수집 실패는 무반응이 아니다 -
AI 기반 문화수용 연구에서 수용 가시성의 인프라적 선택성

남병수 (연세대학교 인문학연구원 산하 비교문학연구소, 전문연구원)

〈목 차〉

Ⅰ. 서론
Ⅱ. 플랫폼 인프라와 문화수용의 가시성
Ⅲ. 연구 설계와 감사 방법
Ⅳ. 감사 결과
Ⅴ. 수용 가시성의 인프라적 선택성
Ⅵ. 결론

국문초록

이 논문은 온라인 문화수용 자료의 수집 실패가 대중의 무반응으로 해석될 위험을 검토한다. 홍상수 영화 22편과 최인훈 소설 14편의 작품-출처 기록 1,872개를 대상으로, 수집과 판정 과정에서 무엇이 반응의 증거로 남는지 분석한다. 원시 347,992행의 중복 진단과 문맥·분류 조건별 집계를 구분하고, 파일의 보존과 수용 자료의 타당성이 별개의 검증임을 밝힌다. 이를 통해 접근 실패, 문맥 실패, 분류 실패, 한정된 범위에서의 미관찰을 구별할 것을 제안한다. '수용 가시성의 인프라적 선택성'은 플랫폼과 연구자의 규칙이 함께 관찰 범위를 구성한다는 뜻이다. 의의는 기술적 누락이 문화적 침묵의 판단으로 넘어가는 지점을 드러내고, 그 판단에 대한 인문학적 책임을 자료 감사의 절차로 구체화하는 데 있다.

키워드 : AI 기반 문화연구, 플랫폼 인프라, 수용 가시성, 데이터 감사, 실패 생태, 구성타당성, 온라인 수용

www.kci.go.kr

ⓒ 2026 이화인문과학원

Ⅰ. 서론

1. 수집량에서 가시성의 조건으로

온라인에서 한 작품에 관한 자료를 찾지 못했을 때, 연구자가 먼저 확인해야 할 것은 그 작품이 외면받았는지가 아니라 무엇을 어디까지 관찰했는지이다. 검색 결과와 상품 정보, 게시물과 댓글을 같은 표에 저장하면 행 수는 늘어난다. 그러나 이들은 서로 다른 행위의 흔적이다. 접속에 실패하여 빈 결과를 얻은 경우와, 접근 가능한 범위에서 작품에 관한 발화를 찾지 못한 경우 또한 구별되어야 한다. 본 논문은 이 차이가 수집과 분류의 과정에서 지워질 때 발생하는 해석상의 위험을 검토한다. 이 위험을 문화수용 연구 전반의 확립된 관행으로 전제하지 않고, 본 연구의 수집물과 판정 규칙에서 검증할 문제로 한정한다.

이 구별은 전처리의 정확성에만 걸린 문제가 아니다. 누군가의 말이 자료로 기록되지 않았다는 이유로 그 사람이 말하지 않았다고 판단한다면, 연구 장치의 한계가 문화적 주체의 속성으로 옮겨가기 때문이다. Star와 Ruhleder가 인프라를 실천 속에서 성립하는 관계로 파악한 논의는 이 전환을 살피게 한다.[1] Gillespie는 플랫폼이 콘텐츠의 접근과 발견 가능성, 수익화에 개입하는 양상을 논의했다.[2] 이 논의를 자료 수집에 연결하면 플랫폼의 공개 범위와 수집기의 작동 방식, 연구자가 정한 분류 규칙은 발화를 단순히 전달하는 데 그치지 않고 무엇을 수용의 증거로 읽을지에 관여한다. 따라서 본 논문이 묻는 인문학적 문제는 관찰되지 않은 경험에 어떤 판단을 부여할 수 있는가, 그리고 그 판단의 책임을 어디에 배정해야 하는가이다.

본 연구에서 '수용 가시성의 인프라적 선택성'은 어떤 반응이 연구자의

1) Susan Leigh Star·Karen Ruhleder (1996), "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces," *Information Systems Research*, 7(1), pp. 111~134, §1.1. 관계적 정의는 공개 저자 원고의 4~6쪽을 대조하였다.

