헌정보학 연구에서의 표집 방법론에 대한 대규모 언어모델 기반 내용 분석

- 과업 유형에 따른 모델 간 코딩 수행 비교 -

LLM-Based Content Analysis of Sampling Methodology in Library and Information Science Research: A Cross-Model Comparison of Coding Performance by Task Type

민 세 인(Sein Min)*
김 은 기(Eungi Kim)**

〈 목 차 〉

Ⅰ. 서 론
Ⅱ. 선행연구
Ⅲ. 연구 방법
Ⅳ. 연구 결과
Ⅴ. 논의 및 한계
Ⅵ. 결 론

요 약 : 본 연구의 목적은 문헌정보학 연구 방법 분석의 맥락에서 과업 유형에 따른 대규모 언어모델(LLM) 기반 내용 분석의 적용 가능 조건을 차원별로 비교·검토하는 데 있다. 이를 위해 2020년부터 2024년까지 국내 4대 문헌정보학 학회지에 게재된 설문 및 인터뷰 연구 100편을 층화무작위표집 방식으로 선정하고, 표집 방법론을 구성하는 12개 차원에 대해 인간 코더 1인과 4개 대규모 언어모델(Claude-3.5-Haiku, GPT-4o-Mini, Gemini-2.0-Flash, Grok-4-Latest)의 코딩 결과를 비교하였다. 분석 결과, 명시적 기준에 따라 분류가 가능한 차원에서는 상대적으로 높은 일치도가 나타난 반면, 추론적·평가적 판단을 요구하는 차원에서는 일관되게 낮은 수준의 일치도가 확인되었다. 이러한 결과는 LLM 기반 자동화 코딩의 성과가 모델 성능 자체보다는 과업의 판단 구조와 정보의 명시성에 더 크게 영향을 받음을 시사한다. 따라서 LLM의 활용 범위는 과업 유형 및 판단 특성 차원에서 보다 정교하게 검토될 필요가 있으며, 인간-AI 혼합 검증 전략의 체계적 설계가 요구된다.

주제어 : 대규모 언어모델, 자동화 내용 분석, 코딩 신뢰도, 문헌정보학 연구방법론, 표집 방법론

ABSTRACT : The purpose of this study is to compare and examine, across multiple dimensions, the conditions under which large language model (LLM)-based content analysis can be applied according to task type in the context of research methods analysis in library and information science. To this end, 100 survey and interview studies published between 2020 and 2024 in four major Korean journals in library and information science were selected using stratified random sampling. The coding results produced by one human coder and four large language models (Claude-3.5-Haiku, GPT-4o-Mini, Gemini-2.0-Flash, and Grok-4-Latest) were compared across twelve dimensions constituting sampling methodology. The results show that relatively high levels of agreement were observed in dimensions where classification could be made based on explicit criteria, whereas consistently lower levels of agreement appeared in dimensions requiring inferential or evaluative judgment. These findings suggest that the performance of LLM-based automated coding is influenced more by the decision structure of the task and the explicitness of the available information than by model performance itself. Therefore, the scope of LLM application should be more carefully examined from the perspectives of task type and judgment characteristics, and the systematic design of human-AI hybrid validation strategies is required.

KEYWORDS : Large Language Models, Automated Content Analysis, Coding Reliability, Library and Information Science Research Methods, Sampling Methodology

* 계명대학교 문헌정보학과 박사과정(seinmin@kmu.kr / ISNI 0000 0005 3015 1763) (제1저자)
** 계명대학교 문헌정보학과 교수(egkim@kmu.ac.kr / ISNI 0000 0004 9410 6454) (공동저자)

• 논문접수: 2026년 2월 21일 • 최초심사: 2026년 3월 7일 • 게재확정: 2026년 3월 22일
• 한국도서관·정보학회지, 57(1), 413-438, 2026. http://dx.doi.org/10.16981/kliss.57.1.202603.413

Ⅰ. 서 론

경험적 연구에서 표집 방법론은 연구 결과의 해석 범위와 일반화 가능성을 규정하는 핵심 요소이다. 표집 방법, 표집 프레임, 표본 크기, 응답률과 같은 설계 요소는 연구의 외적 타당성과 재현 가능성에 직접적인 영향을 미치며(Cochran, 1977; Groves et al., 2009), 분석 결과가 모집단에 대해 어느 수준까지 확장될 수 있는지를 구조적으로 제한한다. 그럼에도 불구하고 실제 학술 논문에서는 표집 절차와 관련된 정보가 불충분하거나 개념적으로 모호하게 기술되는 경우가 적지 않으며, 이는 연구 결과의 비교 가능성과 지식 축적의 안정성을 저해하는 요인으로 작용한다. 이러한 맥락에서 표집 방법론에 대한 체계적 코딩과 비교 분석은 개별 연구의 질 평가를 넘어, 학문 분야 전반의 연구 관행을 진단하는 방법론적 장치로 기능할 수 있다.

최근 자동화된 텍스트 분석 기법의 발전은 이러한 진단을 대규모 문서 집합 수준에서 수행할 수 있는 환경을 조성하고 있다. 특히 대규모 언어모델(LLM)은 소수 예시 학습만으로도 다양한 분류 과업에서 높은 일반화 성능을 보이는 것으로 보고되어 왔으며(Brown et al., 2020), 일부 연구에서는 사회과학적 텍스트 주석 과제에서 인간 주석자에 근접한 성과를 보일 수 있음이 제시되었다(Gilardi et al., 2023). 이러한 기술적 진전은 학술 논문에 포함된 방법론적 정보를 자동으로 추출·분류할 가능성을 실질적으로 확대하고 있다.

그러나 기존 연구는 주로 예측 정확도나 F1-score와 같은 성능 지표를 중심으로 모델을 평가해 왔으며, 학술 연구 맥락에서 요구되는 내용 분석의 신뢰도와 판단 구조의 안정성을 체계적으로 검토한 연구는 제한적이다. 내용 분석 전통에서 신뢰도는 단순한 일치율이 아니라 범주 체계와 코딩 규칙의 구조적 안정성을 반영하는 개념으로 이해되어 왔고(Krippendorff, 2018), 해석적 판단이 개입되는 차원에서는 판단의 투명성과 일관성이 핵심 기준으로 강조되어 왔다(Lincoln & Guba, 1985; Tracy, 2010). 특히 선행연구는 단일 모델 평가나 단순 분류 과제에 국한되는 경향이 있어, 명시적 정보 추출(텍스트에 직접 기술된 정보를 범주에 귀속하는 과업), 추론적 판단(명시되지 않은 정보를 맥락으로부터 유추하는 과업), 평가적 판단(연구 설계의 질이나 적절성에 대한 가치 판단을 요구하는 과업)과 같이 과업의 판단 요구 수준에 따라 신뢰도가 어떻게 달라지는지를 체계적으로 비교한 연구는 부족하다. 따라서 LLM 기반 자동화 코딩의 타당성을 평가하기 위해서는 개별 모델의 예측 성능을 비교하는 수준을 넘어, 코딩 과업이 요구하는 판단 구조와 정보의 명시성 수준에 따라 신뢰도 패턴이 어떻게 달라지는지를 경험적으로 분석할 필요가 있다.

본 연구의 목적은 문헌정보학 연구 논문에 보고된 표집 방법론을 대상으로 LLM 기반 내용 분석을 적용하여, 표집 정보의 특성이 코딩 과업의 판단 요구 수준과 결합하는 방식, 그리고 그 결과가 차원별 신뢰도 변동으로 어떻게 나타나는지를 체계적으로 비교·분석하는 데 있다.

이를 위해 본 연구는 첫째, 문헌정보학 연구 논문의 표집 방법론 코딩이라는 구체적 사례를 통해 LLM 기반 내용 분석의 수행 특성을 비교하고, 둘째, 명시적 정보 추출, 추론적 판단, 평가적 판단이라는 과업 유형에 따라 자동화 코딩 결과가 어떻게 차별적으로 나타나는지를 검토하며, 특히 인간과 LLM 간 판단 차이의 패턴을 차원별 κ 분석 및 질적 사례 분석을 통해 탐색적으로 분석한다. 이를 위해 2020년부터 2024년까지 국내 문헌정보학 주요 4개 학회지에 게재된 설문 및 인터뷰 연구 100편을 층화무작위표집 방식으로 선정하고, 연구 설계 유형, 표집 구조, 표집 보고의 충실성, 평가 및 성찰을 포함한 12개 차원에 대해 인간 코더와 4개 LLM의 코딩 결과를 비교 분석하였다. 궁극적으로 본 연구는 LLM 기반 내용 분석의 쟁점을 특정 모델의 기술적 성능 문제로 환원하기보다, 코딩 과업의 판단 요구 수준과 정보의 명시성을 조건 변수로 설정하여 LLM의 수행 특성을 맥락 의존적 현상으로 재해석한다. 이를 통해 학술 연구 맥락에서 AI 활용의 적용 가능성과 한계를 과업 유형 및 판단 특성 차원에서 정교화하고자 한다.

Ⅱ. 선행연구

대규모 언어모델의 발전은 텍스트 분석 연구의 방법론적 지형을 빠르게 변화시키고 있다. Transformer 기반 사전학습 모델의 도입 이후(Devlin et al., 2019), GPT 계열 모델은 소수 예시 학습만으로도 다양한 분류 과제에서 높은 일반화 성능을 보이는 것으로 보고되었다(Brown et al., 2020). 이러한 기술적 진전은 감정 분석이나 주제 분류와 같은 전통적 자연어처리 과제를 넘어, 정책 프레이밍 분석, 정치적 메시지 분류, 사회적 담론 분석 등 사회과학적 영역으로 확장되고 있다(Gilardi et al., 2023; Grimmer et al., 2022; Törnberg, 2024). 그러나 기존 연구의 다수는 예측 정확도 중심의 성능 비교에 초점을 두고 있으며, 학술적 내용 분석 맥락에서 요구되는 신뢰도와 재현성 문제를 충분히 다루지 못하고 있다. 자동화된 텍스트 분석은 확장성과 효율성 측면에서 강점을 지니지만, 범주 정의와 코딩 규칙의 구조적 설계가 명확하지 않을 경우 해석의 일관성과 타당성이 약화될 수 있다는 점은 이미 지적되어 왔다(Grimmer & Stewart, 2013). 특히 선행연구는 단일 모델 평가나 단순 분류 과제에 국한되는 경향이 있어, 명시적 정보 추출, 추론적 판단, 평가적 판단과 같이 과업의 판단 요구 수준에 따라 신뢰도가 어떻게 달라지는지를 체계적으로 비교한 연구는 부족하다.

문헌정보학 분야에서 내용 분석은 연구 동향 파악과 주제 구조 해석을 위한 핵심 방법론으로 자리 잡아 왔다. 질적 내용 분석은 체계적 범주화 절차를 따르면서도 연구자의 해석적 판단이 개입되는 방법론이며, 그 엄밀성은 해석 과정의 투명성과 일관성에 의해 평가된다(정은경, 2021;

최성호 외, 2016; Lincoln & Guba, 1985; Tracy, 2010). Krippendorff(2018)는 내용 분석의 신뢰도를 범주 체계의 안정성과 재현 가능성의 문제로 정의하며, 단순한 합의율을 넘어선 구조적 기준을 제시하였다. 이러한 논의는 LLM 기반 자동화 코딩 역시 동일한 신뢰도 기준 아래에서 검토되어야 함을 시사한다.

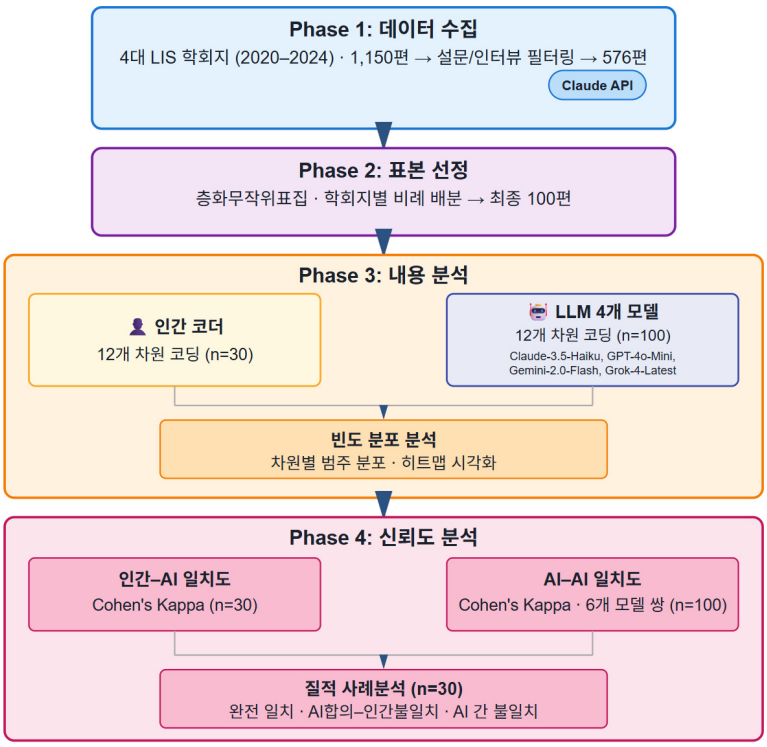
본 연구의 코딩 프레임워크는 이러한 이론적 전통을 토대로, 표집 방법론과 총조사오차(Total Survey Error) 이론에 근거하여 구성되었다(Cochran, 1977; Groves et al., 2009; Kalton, 1983). 또한 American Association for Public Opinion Research(AAPOR) Standard Definitions와 American Psychological Association(APA) Journal Article Reporting Standards를 참조하여 표집 절차, 응답률, 한계점 논의 등 연구 재현성과 직결되는 요소를 코딩 차원에 반영하였다(AAPOR, 2023; APA, 2020). 나아가 정량·정성·혼합 연구의 질 평가 기준을 종합하여(Pluye et al., 2009), 표본 크기 정당화와 편향에 대한 성찰을 관찰 가능한 지표로 조작화하였다. 이러한 선행 논의를 종합하면, LLM 기반 내용 분석의 신뢰도는 단순한 모델 성능 문제가 아니라 코딩 과업의 구조와 판단 요구 수준에 의해 조건적으로 매개되는 현상으로 이해될 필요가 있다. 특히 해석적·평가적 차원에서는 모델 간 분산이 구조적으로 발생할 가능성이 높으며, 따라서 전면적 자동화보다는 과업 유형에 기반한 조건부 활용 전략이 요구된다.

Ⅲ. 연구 방법

1. 연구 설계

본 연구는 인간 코더와 4개 대규모 언어모델(LLM)의 코딩 결과를 비교함으로써, LLM 기반 내용 분석의 신뢰도를 체계적으로 검증하는 방법론적 연구이다. 분석 대상은 2020년부터 2024년까지 국내 문헌정보학 주요 4개 학회지에 게재된 논문 중 설문조사 또는 인터뷰 방법을 활용한 연구 유형으로 한정하였다.

연구는 〈그림 1〉에 제시된 4단계 절차에 따라 수행되었다. 각 단계는 데이터 수집(Phase 1), 표본 선정(Phase 2), 내용 분석(Phase 3), 신뢰도 분석(Phase 4)으로 구성되며, 단계별 세부 내용은 다음과 같다.



〈그림 1〉 LLM 기반 내용 분석 연구 절차

가. Phase 1: 데이터 수집

초기 데이터 수집은 2020년 1월부터 2024년 12월까지 국내 4대 문헌정보학 학회지—『정보관리학회지』, 『한국도서관·정보학회지』, 『한국문헌정보학회지』, 『한국비블리아학회지』—에 게재된 전체 논문 1,150편을 대상으로 하였다. 스크리닝은 논문의 제목·초록·서론 초반부에 해당하는 선두 3,000자를 입력 범위로 설정하고, Claude API를 활용한 자동화 방식으로 수행하였다. 이 구간은 연구 방법 식별에 필요한 핵심 정보가 집중되는 최소 범위로, 1,150편 전체에 대한 처리 비용과 API 호출 속도를 고려하여 설정하였으며, 사전 파일럿 테스트(n=30)에서 설문·인터뷰 여부 식별에 충분함을 확인하였다. 검토 결과 설문조사 또는 인터뷰 방법을 명시적으로 활용한 논문 576편을 선별하였으며(부록 1 참조), 문헌 연구, 시스템 개발, 계량서지 분석 등 표집 방법론이 적용되지 않는 연구 유형은 제외하였다.

나. Phase 2: 표본 선정

최종 분석 대상 100편은 층화무작위표집(stratified random sampling) 방법으로 선정하였다.

표집 프레임은 연구 유형별 필터링을 거친 576편이며, 층화 기준은 학회지별 게재 비율이다. 각 학회지의 모집단 비율을 유지하는 비례 층화 방식을 적용하여 무작위 추출을 실시하였고, 재현 가능성을 확보하기 위해 난수 시드(random seed)는 2024로 고정하였다.

표본 규모는 100편으로 설정하였다. 이는 인간 코더 검증 범위(30편)의 약 3배에 해당하며, 선행 연구에서 LLM 기반 코딩 신뢰도 검증에 활용된 표본 규모와도 부합한다(Gilardi et al., 2023; Törnberg, 2024). 본 연구의 목적이 576편에 대한 전수 내용 분석 결과 산출보다 LLM 기반 코딩의 신뢰도 패턴과 적용 조건 규명에 있음을 고려할 때, 100편 규모는 분석 목적에 부합하는 방법론적 선택이다. 학회지별 표본 추출 현황은 〈표 1〉에 제시하였다.

연구 유형별 필터링된 모집단(576편) 대비 약 17.4%를 비례 배분하여 추출하였으며, 각 학회지에서도 17% 내외의 표집률을 유지하였다. 이러한 비례 층화 방식은 특정 학회지의 과대표집이나 과소대표집을 방지하고, 분야 내 방법론적 분포를 모집단 구조에 가깝게 재현하기 위한 설계이다.

〈표 1〉 학회지별 표본 추출 현황(2020-2024)

학회지명	검색된 전체 논문 (n=1,150)	설문·인터뷰 연구로 선별된 논문(n=576)	표본(표집률) (n=100)
정보관리학회지	273	119 (100%)	21 (17.6%)
한국도서관·정보학회지	289	151 (100%)	26 (17.2%)
한국문헌정보학회지	339	155 (100%)	27 (17.4%)
한국비블리아학회지	249	151 (100%)	26 (17.2%)
합계	1,150	576 (100%)	100(17.4%)

주: 표집률은 각 학회지별 표본 수를 해당 학회지 모집단으로 나눈 값이다. 층화무작위표집을 통해 학회지별 모집단 비율을 표본에 반영하였다(random seed: 2024).

다. Phase 3: 내용 분석

내용 분석은 인간 코더와 4개 LLM 코더의 이원적 체계로 수행하였다.

(1) 인간 코더 작업 절차

인간 코더(제1저자)는 표집 방법론, 총조사오차 이론, 주요 연구 보고 지침(AAPOR, 2023; APA, 2020)에 근거하여 12개 차원의 코딩 프레임워크를 개발하였다. 프레임워크는 연구 설계 유형(A), 표집 구조(B), 표집 보고의 충실성(C), 평가 및 성찰(D)의 네 범주로 구성된다(〈표 2〉 참조). 전체 100편 중 30편을 학회지별 층화무작위표집으로 검증용 표본으로 선정하여, 연구방법(Method) 섹션을 중심으로 차원별 코딩을 수행하였다. 검증 표본 30편은 차원별 일치도 추정에 필요한 최소 규모를 충족하면서도 분석 부담을 고려한 절충적 규모이며, 선행연구에서 일반적으로 활용된 검증 비율과도 부합한다(Gilardi et al., 2023; Krippendorff, 2018; Törnberg, 2024).

(2) LLM 코더 작업 절차

4개 LLM 모델(Claude-3.5-Haiku, GPT-4o-Mini, Gemini-2.0-Flash, Grok-4-Latest)의 선정 기준은 (1) 공식 API를 통한 접근 가능성과 (2) 2024년 기준 학술 및 실무 영역에서의 대표성으로, (3) 오픈소스 모델은 API 기반 재현 가능성과 안정적 버전 관리의 어려움을 고려하여 분석 대상에서 제외하였다. 각 모델에는 인간 코더가 사용한 동일한 코딩 프레임워크와 범주 정의를 프롬프트 형식으로 제공하였다. 프롬프트는 (1) 연구 목적 및 과업 설명, (2) 〈표 2〉에 제시된 12개 차원별 정의 및 범주 기준, (3) 구조화된 출력 형식 지정의 세 부분으로 구성되었다. 1차 스크리닝 프롬프트는 YES/NO 이진 판단 및 방법 유형 분류를 지시하는 단순 구조이며, 2차 코딩 프롬프트는 12개 필드 각각에 대해 정해진 범주 목록 중 하나를 선택하도록 지시하는 구조화된 형태이다([부록 1, 2] 참조).