2) Tarleton Gillespie (2010), "The Politics of 'Platforms'," *New Media & Society*, 12(3), pp. 358~359.

www.kci.go.kr

자료에 남는지가 플랫폼의 접근 조건과 연구자의 처리 규칙에 따라 달라진다는 뜻이다. 여기서 가시성은 화면에 보이는 정도가 아니라 분석 대상으로 기록될 가능성을 가리킨다. 이 개념은 선택의 원인을 플랫폼에만 돌리지 않는다. 무엇을 본문으로 추출하고 어느 문맥을 관련 자료로 인정할지 정하는 연구자 역시 관찰 범위의 형성에 참여한다. 그러므로 빈 결과를 기록할 때에는 제대로 접속하지 못했는지, 작품 관련성을 판정하지 못했는지, 분석 범주를 부여하지 못했는지, 아니면 정해진 관찰 범위에서 발화를 발견하지 못했는지를 함께 남겨야 한다.

본 연구는 홍상수 영화 22편과 최인훈 소설 14편에 관한 온라인 수집물을 대상으로, 작품의 수용을 관찰하려는 절차 자체를 감사한다. 분석 단위는 36개 작품과 52개 출처의 조합으로 구성된 기록 및 여기에서 보존된 347,992행이다. 더불어 파일과 집계가 보존되었다는 사실만으로 각 행을 관객이나 독자의 독립적인 발화로 승인할 수 있는지 또한 검토한다. 따라서 결과는 수집물의 구성과 판정 조건에 관한 것이다. 두 작가군의 인기나 감정 분포, 작품 형식이 수용에 미치는 인과적 영향은 이 논문의 판단 범위에 포함하지 않는다.

2. 연구 질문과 논문의 차별성

연구 질문은 총 네 가지이다. 첫째, 수집된 행은 정확히 같은 텍스트의 반복이라는 진단과 작품 문맥·분류 조건에 관한 진단에서 각각 어떤 구성을 보이는가. 둘째, 행을 얻지 못했거나 판정을 내리지 못한 상태를 어떤 기록에 근거하여 구분할 수 있는가. 셋째, 이 구분은 관찰되지 않은 수용을 해석하는 연구자의 책임을 어떻게 구체화하는가. 넷째, 자료의 지시 대상을 판별하는 불확실성과 작품 경험의 최적 모호성을 구별할 때, 문화수용의 인지적 해석에 앞서 충족되어야 할 증거 조건은 무엇인가.

이 질문들을 관통하는 관류 명제는 수집 장치의 산출물과 문화수용의 증거 사이에 별도의 정당화 과정이 필요하다는 것이다. 반복 여부, 작품 관련성, 문서 장르, 작성 주체는 서로 다른 판정이다. 파일을 동일하게 보

www.kci.go.kr

존했다는 확인은 이 판정들을 대신하지 않는다. 본 논문은 관찰의 실패가 수용자의 침묵으로 전가되지 않도록, 각 단계가 허용하는 판단을 구분하는 절차를 제시한다.

더 중요한 지점은 홍상수와 최인훈의 작품이 미학적 우열을 가리기 위한 비교쌍으로 배치되지 않았다는 점이다. 본 연구는 원장(기존 수집 기록)에 포함된 두 작품군을 제한된 감사 사례로 제시한다. 예술가별로 할당된 작품 수가 서로 다르고 또 전체 창작 목록을 포괄하지 않으므로, 결과를 두 작가나 두 매체의 대표적 수용 양상으로 일반화할 수 없다. 다만 영화의 정보·영상 유통 경로와 소설의 판본·판매·독서 기록이 같은 수집표로 옮겨질 때 생기는 차이는 구체적으로 물을 수 있다. 작품명과 수집 경로, 출처 유형을 함께 제시하는 까닭도 여기에 있다. 다만 수집 목표명만으로 소설의 특정 판본을 확정할 수는 없으므로 판본 식별은 추가 확인 과제로 남긴다. 작품의 고유성은 숫자의 뒤안길로 사라져도 되는 잉여가 아니라, 어떤 자료를 해당 작품에 관한 반응으로 인정할지 판단하는 근거로 봉사한다.

디지털 인문학이 자료의 기계 가독성과 분석 가능성을 확장해 온 만큼, 자료가 무엇을 대표하는지 묻는 데이터 리터러시가 함께 요구된다. 이재연은 한국현대문학 분야의 디지털 인문학 수용 과정과 연구 방법의 재편, 연구·교육의 과제를 검토했고, 한상우는 인문학 기반 데이터 리터러시에 데이터의 비판적 이해와 활용 역량을 포함시켰다. 본 연구는 일련의 논의를 '수집 이전과 이후의 감사 가능성'으로 구체화한다. 이는 단지 데이터를 많이 모은 뒤 분석하는 순서를 따라 진행되는 것이 아니라, 무엇이 어떤 조건에서 보이도록 가시화되었는지를 먼저 기록하고 그 조건이 확보된 범위에서만 분석의 권리를 주장하는 순서를 따른다.[3)]