모든 모델에 공통 프롬프트와 입력 구조를 적용하여 전체 100편을 독립적으로 코딩하도록 하였다. 각 모델은 공식 API(Claude: Anthropic, GPT-4o-Mini: OpenAI, Gemini-2.0-Flash: Google, Grok-4-Latest: xAI)를 통해 호출하였으며, temperature는 0으로 설정하여 결정론적 출력을 확보하였다. 논문 원문은 PDF 형식으로 입력하였고, 텍스트 추출에는 pdfplumber를 사용하였다. 입력 범위는 선두 12,000자로 설정하였으며, 이는 연구 방법 섹션 전체를 포함할 수 있는 구간으로, 사전 파일럿 테스트에서 분석 대상 100편 중 93편의 방법 섹션이 해당 범위 내에 포함됨을 확인하였다. 다만 모델별 API 환경의 차이에 따라 응답 수집 방식에는 일부 구현상의 차이가 존재할 수 있으며, 수집된 출력은 후처리 과정을 거쳐 동일한 차원 체계로 정형화하였다. 이를 바탕으로 차원별 빈도 분석과 인간-LLM 및 LLM-LLM 간 일치도(κ) 분석을 수행하였다.

(3) 코딩 차원 구성 원리

본 연구는 표집 방법론의 핵심 구성 요소와 연구 보고 관행, 평가적 성찰을 포괄적으로 반영하기 위해 총 12개 코딩 차원(부록 2 참조: 총 12개 코딩 차원/ A 범주 1개, B 범주 4개, C 범주 4개, D 범주 3개)을 설정하였다. 이들 차원은 기능적 성격에 따라 연구 설계 유형(A), 표집 구조(B), 표집 보고의 충실성(C), 평가 및 성찰(D)의 네 범주로 집합화하였다.

범주 A(연구 설계 유형)와 B(표집 구조)의 범주 체계는 Cochran(1977), Groves et al.(2009) 및 AAPOR(2023)의 표집 유형 분류 체계를 기반으로 구성하였다. 범주 C(표집 보고의 충실성)는 APA(2020)의 연구 보고 기준 및 Pluye et al.(2009)의 혼합 연구 질 평가 지표를 참조하였다. 범주 D(평가 및 성찰) 중 주관적 판단이 개입될 수 있는 D_SAMPLING_BIAS와 D_OVERALL_QUALITY는 코딩 지침에 편향 유형별 정의와 품질 수준별 기술 기준을 예시와 함께 명시하여 코더 간 해석의 일관성을 최대한 확보하고자 하였다([부록 2] 참조).

연구 설계 유형(A)은 설문이나 인터뷰와 같은 기본 방법적 틀을 구분하는 차원이다. 표집 구조(B)는 표집 방법, 표집 프레임, 모집 방식 등 참여자가 어떠한 경로와 절차를 통해 선정·모집되

었는지를 포착한다. 표집 보고의 충실성(C)은 표본 크기 정당화, 응답률, 포함 기준, 인구통계 보고 여부 등 표집 관련 핵심 정보가 논문 내에 얼마나 명시적으로 기술되었는지를 평가하는 차원이다. 평가 및 성찰(D)은 표집 편향, 한계점 논의, 전반적 품질과 같이 연구자가 자신의 표집 설계와 한계를 어떻게 성찰적으로 다루고 있는지를 포착하는 평가 차원으로 구성된다. 코딩 차원을 영문으로 설계한 것은 이론적 출처(AAPOR, 2023; APA, 2020)와의 개념적 정합성을 유지하고 번역 과정의 모호성을 제거하기 위한 의도적 선택이다. 또한 모든 모델이 동일하게 한국어 입력을 영문 범주로 매핑하는 조건에 놓였으므로, 한국어 처리 능력의 차이는 모델 간 비교의 공정성을 부분적으로 통제하는 효과가 있다. 그러나 이것이 한국어 이해 능력의 차이를 완전히 제거한다고 볼 수는 없으며, 이 점은 본 연구의 해석 범위를 제한하는 요인으로 작용한다.

〈표 2〉 코딩 프레임워크

방법론적 영역 (유형 A-D)	차원 코드	설명	범주 예시(축약)
연구 설계 유형 (A)	A_METHOD_TYPE	연구 방법	Survey, Interview, Both
표집 구조 (B)	B_SAMPLING_METHOD	표집 방법	Probability_Random, Convenience, Purposive
	B_SAMPLING_FRAME	표집 프레임	Institutional_List, Membership_Database, Not_Stated
	B_SAMPLE_SIZE_CATEGORY	표본 크기 범주	Very_Small(〈10), Small(10-30),Large(〉100)
	B_RECRUITMENT_METHOD	모집 방식	Email, In_Person, Not_Stated
표집 보고의 충실성 (C)	C_SAMPLE_SIZE_JUSTIFICATION	표본 크기 정당화	Power_Analysis, No_Justification
	C_RESPONSE_RATE	응답률 범주	Very_High(〉70%), Low(〈40%), Not_Stated
	C_INCLUSION_CRITERIA	참여자 포함 기준	Clear_Stated, Vague, Not_Stated
	C_DEMOGRAPHICS_REPORTED	인구통계 보고	Comprehensive, Moderate, None
평가 및 성찰 (D)	D_SAMPLING_BIAS	표집 편향	Convenience_Bias, Self_Selection_Bias
	D_LIMITATIONS_DISCUSSED	한계점 논의	Comprehensive, Brief, Not_Discussed
	D_OVERALL_QUALITY	전반적 품질	2_Poor, 3_Adequate, 4_Good

(4) 차원별 빈도 분포 비교

빈도 분석은 전체 100편을 대상으로 4개 LLM의 코딩 결과를 차원별·범주별로 집계하여 모델 간 범주 선택 분포를 비교하기 위해 수행하였다. 이를 통해 각 모델이 특정 범주를 체계적으로 과다 또는 과소 선택하는 경향이 존재하는지를 먼저 점검하였다. 분석 결과는 차원×모델 구조의 히트맵(heatmap)으로 시각화하여 범주 분포의 상대적 편차를 직관적으로 파악할 수 있도록 구성하였다. 이러한 빈도 분포 분석은 이후 수행되는 인간-LLM 및 LLM-LLM 간 신뢰도 분석 결과를 해석하기 위한 기초 자료로 활용되었다.

라. Phase 4: 신뢰도 분석

신뢰도 분석은 AI-AI 간 코더 일치도 분석과 인간-AI 간 일치도 분석 및 질적 사례 분석의

두 수준에서 수행하였다.

(1) AI-AI 간 코더 일치도 분석

AI-AI 간 일치도는 전체 100편을 대상으로 4개 LLM 간 모든 모델 쌍(총 6개 조합)에 대해 차원별 Cohen's kappa를 산출하여 비교하였다. 범주형 자료의 코더 간 일치도 지표로 Cohen's kappa(Cohen, 1960)를 사용하였으며, 우연적 일치를 보정하는 특성을 고려하여 단순 일치율보다 보수적인 추정을 제공한다. Kappa 값의 해석은 Landis와 Koch(1977)의 기준을 참고하되, 절대적 판정 기준이 아닌 차원 간 상대 비교를 위한 해석 틀로 활용하였다. 이를 통해 모델 간 일관성과 분산 구조를 분석하고, 코딩 과업의 구조적 특성에 따른 변동을 파악하였다.

(2) 인간-AI 간 질적 사례 분석

인간-AI 간 일치도는 검증용 표본 30편을 대상으로 질적 사례 분석을 통해 보완적으로 검토하였다. 사례는 (1) 인간과 4개 AI 모델이 모두 일치한 경우, (2) AI 모델 간 합의가 있으나 인간과 불일치한 경우, (3) AI 모델 간 판단이 크게 분기된 경우의 세 유형으로 분류하였다. 해당 사례들은 코딩 과업의 구조적 특성과 판단 기준의 차이를 질적으로 해석하기 위한 분석적 자료로 활용되었다. 이를 보완하기 위해 검증용 표본 30편을 대상으로 인간 코더와 4개 AI 모델 간 차원별 Cohen's kappa를 별도 산출하여 〈그림 7〉로 제시하였으며, 이는 AI-AI 간 일치도 분석(n＝100)과 표본 크기가 상이함을 유의할 필요가 있다.

Ⅳ. 연구 결과

1. 차원별 빈도 분포 비교

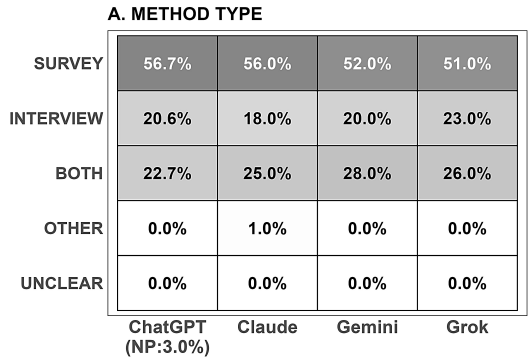
본 절에서는 100편 전체 표본에 대해 4개 LLM(GPT-4o-Mini, Claude-3.5-Haiku, Gemini-2.0-Flash, Grok-4-Latest)의 코딩 결과 분포를 비교하여, 차원별 판단 경향의 구조적 차이를 분석하였다. 이는 신뢰도 분석 이전 단계에서 각 모델이 어떤 범주를 선호하는지를 확인하기 위한 것이다. 이하 그림에서는 가독성을 위해 모델명을 ChatGPT, Claude, Gemini, Grok으로 약칭하여 표기하였다.

〈그림 2〉에서 〈그림 5〉까지 표기된 ChatGPT의 각 차원별 NP(Not Parsed) 비율은 해당 차원에서 모델이 범주를 반환하지 못하거나 분석 대상에서 제외된 사례의 비율을 의미한다. 이는 무응답률과는 구분되며, 모델 출력의 파싱 가능성에 대한 기술적 지표로 제시된다.

가. 연구 설계 유형 내용 분석

〈그림 2〉는 4개 LLM이 100편의 논문을 대상으로 분류한 연구 설계 유형(A_METHOD_TYPE)의 분포를 비교한 결과를 제시한다. 모든 모델에서 설문(SURVEY) 연구가 가장 높은 비율을 차지하였으며, ChatGPT(56.7%), Claude(56.0%), Gemini(52.0%), Grok(51.0%)으로 4개 모델 모두 51%-57% 범위 내에서 비교적 안정적으로 수렴하였다. 인터뷰(INTERVIEW) 연구는 전 모델에서 18.0%-23.0%로 유사한 분포를 보였으며, 설문과 인터뷰를 병행한 혼합형(BOTH) 연구는 Gemini(28.0%), Grok(26.0%), Claude(25.0%), ChatGPT(22.7%)로 모든 모델에서 22%-28% 범위로 나타났다.

OTHER 범주는 Claude(1.0%)를 제외한 전 모델에서 0.0%로 나타났으며, UNCLEAR 범주는 전 모델에서 0.0%로 확인되었다. 종합하면, 연구 설계 유형 분류에서는 네 모델 모두 비교적 일관된 판단 경향을 보였으며, 이는 연구 방법이 논문 내에 비교적 명시적으로 기술되는 차원에서 모델 간 판단이 안정적으로 수렴함을 시사한다.



〈그림 2〉 연구 설계 유형(A_METHOD_TYPE)에 대한 LLM별 분류 분포 히트맵

나. 표집 구조 내용 분석

〈그림 3〉은 표집 구조(B)에 해당하는 네 개 차원에 대한 LLM별 분류 분포를 제시한다.

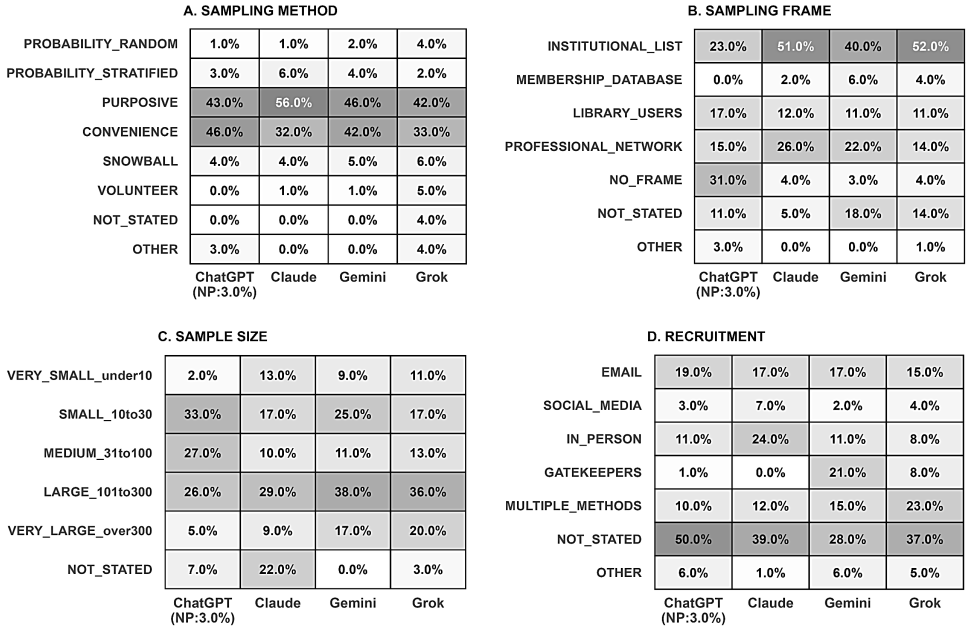
A. 표집방법(B_SAMPLING_METHOD)에서는 목적표집(PURPOSIVE)과 편의표집(CONVENIENCE)이 모든 모델에서 높은 비중을 차지하였으나, 모델별 상대적 비중은 상이하였다. Claude는 목적표집을 56%로 가장 높게 분류한 반면, ChatGPT는 편의표집(CONVENIENCE)을 46%로 가장 높게 분류하여 우세 범주의 방향성이 달랐다. OTHER 범주는 전 모델에서 0%-4% 수준으로 낮게 나타나 범주 귀속 불가 사례는 전반적으로 드물었다.

B. 표집프레임(B_SAMPLING_FRAME)에서는 Claude(51%)와 Grok(52%)이 기관목록(INSTITUTIONAL_LIST)을 가장 많이 분류한 반면, ChatGPT는 기관목록(23%)보다 NO_FRAME(31%)을 더 높게 분류하여 표집 프레임 부재 판단 경향이 두드러졌다. Gemini는

40%로 기관 목록 비율이 중간 수준이었다.

C. 표본 크기(B_SAMPLE_SIZE)에서는 Gemini와 Grok이 LARGE_101to300을 각각 38%와 36%로 가장 높게 분류한 반면, ChatGPT는 SMALL_10to30(33%), MEDIUM_31to100(27%), LARGE_101to300(26%)에 비교적 고르게 분산되었고 NOT_STATED는 7%로 낮았다. Claude는 NOT_STATED를 22%로 상대적으로 높게 분류하였다.

D. 모집 방식(B_RECRUITMENT_METHOD)에서는 전 모델에 걸쳐 NOT_STATED 비율이 높았으나, ChatGPT(50%)와 Grok(37%)이 특히 높게 나타났다. Claude는 IN_PERSON(24%)과 EMAIL(17%)을 비교적 적극적으로 분류하여 모집 경로 식별 경향이 상대적으로 강하였다. 이러한 차이는 동일한 텍스트 조건에서도 표집 구조와 관련된 정보 해석 기준이 모델별로 체계적으로 다를 수 있음을 시사한다.



〈그림 3〉 표집 구조에 대한 LLM별 분류 분포 히트맵

다. 표집 보고의 충실성 내용 분석

〈그림 4〉는 표집 보고의 충실성 범주 네 개 차원에 대한 LLM별 분류 분포를 제시한다.

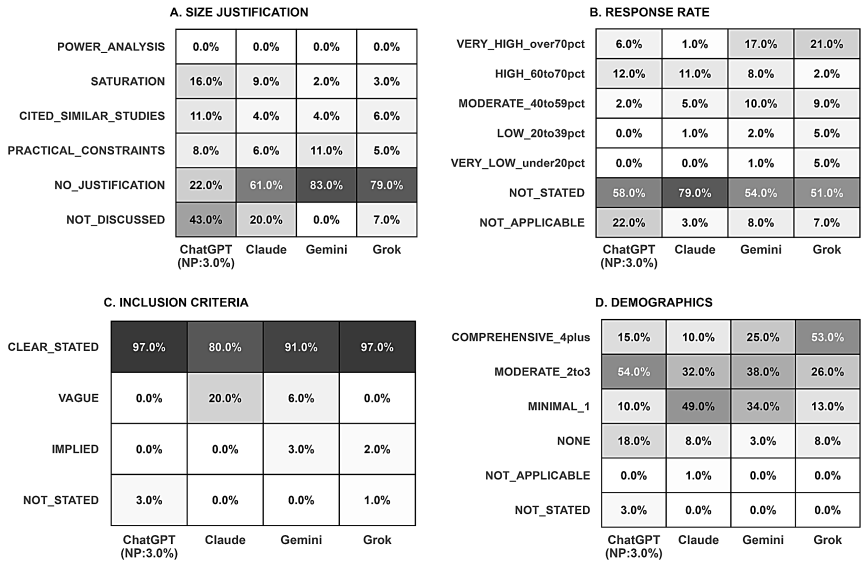
A. 표본 크기 정당화(C_SAMPLE_SIZE_JUSTIFICATION)는 모든 모델에서 정당화 없음(NO_JUSTIFICATION) 또는 미논의(NOT_DISCUSSED)가 지배적이었으나, 그 분포 양상은 모델 간에 상이하였다. Claude(61%)와 Gemini(83%), Grok(79%)은 NO_JUSTIFICATION

을 최빈 범주로 분류한 반면, ChatGPT는 NOT_DISCUSSED(43%)를 가장 높게 분류하였다. ChatGPT는 또한 SATURATION(16%), CITED_SIMILAR_STUDIES(11%)를 다른 모델에 비해 상대적으로 높게 분류하여 정당화 방식 판단 기준이 다른 모델과 구조적으로 상이함을 보여주었다.

B. 응답률(C_RESPONSE_RATE)은 전 모델에서 미보고(NOT_STATED) 비율이 51%-79%로 전반적으로 높았으며, ChatGPT는 NOT_APPLICABLE을 23%로 가장 높게 분류하여 응답률 개념 자체를 적용 불가로 판단하는 경향이 두드러졌다. Grok(21%)도 VERY_HIGH_over70pct를 비교적 높게 분류하여 응답률을 명시적으로 추출하는 경향이 있었다.

C. 참여자 포함 기준(C_INCLUSION_CRITERIA)은 ChatGPT(97%), Grok(97%), Gemini(91%)에서 명확히 제시된 경우(CLEAR_STATED)가 압도적으로 높았으나, Claude만 VAGUE를 20%로 분류하여 다른 모델과 상이한 판단 기준을 적용하였다.