3) 이재연 (2018), 「디지털 시대의 인문학에서 디지털 인문학 시대로 - 한국문학에서 본 디지털 인문학 연구」, 『역사학보』, 240, 158~159, 183~185쪽; 한상우 (2020), 「인문학 기반 데이터 리터러시 모형 설계에 관한 연구」, 『정보관리학회지』, 37(1), 188~191쪽.

www.kci.go.kr

II. 플랫폼 인프라와 문화수용의 가시성

1. 관계적 인프라와 분류의 정치

인프라는 데이터가 이동하는 수동적 배경이 아니다. 검색 API의 제공 여부, 요청량 제한, 사용자 에이전트 정책, 자바스크립트 렌더링, 로그인 벽, 국가별 접속 조건, 결과 정렬, 페이지네이션은 무엇이 연구 가능한 객체가 되는지를 선별한다. Plantin 등은 Google과 Facebook 등을 통해 일부 플랫폼이 인프라적 지위를 획득하고 기존 인프라가 플랫폼화되는 상호 전환을 논의하면서, 플랫폼의 경제적 논리를 인프라의 공공적 역할과 혼동하지 말아야 한다고 지적했다.[4] van Dijck·Poell·de Waal은 플랫폼 메커니즘을 데이터화·상품화·선별로 구분한다. 본 연구는 이 구분을 공적 가치와 자료 가시성의 관계를 검토하는 틀로 사용한다.[5] 국내에서도 김홍중은 플랫폼 자본주의와 알고리즘 통치성을 함께 고찰하며 플랫폼을 사회관계의 조직 원리로 다룬 바 있다.[6]

문화수용 자료가 만들어지는 과정은 접근, 표현, 판정의 세 층으로 나누어 살필 수 있다. 접근 층에서는 플랫폼이 공개하는 범위와 질의 방식이 자료의 입구를 정한다. 표현 층에서는 수집기가 페이지의 여러 요소 중 무엇을 한 행의 텍스트로 옮길지 결정한다. 판정 층에서는 연구자가 작품 관련성과 분석 범주를 정한다. 같은 작품명이 들어 있어도 상품 정보와 독자의 해석은 서로 다른 문서 장르이며, 이 차이가 행으로 변환되는 과정에서 보존되었는지가 중요하다.

Bowker와 Star의 분류 논의는 세 번째 층을 이해하는 데 도움을 준

4) Jean-Christophe Plantin, Carl Lagoze, Paul N. Edwards, Christian Sandvig (2018), "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook," *New Media & Society*, 20(1), pp. 293~310. 관련 논의는 공개 저자 원고(2016), pp. 19~20(원고 자체 쪽수) 참조.

5) José van Dijck, Thomas Poell, Martijn de Waal (2018), *The Platform Society: Public Values in a Connective World*, Oxford University Press, pp. 4~5.

6) 김홍중 (2022), 「플랫폼의 사회이론: 플랫폼 자본주의와 알고리즘 통치성을 중심으로」, 『사회와 이론』, 41, 17~19, 29~30쪽.

www.kci.go.kr

다. 분류 체계는 자료를 정돈하면서 동시에 체계 안에 들어오지 않는 경험을 남긴다.[7] 이를 본 연구에 적용하면, '분류 가능'이라는 값은 수용자의 태도 자체보다 텍스트와 판정 규칙의 관계를 먼저 보여 준다. 연구자의 범주에 들어오지 않은 발화를 중립이나 무반응으로 옮기지 않고 별도 상태로 보존하는 일은, 해석되지 않은 경험이 있었다는 가능성을 남겨 두는 조치이다. 본 논문의 자료 감사는 바로 이 가능성을 계산 과정에서 지우지 않기 위한 실천이다.

소셜 빅데이터 연구는 오래전부터 대표성과 타당성의 함정을 경고해 왔다. boyd와 Crawford는 크기가 데이터의 객관성이나 전체성을 보장하지 않는다고 지적했고, Tufekci는 플랫폼이 제공하는 표본과 연구자가 원하는 모집단 사이의 간극을 강조했다.[8] Ruths와 Pfeffer는 플랫폼별 인구 구성, 행동, 인터페이스와 데이터 접근 방식이 관찰 결과를 편향할 수 있음을 보였고, Olteanu 등은 사회 데이터의 편향이 수집·처리·분석의 각 단계에서 발생할 수 있다고 정리했다.[9] 본 연구는 이 경고를 작품–출처 기록 및 서로 다른 단위의 감사 집계로 번역한다.