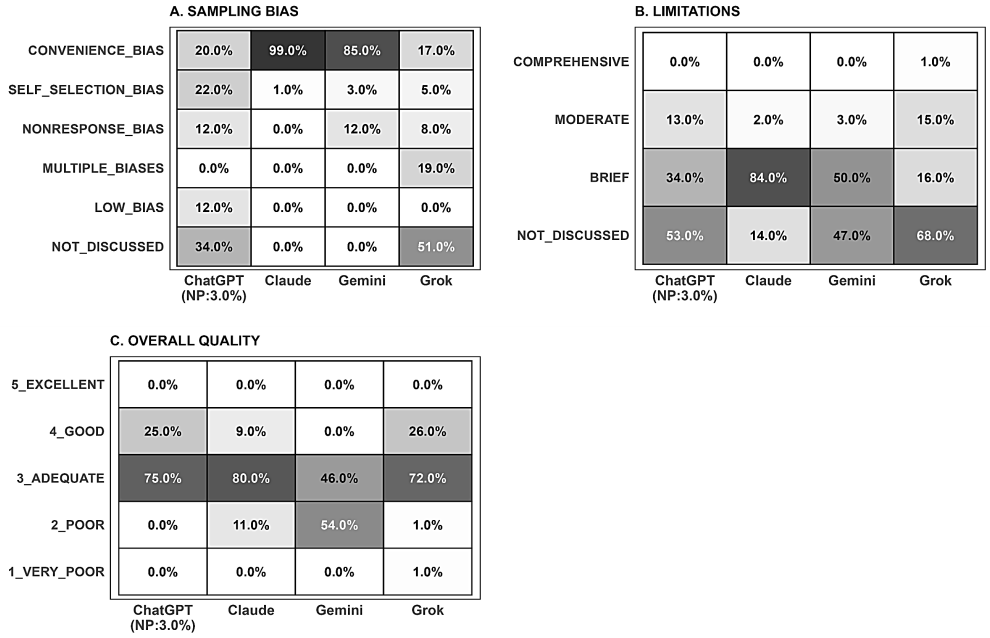
D. 인구통계 보고(C_DEMOGRAPHICS)에서는 Grok이 포괄적 보고(COMPREHENSIVE_4plus)를 53%로 가장 높게 분류한 반면, Claude는 최소 보고(MINIMAL_1)를 49%로 가장 높게 분류하였다. ChatGPT는 MODERATE_2to3를 54%로 가장 높게 분류하여 중간 수준의 인구통계 보고를 주로 식별하였다. 이는 해석 판단이 개입되는 차원일수록 모델 간 분류 기준이 체계적으로 달라짐을 보여준다. 인구통계 보고 차원에서는 해당 코딩 체계에 'NOT_STATED' 범주가 정의되어 있지 않음에도 불구하고, ChatGPT는 3%의 응답을 해당 범주로 분류하였다. 이는 프롬프트 지시를 벗어난 범주를 모델이 임의로 생성한 사례로, ChatGPT의 지시 준수 충실도에 관한 방법론적 검토가 필요함을 시사한다.



〈그림 4〉 표집 보고의 충실성 차원에 대한 LLM별 분류 분포 히트맵

라. 평가 및 성찰 내용 분석

〈그림 5〉는 평가 및 성찰(D)에 해당하는 세 차원, 즉 표집 편향(D_SAMPLING_BIAS), 한계점 논의(D_LIMITATIONS), 전반적 품질(D_OVERALL_QUALITY)에 대한 LLM별 분류 분포를 제시한다.



〈그림 5〉 평가 및 성찰 차원에 대한 LLM별 분류 분포 히트맵

A. 표집 편향 차원에서 Claude와 Gemini는 편의표집 편향(CONVENIENCE_BIAS)을 각각 99.0%와 85.0%로 집중적으로 분류한 반면, ChatGPT는 자기선택 편향(SELF_SELECTION_BIAS, 22%), 편의표집 편향(20%), 무응답 편향(NONRESPONSE_BIAS, 12%), 낮은 편향(LOW_BIAS, 12%), 미논의(NOT_DISCUSSED, 34%)에 비교적 분산된 분포를 보였다. Grok은 미논의(NOT_DISCUSSED, 51%)와 복수 편향(MULTIPLE_BIASES, 19%)을 높게 분류하여 편향 귀속 기준이 모델 간에 체계적으로 상이함을 보여준다.

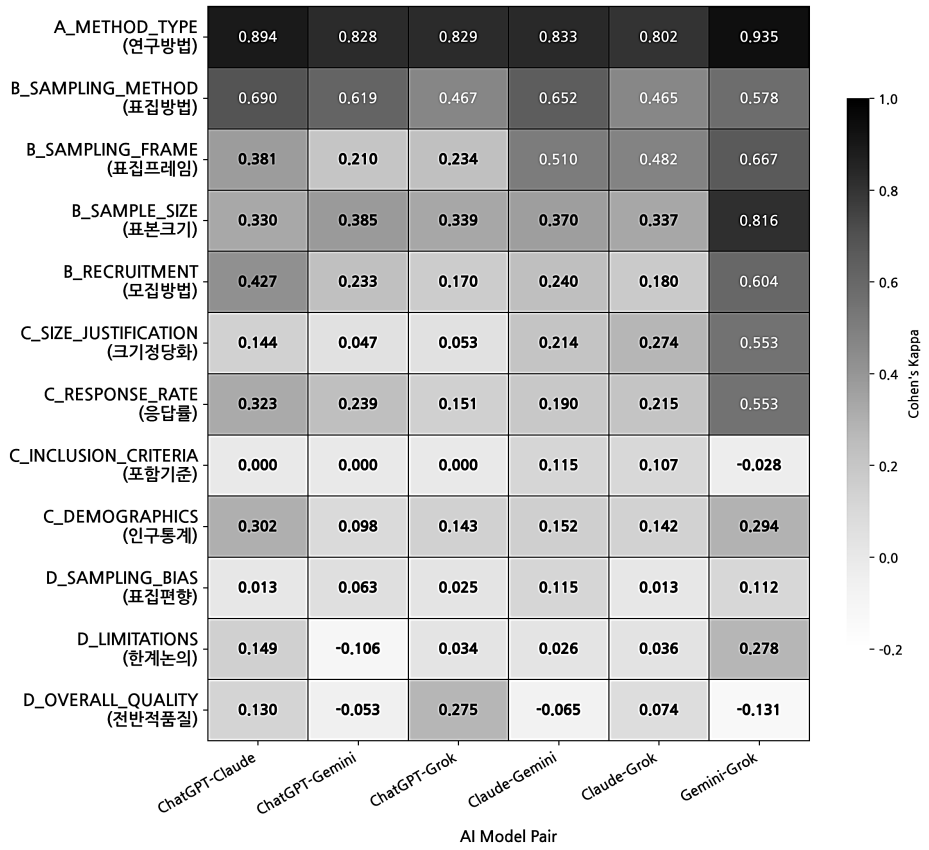
B. 한계점 논의 차원에서는 Claude가 BRIEF를 84.0%로 분류한 반면, ChatGPT(53%)와 Grok(68%)은 NOT_DISCUSSED를 최빈 범주로 분류하여 동일한 기술을 두고 '논의'로 판단하는 기준이 모델 간에 현저히 달랐다. Gemini는 BRIEF(50%)와 NOT_DISCUSSED(47%)가 거의 동등하게 분포하여 중간적 경향을 보였다.

C. 전반적 품질(D_OVERALL_QUALITY) 차원에서는 모든 모델에서 3_ADEQUATE가

최빈 범주였으나 세부 분포에서 유의한 차이가 관찰되었다. 주목할 점은 Gemini의 이례적 엄격성으로, 2_POOR 비율이 54.0%로 가장 높게 나타났다. 질적 사례 분석에서 확인한 바에 따르면, Gemini는 표집 방법이 명확히 기술되지 않은 경우 이를 방법론적 결함으로 적극 해석하여 낮은 품질 점수를 부여하는 경향이 있었다. 반면 동일 논문에 대해 ChatGPT와 Grok은 기술된 정보를 있는 그대로 수용하여 3_ADEQUATE 또는 4_GOOD으로 판단하였으며, 두 모델의 4_GOOD 비율은 각각 25.0%와 26.0%로 비교적 관대한 평가 기준을 적용하였다. 이는 평가적 판단 차원에서의 모델 간 기준 편차가 단순한 이진 분류 오류가 아니라 평가 철학 수준의 구조적 차이에서 비롯될 수 있음을 시사한다. 이들 세 차원은 모두 해석적 귀속과 규범적 판단을 요구하는 영역으로, 판단 요구 수준이 높을수록 모델 간 기준 편차가 구조적으로 확대되는 양상이 확인된다.

2. AI-AI(모델-모델) 간 신뢰도

AI-AI 간 신뢰도는 전체 평균 $\kappa \approx 0.25$로 낮게 나타났다. 이는 빈도 분포 분석에서 확인된 모델별 범주 귀속 경향의 체계적 차이와 직결된다. 예컨대 표집 편향(D_SAMPLING_BIAS) 차원에서 Claude와 Gemini는 편의표집 편향에 집중된 반면, ChatGPT는 다양한 범주로 분산되었고 Grok은 미논의 비율이 상대적으로 높았다. 이처럼 동일한 텍스트 조건에서도 범주 선택 기준이 구조적으로 상이할 경우 모델 간 합의는 낮아지며, 이는 κ값 하락으로 반영된다. 다만 이러한 변동은 모델 조합의 차이보다 코딩 차원별 판단 구조의 차이가 더 크게 설명한다. 〈그림 6〉은 차원별 · 모델 쌍별 Cohen's kappa를 제시하며, 일치도 수준이 차원에 따라 계층적으로 구분됨을 보여준다. 연구 설계 유형(A_METHOD_TYPE)은 모든 모델 쌍에서 가장 높은 일치도를 보였으며, 대부분의 조합에서 $\kappa = 0.808 \sim 0.952$로 안정적 수렴이 관찰되었다. 표집 구조(B)와 보고 충실성(C) 차원은 중간 수준의 변동을 나타냈는데, 표본 크기 범주(B_SAMPLE_SIZE_CATEGORY)는 Gemini-Grok 조합에서 0.816으로 비교적 높은 반면, 모집 방식(B_RECRUITMENT_METHOD)과 응답률(C_RESPONSE_RATE)은 일부 조합에서 0.180~0.330에 머물렀다. 이는 정보의 명시성과 범주 경계의 해석 가능성에 따라 차원별 안정성이 달라짐을 보여준다. 가장 큰 변동은 평가 및 성찰(D) 범주에서 확인되었다. 표집 편향(D_SAMPLING_BIAS)은 대부분의 모델 쌍에서 $\kappa = 0.015 \sim 0.112$로 우연 수준에 가까웠으며, 전반적 품질(D_OVERALL_QUALITY)에서는 Claude-Gemini(-0.065), Gemini-Grok(-0.131) 등 음의 Kappa가 관찰되었다. 이는 단순한 불일치를 넘어 평가의 방향성과 기준 자체가 모델 간에 상이함을 시사한다. 종합하면, AI-AI 간 신뢰도 변동은 특정 모델의 기술적 성능 차이보다 코딩 차원의 해석 요구 수준과 정보 구조의 명시성에 조건적으로 의존하며, 평가적 판단을 요구하는 차원일수록 모델 간 기준 편차가 확대됨이 확인된다.

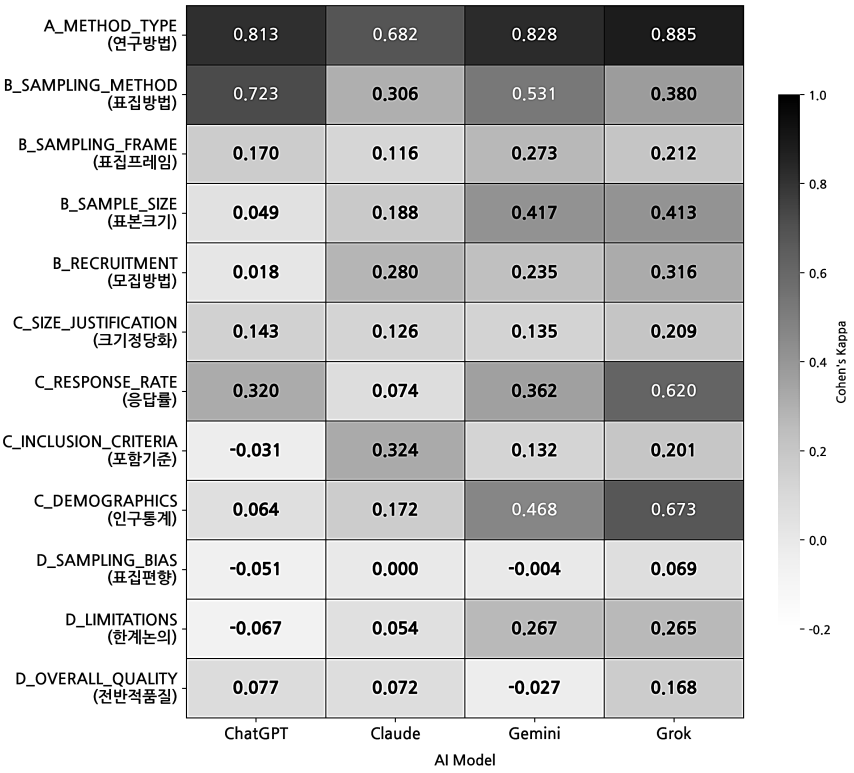


주. 전체 평균 κ값은 약 0.25이다(12개 차원 × 6개 모델 쌍, 총 72개 κ 값의 단순 평균).

〈그림 6〉 차원별 AI 모델 간 Cohen's kappa 비교 히트맵(n=100)

3. 인간-AI 간 질적 사례 분석

본 절은 정량적 신뢰도 분석을 보완하기 위해, 검증용 표본 30편을 대상으로 인간 코더와 LLM 간 판단 차이가 발생하는 양상을 과업 유형별로 분석한다. 〈그림 7〉은 검증용 표본 30편을 대상으로 산출한 인간-AI 간 차원별 Cohen's kappa를 제시한다. A_METHOD_TYPE에서는 전 모델에서 $\kappa = 0.682 \sim 0.885$로 높은 일치도가 확인된 반면, D_SAMPLING_BIAS에서는 $\kappa = -0.051 \sim 0.069$로 우연 수준에 가까웠으며, D_OVERALL_QUALITY 역시 $\kappa = -0.027 \sim 0.168$로 낮게 나타났다. 이러한 패턴은 〈그림 6〉의 AI-AI 간 신뢰도 구조와 수렴하며, 판단 요구 수준에 따른 일치도 계층화가 인간-AI 쌍에서도 동일하게 나타남을 보여준다. 이하에서는 이 정량적 패턴의 구조적 원인을 질적 사례 분석을 통해 해석한다.



〈그림 7〉 인간-AI 모델 간 Cohen's kappa 비교 히트맵(n=30)

질적 사례 분석의 목적은 개별 논문의 방법론적 적절성을 평가하는 데 있지 않으며, 동일한 텍스트 조건에서 인간과 AI의 판단이 어떠한 과업 구조적 조건 아래에서, 어떠한 방향으로 분기하는지를 유형화하는 데 있다. 코딩 일치 패턴은 (1) 인간과 AI 4개 모델이 모두 일치한 경우(Type 1), (2) AI 모델 간 합의가 이루어졌으나 인간과 불일치한 경우(Type 2), (3) AI 모델 간 판단이 크게 분기된 경우(Type 3)의 세 유형으로 분류하였다. 각 유형의 분포와 구조적 특성은 〈표 3〉에 요약하였다.

〈표 3〉 질적 사례 유형별 분포 (n=30)

유형	정의	대표 차원	사례 수(%)	판단 차이의 구조적 원인
Type 1	인간·AI 모두 일치	A_METHOD_TYPE	23편 (76.7%)	정보 명시성 높음 → 해석 개입 불필요
Type 2	AI 간 합의, 인간 불일치	B_SAMPLING_METHOD	3편 (10.0%)	AI: 텍스트 표면 해석 인간: 연구 맥락 종합
Type 3	AI 간 불일치	D_SAMPLING_BIAS	25편 (83.3%)	평가 기준 자체가 부재 → 코더별 해석 원칙 상이

주. 각 유형은 상이한 코딩 차원을 대표 사례로 제시한 것으로, n=30이 각 유형의 100% 기준임.

가. 명시적 추출 과업: 판단 수렴(Type 1)

명시적 정보 추출 과업에 해당하는 A_METHOD_TYPE 차원에서는 검증용 표본 30편 중 23편(76.7%)에서 인간과 AI 4개 모델이 모두 동일한 판단을 보였다. 이들 논문은 대체로 연구방법을 제목이나 초록에 명시적으로 기술하는 공통적 특징을 지니고 있었다. 대표 사례로, 김나연과 정은경(2020)의 연구는 의도적 표집(purposive sampling)과 10명 미만 규모의 인터뷰라는 질적 연구의 전형적 특성을 명확히 제시하고 있어, 모든 코더가 일관되게 연구 방법 및 표집 방법을 분류할 수 있었다. 이 유형에서 판단 수렴의 조건은 코더의 종류(인간 또는 AI)가 아니라 텍스트 내 정보의 명시성 수준임이 확인된다. 즉, 판단 기준이 텍스트에 충분히 고정되어 있을 때 인간과 AI의 해석 경로는 동일한 결과로 수렴한다.

나. 추론적 판단 과업: 해석 기준의 구조적 분기(Type 2)

표집 방법 정보가 명시되지 않은 상황에서 요구되는 추론적 판단 과업에서는 AI와 인간 코더 사이에 체계적인 불일치 패턴이 관찰되었다. 표집 방법(B_SAMPLING_METHOD) 차원에서 3편이 확인되었는데, 이선애(2023)와 송지애(2021) 논문에서 AI 4개 모델이 모두 편의표집(CONVENIENCE)으로 합의한 반면, 인간 코더는 자원자 표집(VOLUNTEER)으로 판단하였다. 이선애(2024)에서도 AI 4개 모델이 편의표집으로 합의하였으나 인간 코더는 목적표집(PURPOSIVE)으로 분류하였다. AI는 논문 내 텍스트의 표면적 표현을 근거로 범주를 귀속하는 반면, 인간 코더는 연구 목적, 참여자 선정 맥락, 방법론 관행을 종합적으로 고려하여 판단한다. 이 차이는 AI의 오류가 아니라 해석 기준의 구조적 차이로 이해되어야 한다. 동시에 인간 코더 역시 맥락 의존적 해석으로 인한 편향에서 자유롭지 않으며, 이는 인간 코딩을 절대 기준으로 설정하는 관행의 한계를 함께 드러낸다.

다. 평가적 판단 과업: AI 간 판단 원칙의 분기(Type 3)

평가적 판단을 요구하는 D_SAMPLING_BIAS에서는 AI 모델 간 판단이 크게 분기되는 사례가 다수 관찰되었다. 검증용 표본 30편 중 25편(83.3%)에서 AI 모델 간 불일치가 나타났으며, 이는 정량 분석에서 이 차원의 AI-AI 간 κ값이 0.015~0.112(그림 6), 인간-AI 간 κ값이 −0.051~0.069(〈그림 7〉 참조)로 모두 우연 수준에 가깝게 나타난 결과와 일치한다. 대표 사례인 Kim(2021) 논문에서 인간 코더는 자기선택 편향(SELF-SELECTION BIAS)을 주요 편향으로 판단한 반면, Claude는 편의표집 편향(CONVENIENCE BIAS), Gemini는 무응답 편향(NONRESPONSE BIAS), Grok은 복합 편향(MULTIPLE BIASES), ChatGPT는 편의표집 편향(CONVENIENCE BIAS)으로 각각 분류하였다. 해당 논문은 편의표집, 낮은 응답률, 자발적 참여 등 복수의 편향 요소를 동시에 포함하고 있었으며, 각 코더가 서로 다른 편향 요인을 우선적으로 강조한 결과 판단이

분기되었다. 이 유형에서 주목할 점은 인간 코더의 판단 역시 다수 AI 모델과 불일치한다는 점이다. 이는 AI의 성능 한계가 아니라 평가 기준 자체가 객관적으로 고정되기 어려운 과업 특성에서 비롯된다. 복수의 편향 요소가 공존하는 경우 어떤 편향을 '주요 편향'으로 귀속할지에 대한 합의 가능한 기준이 존재하지 않으며, 이로 인해 코더 유형(인간/AI)과 무관하게 판단 분기가 구조적으로 발생한다.

라. 종합: 과업 구조가 판단 차이의 일차 결정 요인

세 유형의 분석을 종합하면, 인간-AI 판단 차이의 성격과 방향은 코더의 종류보다 과업이 요구하는 판단 구조에 의해 일차적으로 결정됨이 확인된다. 명시적 추출 과업에서는 인간과 AI 모두 텍스트에 고정된 정보를 동일하게 귀속하여 판단이 수렴한다. 추론적 판단 과업에서는 정보가 명시되지 않은 상황에서 AI의 문자적 해석과 인간의 맥락적 해석이 구조적으로 분기하며, 이 차이는 AI의 한계인 동시에 인간 판단의 편향 가능성을 함께 드러낸다. 평가적 판단 과업에서는 평가 기준 자체의 부재로 인해 인간과 AI 모두에서 판단 불안정성이 발생하며, 이는 코딩 과업의 설계 수준에서 해결되어야 할 문제다. 아울러 〈그림 7〉의 인간-AI κ 패턴이 〈그림 6〉의 AI-AI κ 패턴과 구조적으로 수렴한다는 점은, 일치도 변동의 주된 원인이 특정 코더 유형이 아니라 코딩 과업의 판단 구조에 있음을 정량적으로 뒷받침한다.