2. 자료 식별의 불확실성과 해석의 조건

자료를 분류하기 전에 해소해야 할 불확실성과 작품을 경험하며 유지할 수 있는 모호성은 서로 다른 층위에 있다. 전자는 이 문서가 어느 작품을

7) Geoffrey C. Bowker, Susan Leigh Star (1999), *Sorting Things Out: Classification and Its Consequences*, MIT Press, pp. 5~6.

8) danah boyd, Kate Crawford (2012), "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society*, 15(5), pp. 666~669; Zeynep Tufekci (2014), "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls," *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pp. 506~509.

9) Derek Ruths, Jürgen Pfeffer (2014), "Social Media for Large Studies of Behavior," *Science*, 346(6213), pp. 1063~1064; Alexandra Olteanu, Carlos Castillo, Fernando Diaz, Emre Kıcıman (2019), "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *Frontiers in Big Data*, 2, Article 13, §2.3 및 §§5~7.

www.kci.go.kr

가리키는지, 누가 어떤 종류의 말을 남겼는지에 관한 식별 문제이다. 작품명이 본문에 포함되더라도 상품 소개인지 독자의 감상인지에 따라 증거의 지위가 달라진다. 반대로 제목이 생략된 실제 감상은 정확 제목 규칙을 통과하지 못할 수 있다. 따라서 본 연구의 문맥 검사를 작품 경험의 모호성을 측정하는 절차로 해석해서는 안 된다.

최적 모호성은 이러한 구별 이후에 제기할 수 있는 후속 질문으로 남는다. 이 개념은 단지 중립이나 중간량을 지시하는 용어가 아니라, 미결정성이 유지되면서도 이해의 재구성이 가능해지는 관계론적인 조건을 가리킨다. 예측처리와 미학의 접점을 다루는 연구들은 이러한 질문을 조직하는 이론적 자원이 될 수 있지만,[10] 현재의 수집물만으로 감상자의 인지 변화나 그 최적점을 확인할 수는 없다. 본 논문은 이 가능성을 폐기하는 대신, 그 가능성에 관한 논증이 어떤 자료에서 출발해야 하는지를 묻는다.

그 선행조건은 발화의 대상을 식별하고 문서 장르와 작성 주체를 확인하는 것이다. 수집기의 누락을 미학적 난해함에 대한 회피로 설명하거나, 분류값의 부재를 이해의 중단으로 읽는 순간 기술적 상태와 경험적 상태가 뒤섞인다. 본 논문이 자료의 계보와 판정 조건을 전면에 두는 이유는, 이후의 수용 해석이 플랫폼이나 분류기의 산출을 사람의 경험으로 오인하지 않도록 하기 위해서이다.

Ⅲ. 연구 설계와 감사 방법

1. 작품–출처 계획 행렬과 데이터 계보

연구 대상은 홍상수 영화 22편과 최인훈 소설 14편이다. 각 작품에 52

10) Jacopo Frascaroli, Helmut Leder, Elvira Brattico, Sander Van de Cruys (2024), “Aesthetics and Predictive Processing: Grounds and Prospects of a Fruitful Encounter,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1895), Article 20220410, §2.

www.kci.go.kr

개 수집 경로를 배정하여 1,872개의 작품–출처 셀을 구성했다. 영화 1,144 셀과 소설 728셀은 계획 행렬의 크기이며 실제 HTTP 요청 횟수가 아니다. 출처 시도 카운터의 합 4,856도 모든 네트워크 요청의 수로 확인된 값은 아니다. 매체 구분은 자료 계보를 설명하는 변수로 사용한다.

(표 1) 감사 대상과 원시 산출물

구분	작품 수	출처 수	계획된 작품–출처 셀	생산적 출처 인스턴스	원시 행
영화 질의 집합	22	52	1,144	265	245,613
소설 질의 집합	14	52	728	112	102,379
합계	36	52	1,872	377	347,992

보충표 S-1은 기존 자료 수집 기록표에 등재된 36개 작품의 집계를, 보충표 S-2는 52개 수집 경로의 목록과 산출량을 제시한다. 두 표는 부록 D에 수록하였으며, 대응하는 전체 열과 파일 식별정보는 별도 데이터 보충자료의 table_s1_works.csv와 table_s2_sources.csv에서 확인할 수 있다. 작품명은 수집 목표명이며 판본의 출판연도나 문학사적 대표성을 보증하지 않는다. 분석 대상은 이 기록표의 작품과 파일로 한정하며, 내부 작품 분석의 수치를 온라인 수용의 타당성 근거로 사용하지 않는다.