Ⅴ. 논의 및 한계

본 연구의 핵심 발견은 LLM 기반 내용 분석의 일치도가 코딩 과업의 판단 구조에 따라 체계적으로 달라진다는 점이다. AI-AI 간 평균 일치도는 $\bar{\kappa} \approx 0.25$로 나타났으며, 이는 특정 모델의 성능 한계보다 차원별 판단 요구 수준의 차이에 의해 구조적으로 나타난 현상으로 해석된다. 표집 편향이나 전반적 품질과 같이 해석 개입이 필요한 차원에서는 낮은 일치도가 반복적으로 관찰된 반면, 연구 설계 유형과 같이 명시적 정보에 기반한 차원에서는 모델 간 판단이 비교적 안정적으로 수렴하였다. 코딩 차원의 기능적 성격에 따라 일치도가 계층적으로 분화되는 이러한 패턴은, 자동화가 가능한 영역과 인간 검증이 필요한 영역을 차원 수준에서 구분하는 경험적 근거를 제공한다.

본 연구는 인간 코딩을 절대적 기준으로 전제해 온 관행에 대해서도 재검토의 필요성을 제기한다. Type 2 사례에서처럼 AI 4개 모델이 모두 합의한 판단이 인간 코더의 판단과 불일치한 경우도 존재하며, 이를 단순히 AI의 오류로 규정하기는 어렵다. 표집 방법 차원에서 AI 모델들이 모두 편의표집으로 합의하였으나 인간 코더가 자원자 표집 또는 목적표집으로 판단한 사례가 확인되었는데, 이는 두 가지를 동시에 드러낸다. 하나는 모집 절차의 표면적 기술에 근거하여 범주를 귀속

하는 AI의 경향과 연구 목적 및 맥락을 종합적으로 고려하는 인간 판단 간의 구조적 차이다. 다른 하나는 방법론 용어의 비표준적 사용이라는 문헌 수준의 문제와 인간 코딩 역시 해석 편향으로부터 자유롭지 않다는 점이다. Type 3 사례에서 확인된 모델 간 판단 분기 역시 특정 모델의 오류라기보다, 평가적 판단 차원에서 객관적 기준이 부재할 때 각 모델이 서로 다른 해석 원칙을 적용하는 데서 기인한다. 이러한 분석은 인간-AI 불일치를 단순히 AI의 정확도 문제로 환원하기보다, 판단 과업의 구조적 성격과 정보 명시성 수준에 따라 조건적으로 해석해야 함을 시사한다.

이러한 결과는 완전 자동화나 전면적 인간 코딩 모두 현실적 대안이 아님을 보여준다. 명시적 분류 차원은 LLM 단독 처리로 충분하고, 추론적 코딩 차원은 LLM 초안을 인간이 검토하는 방식이 적절하며, 평가적 판단 차원은 인간 주도 분석이 요구된다. 이러한 차원별 역할 분담 전략은 LLM의 확장성과 인간의 맥락적 판단력을 상호 보완적으로 활용하면서 분석 효율성과 방법론적 엄밀성을 함께 확보할 수 있는 현실적 대안이다. 다만 표집 편향 차원에서 Claude와 Gemini는 편의표집 편향에, ChatGPT는 여러 편향 유형에 분산된 분포를 보였고 Grok은 미논의 비율이 높게 나타나는 등, 과업 구조 외에 모델 수준의 범주 귀속 기준 차이도 일치도 변동에 부분적으로 기여할 수 있다. 두 요인의 상대적 기여는 후속 연구에서 정밀하게 검토될 필요가 있다.

본 연구는 몇 가지 한계를 지닌다. 첫째, 단일 인간 코더를 기준으로 분석하여 인간-인간 간 일치도를 직접 검증하지 못하였으며, 인간 코더의 코딩 범위가 30편으로 한정되어 AI-AI 분석(n=100)과 동일한 수준의 차원별 인간-AI 비교가 불가능하였다. 이로 인해 인간-AI 간 일치도 패턴은 검증용 표본 30편을 대상으로 산출한 차원별 κ(〈그림 7〉)와 질적 사례 분석을 통해 보완적으로 제시하였으나, 표본 크기의 차이가 결과 해석에 영향을 미칠 수 있다. 아울러 초기 표집 프레임 구축 단계에서 단일 LLM에 의존하였다는 점도 방법론적 한계로 남는다. 후속 연구에서는 복수의 인간 코더를 활용한 인간-인간 간 일치도 검증과 함께 전면적인 인간-AI 비교 분석을 수행함으로써 본 연구의 결과를 보다 견고하게 검증할 필요가 있다. 둘째, 한국어 문헌정보학 논문에 한정되어 다른 언어권이나 학문 분야로의 일반화에 제약이 있다. 셋째, 프롬프트 설계가 차원별 결과에 미치는 영향을 충분히 검토하지 못하였다. 특히 평가적 판단 차원의 낮은 일치도는 코딩 루브릭의 조작화 수준에 부분적으로 기인할 가능성이 있으며, temperature=0 조건을 통해 출력의 결정론적 안정성을 확보하였음에도 모델별 프롬프트 해석 방식의 구조적 차이는 잔존할 수 있다. 이러한 프롬프트 민감도 문제는 후속 연구에서 체계적으로 검토될 필요가 있다. 넷째, LLM 기술의 급속한 발전으로 인해 결과의 시간적 일반화에 한계가 있으며, 모델 버전 갱신에 따른 반복 검증이 요구된다. 다섯째, Cohen's kappa는 범주 분포가 불균형한 경우 해석상 제약이 발생할 수 있으므로, 후속 연구에서는 Krippendorff's alpha 등 대안적 신뢰도 지표를 병행 활용할 필요가 있다. 여섯째, 본 연구는 한국어로 작성된 학술 논문을 분석 대상으로 하였으나, 사용된 4개 모델(Claude-3.5-Haiku, GPT-4o-Mini, Gemini-2.0-Flash, Grok-4-Latest) 간 한국어 처리

성능의 차이가 체계적으로 통제되지 않았다. 코딩 차원을 영문으로 설계하여 출력 형식의 언어적 조건을 통일하였으나, 한국어 입력 텍스트에 대한 각 모델의 의미 파악 수준은 모델별로 상이할 수 있다. 따라서 본 연구에서 관찰된 모델 간 코딩 결과의 차이는 코딩 과업의 판단 구조 차이에서 주로 기인하는 것으로 해석되나, 모델별 한국어 이해 능력의 차이가 결과에 부분적으로 기여하였을 가능성을 완전히 배제하기는 어렵다. 후속 연구에서는 동일 내용의 영문 번역본과의 비교 분석, 또는 각 모델의 한국어 벤치마크 성능을 통제변수로 설정하는 방식으로 이 문제를 보다 정밀하게 검토할 필요가 있다.

Ⅵ. 결 론

본 연구는 LLM 기반 내용 분석의 신뢰도가 코딩 과업의 판단 구조에 따라 체계적으로 달라짐을 실증적으로 확인하였다. AI-AI 간 일치도 분석, 인간-AI 간 일치도 분석, 그리고 질적 사례 분석을 종합한 결과, 명시적 정보 추출 과업에서는 인간과 AI 모두 판단이 안정적으로 수렴한 반면, 추론적 판단 과업에서는 AI의 문자적 해석과 인간의 맥락적 해석이 구조적으로 분기하였으며, 평가적 판단 과업에서는 객관적 기준의 부재로 인해 인간과 AI 모두에서 판단 불안정성이 발생하였다. 이러한 패턴은 LLM 기반 내용 분석의 성과가 특정 모델의 기술적 성능보다 과업이 요구하는 판단 구조와 정보의 명시성에 의해 일차적으로 규정됨을 시사한다. 아울러 표본 크기 정당화와 응답률 보고가 다수의 논문에서 누락되어 있음을 확인함으로써, LIS 연구 공동체의 표집 보고 관행 개선의 필요성도 제시하였다.

이러한 결과는 코딩 과업의 판단 구조를 기준으로 LLM 활용 범위를 단계적으로 설계해야 함을 시사한다. 명시적 분류 차원은 LLM 단독 처리로 충분하고, 추론적 코딩 차원은 LLM 초안을 인간이 검토하는 방식이 적절하며, 평가적 판단 차원은 인간 주도 분석이 요구된다. 다만 과업 구조 외에 모델 수준의 범주 귀속 기준 차이 역시 일치도 변동에 부분적으로 기여할 수 있으므로, 모델 선정 단계에서 차원별 범주 귀속 경향에 대한 사전 검토를 병행하는 것이 방법론적 엄밀성 확보를 위한 현실적 보완책이다. 본 연구는 과업 유형 및 판단 특성을 기준으로 LLM 활용 조건을 경험적으로 정립하였으며, 향후 복수 인간 코더를 활용한 human-human 일치도 검증과 전면적 human-AI 비교 분석을 통해 이 전략의 적용 가능성을 보다 견고하게 확장할 수 있을 것이다.

참 고 문 헌

김나연, 정은경 (2020). 사회과학 분야 연구자의 데이터요구와 데이터 재이용 행위에 관한 연구. 정보관리학회지, 37(4), 1-26. https://doi.org/10.3743/KOSIM.2020.37.4.001

송지애 (2021). 학교도서관의 블렌디드 러닝 운영에 관한 연구: 자유학기제 연계 독서동아리 프로그램을 중심으로, 55(2), 179-200. https://doi.org/10.4275/KSLIS.2021.55.2.179

이선애 (2023). 사서들의 감정노동과 전문직 삶의 질(ProQOL)에 미치는 영향관계에서 긍정심리자본의 매개효과에 관한 연구: 광역대표도서관 사서들을 대상으로. 한국비블리아학회지, 34(2), 251-274. https://doi.org/10.14699/kbiblia.2023.34.2.251

이선애 (2024). 대학도서관 사서 직업에 대한 경험적 의미와 딜레마: Giorgi의 현상학 방법을 적용하여. 정보관리학회지 41(2), 353-374. https://doi.org/10.3743/KOSIM.2024.41.2.353

정은경 (2021). 네트워크 분석 논문의 고찰: 계량서지적 분석과 내용 분석을 중심으로. 정보관리학회지, 38(1), 169-190.

최성호, 정정훈, 정상원 (2016). 질적 내용 분석의 개념과 절차. 질적탐구, 2(1), 127-155.

American Association for Public Opinion Research (2023). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys (10th ed.).

American Psychological Association. (2020). Publication Manual of the American Psychological Association (7th ed.).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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. M., 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, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), Advances in Neural Information Processing Systems (Vol. 33, pp. 1877-1901). https://doi.org/10.48550/arXiv.2005.14165

Cochran, W. G. (1977). Sampling Techniques (3rd ed.). New York, NY: John Wiley & Sons.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, &

T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Grimmer, J. & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, 21(3), 267-297. https://doi.org/10.1093/pan/mps028

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton, NJ: Princeton University Press.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Survey Methodology (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Kalton, G. (1983). Introduction to Survey Sampling. Beverly Hills, CA: Sage Publications.

Kim, K. (2021). Effects of knowledge of evidence-based practice and organizational culture on innovation behavior of university librarians. Journal of the Korean Library and Information Science Society, 52(1), 129-154.

Krippendorff, K. (2018). Content Analysis: An Introduction to Its Methodology (4th ed.). Thousand Oaks, CA: Sage Publications.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174. https://doi.org/10.2307/2529310

Lincoln, Y. S. & Guba, E. G. (1985). Naturalistic Inquiry. Beverly Hills, CA: Sage Publications.

Pluye, P., Gagnon, M. P., Griffiths, F., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative, and mixed methods primary studies in mixed studies reviews. International Journal of Nursing Studies, 46(4), 529-546. https://doi.org/10.1016/j.ijnurstu.2009.01.009

Törnberg, P. (2024). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. Social Science Computer Review. 43(6), 1181-1195. https://doi.org/10.1177/08944393241286471

Tracy, S. J. (2010). Qualitative quality: Eight “big-tent” criteria for excellent qualitative research. Qualitative Inquiry, 16(10), 837-851. https://doi.org/10.1177/1077800410383121

• 국한문 참고문헌의 영문 표기
(English translation / Romanization of references originally written in Korean)

Choi, Seongho, Jeong, Jeonghoon, & Jung, Sangwon (2016). Concept and procedures of qualitative content analysis. Qualitative Inquiry, 2(1), 127-155.

Chung, Eunkyung (2021). An investigation on the network analysis papers by content analysis and bibliometric analysis. Journal of the Korean Society for Information Management, 38(1), 169-190.

Kim, NaYon & Chung, Eunkyung (2020). An investigation on data needs and data reuse behavior in the field of social sciences. Journal of the Korean Society for Information Management, 37(4), 1-26. https://doi.org/10.3743/KOSIM.2020.37.4.001

Lee, Sunae (2023). A study on the mediating effect of positive psychological capital on the effectiveness relationship of librarians' emotional labor and professional quality of life (ProQOL): including widely representative library librarians. Journal of the Korean Biblia Society for Library and Information Science, 34(2), 251-274. https://doi.org/10.14699/kbiblia.2023.34.2.251

Lee, Sunae (2024). Empirical meanings and dilemmas of the profession of an academic librarian: applying Giorgi's phenomenological method. Journal of the Korean Society for Information Management, 41(2), 353-374. https://doi.org/10.3743/KOSIM.2024.41.2.353

Song, Jiae (2021). A study on the management of blended learning at school library: focusing on reading club program linked with free semester system. Journal of the Korean Library and Information Science Society, 55(2), 179-200. https://doi.org/10.4275/KSLIS.2021.55.2.179

[부록 1] 필터링 프롬프트 (설문·인터뷰 스크리닝용)

```
# 필터링 프롬프트: 설문·인터뷰 방법 사용 여부를 판별하기 위한 1차 스크리닝 도구
# 단일 LLM(Claude)을 사용하여 전체 1,150편 중 해당 연구 유형 576편을 선별하는 데 활용됨
# YES/NO 이진 판단 및 METHOD 유형 분류를 통해 표집 프레임 구축의 기초 자료로 사용됨

# PROMPT TEMPLATE: ──────────────────────────────────────────────
prompt = f"""
Determine if this research paper uses SURVEY or INTERVIEW methods to collect data.

Paper: {filename}
Text:
{pdf_text}

TASK: Answer ONE simple question:
Does this paper collect data using surveys (questionnaires) OR interviews?

SURVEY includes: questionnaires, online surveys, paper surveys, rating scales, structured instruments
INTERVIEW includes: face-to-face interviews, telephone interviews, focus groups, video interviews

Answer in this EXACT format:
ANSWER: [YES or NO]
METHOD: [SURVEY or INTERVIEW or BOTH or NONE]
BRIEF_REASON: [one sentence explaining why]

INSTRUCTIONS:
- Answer YES only if the paper clearly uses surveys or interviews for PRIMARY data collection
- Answer NO for literature reviews, system development, bibliometric analysis, or secondary data studies
- Base judgment on explicit evidence in the abstract and methods section
"""

# ──────────────────────────────────────────────────────────────#
프롬프트 설정 및 실행 조건:
# 1. 사용 모델: Claude 3.5 Haiku (claude-3-5-haiku-20241022)
# 2. 온도(Temperature): 0 (결정론적 출력)
# 3. 최대 토큰 수(Max Tokens): 256
# 4. 입력 범위: PDF 텍스트 중 선두 3,000자 (제목·초록·연구방법 섹션 중심)
# 5. 적용 대상:
```

[부록 2] 코딩 프롬프트 (표집 방법론 분석용)

```
# ============================================
# PROMPT FOR SAMPLING METHODOLOGY CODING
# Used in LLM-Based Content Analysis Study
# ============================================
PROMPT TEMPLATE:
```

Analyze the sampling methods in this LIS research paper. Code each dimension with ONE option from the list.
Paper: [FILENAME]
Text:
[PAPER TEXT - first 12,000 characters]

CODE THESE 12 FIELDS (select ONE option per field):
[Category A: 연구 설계 유형]
A_METHOD_TYPE (Is this a survey, interview, or both?)
Options: SURVEY | INTERVIEW | BOTH | OTHER | UNCLEAR
[Category B: 표집 구조]
B_SAMPLING_METHOD (How were participants selected?)
Options: PROBABILITY_RANDOM | PROBABILITY_STRATIFIED | CONVENIENCE | PURPOSIVE | SNOWBALL | VOLUNTEER | NOT_STATED | OTHER
B_SAMPLING_FRAME (What list/source was used?)
Options: INSTITUTIONAL_LIST | MEMBERSHIP_DATABASE | LIBRARY_USERS | PROFESSIONAL_NETWORK | NO_FRAME | NOT_STATED | OTHER
B_SAMPLE_SIZE_CATEGORY (How many participants?)
Options: VERY_SMALL_under10 | SMALL_10to30 | MEDIUM_31to100 | LARGE_101to300 | VERY_LARGE_over300 | NOT_STATED
B_RECRUITMENT_METHOD (How were participants recruited?)
Options: EMAIL | SOCIAL_MEDIA | IN_PERSON | GATEKEEPERS | MULTIPLE_METHODS | NOT_STATED | OTHER
[Category C: 표집 보고의 충실성]
C_SAMPLE_SIZE_JUSTIFICATION (Was sample size justified?)
Options: POWER_ANALYSIS | SATURATION | CITED_SIMILAR_STUDIES | PRACTICAL_CONSTRAINTS | NO_JUSTIFICATION | NOT_DISCUSSED
C_RESPONSE_RATE (For surveys: what % responded? For interviews: participation rate?)
Options: VERY_HIGH_over70pct | HIGH_60to70pct | MODERATE_40to59pct | LOW_20to39pct | VERY_LOW_under20pct | NOT_STATED | NOT_APPLICABLE
C_INCLUSION_CRITERIA (Were participant criteria stated?)
Options: CLEAR_STATED | VAGUE | IMPLIED | NOT_STATED

C_DEMOGRAPHICS_REPORTED (What demographics were reported?)
Options: COMPREHENSIVE_4plus | MODERATE_2to3 | MINIMAL_1 | NONE | NOT_APPLICABLE
[Category D: 평가 및 성찰]
D_SAMPLING_BIAS (What bias concerns exist?)
Options: CONVENIENCE_BIAS | SELF_SELECTION_BIAS | NONRESPONSE_BIAS | MULTIPLE_BIASES | LOW_BIAS | NOT_DISCUSSED
D_LIMITATIONS_DISCUSSED (Were sampling limitations discussed?)
Options: COMPREHENSIVE | MODERATE | BRIEF | NOT_DISCUSSED
D_OVERALL_QUALITY (Rate overall sampling quality 1-5)
Options: 1_VERY_POOR | 2_POOR | 3_ADEQUATE | 4_GOOD | 5_EXCELLENT

RESPONSE FORMAT (one option per line):
A_METHOD_TYPE: [one option]
B_SAMPLING_METHOD: [one option]
B_SAMPLING_FRAME: [one option]
B_SAMPLE_SIZE_CATEGORY: [one option]
B_RECRUITMENT_METHOD: [one option]
C_SAMPLE_SIZE_JUSTIFICATION: [one option]
C_RESPONSE_RATE: [one option]
C_INCLUSION_CRITERIA: [one option]
C_DEMOGRAPHICS_REPORTED: [one option]
D_SAMPLING_BIAS: [one option]
D_LIMITATIONS_DISCUSSED: [one option]
D_OVERALL_QUALITY: [one option]
INSTRUCTIONS:
- Select EXACTLY ONE option per field
- Use ONLY the options listed above
- If information is missing, use NOT_STATED or NOT_DISCUSSED
- Base coding on explicit evidence in the methods section

프롬프트 설정 및 실행 조건:
1. API 호출 버전 문자열: Claude-3.5-Haiku, GPT-4o-Mini, Gemini-2.0-Flash, Grok-4-Latest (호출시점: 2026년 2월)
2. 온도(Temperature): 0 (결정론적 출력)
3. 최대 토큰 수(Max Tokens): 1024
4. 입력 범위: PDF 텍스트 중 선두 12,000자
5. 각 모델의 공식 API를 통해 독립적으로 호출 (Anthropic API / OpenAI API / Google AI API / xAI API)
6. 텍스트 추출 도구: pdfplumber(주 사용), PyPDF2(보조)
7. 코딩 차원 수: 총 12개 (A 1개 + B 4개 + C 4개 + D 3개)