생산적 출처 인스턴스 377개는 분석 행이 하나 이상 남은 작품–출처 셀로, 계획 행렬의 20.14%이다. 원장에는 성공 243개와 필터 후 부분 성공 134개 외에 검색결과 부재·제공자 오류가 결합된 상태 1,444개, 매체 적용 불가 50개, 수집 후 길이 등의 조건으로 제외된 1개가 있다. 비생산적 1,495셀을 모두 접근 실패나 무반응으로 분류할 수 없다. 특히 매체 적용 불가는 요청 실패와 별도 상태로 남겨야 한다.

보존된 온라인 파일의 동일성과 각 행의 해석 가능성은 구분하여 점검한다. SHA-256은 파일의 바이트열로부터 계산하는 식별값이며, 같은 파

www.kci.go.kr

일의 기준값과 재계산값을 비교하여 변경 여부를 확인하는 데 사용한다. 이 검사는 텍스트 추출의 정확성이나 작품 관련성, 수용 발화 여부를 판정하지 않는다. 또한 영상 원본과 추출 프레임, 소설 파일과 분절 문장은 서로 다른 파일이므로 그 해시가 서로 같아야 하는 관계도 아니다. 내부 작품 코퍼스의 처리 기록은 온라인 자료를 수용의 증거로 승인하는 근거와 분리한다. 온라인 번들의 무결성 결과를 보고할 때에는 비교한 파일의 범위와 기준 해시의 생성 시점을 함께 명시한다. 파일의 보존 상태는 기존 자료 수집 기록표에 등재된 36개 작품의 파일 묶음을 대상으로 확인하였다. 각 파일 묶음을 검증 범위로 고정하고 작품별 원시·출처·요약 CSV 세 개의 SHA-256을 기존 기록과 비교하여 총 108개 파일의 일치를 확인했다. 이 결과는 온라인 파일의 보존 상태에 한정된다.

이 글에서 번들(bundle)은 한 작품에 연결된 원시·출처·요약 파일 묶음을 뜻한다. 체크섬(checksum)은 파일의 변경 여부를 대조하는 검사값이며, 여기서는 SHA-256을 사용한다. 데이터 스키마(data schema)는 표에 필요한 열과 각 값의 형식을 정한 구조이다. 이러한 구조와 파일이 보존되었다는 확인을, 기록된 문장이 실제 수용 발화라는 내용 판단과 구별하여 사용한다.

정규화된 텍스트의 반복은 별도의 진단 대상으로 삼는다. 원고에 보고된 중복 집계는 서로 다른 텍스트 102,189개와 첫 출현을 초과한 반복 행 245,803개로 구성된다. 둘 이상의 작품에 재등장한 텍스트 19,914개는 265,716행을 차지한다. 이 수치는 행 수를 독립적인 수용 발화 수로 곧바로 사용할 수 없음을 보여 준다. 다만 같은 짧은 문장을 서로 다른 사람이 남겼거나 하나의 문서가 여러 작품을 다루었을 가능성도 있으므로, 정확한 텍스트 일치만으로 개별 행의 오문맥 여부를 확정할 수는 없다. 반복의 수량과 반복된 내용의 성격은 각각 확인해야 한다.

자료 감사는 세 갈래로 구분한다. 첫째는 전체 원시 행에서 정규화된 텍스트의 정확 반복을 확인하는 중복 진단이다. 둘째는 원시 행에 제목·문맥·길이 조건을 적용하고, 그 통과 항목에서 라벨 부여 여부를 확인하

www.kci.go.kr

는 진단이다. 셋째는 특정 출처군과 제목·문맥·길이 조건을 통과한 행에서 점수와 출처의 고유 조합을 세는 별도 진단이다. 서로 다른 텍스트 102,189개, 문맥 통과 22,831행, 분류 가능 6,089행, 점수 조합 3,428개를 하나의 연속적인 축소 과정으로 배열해서는 안 된다.

2. 감사 단계와 실패 생태의 분류

기존 수집물에는 작품 관련성을 확정하지 않은 상태에서 저장된 항목이 포함되어 있다. 이를 진단하려면 판정 규칙의 통과 수와 실제 내용 점검을 구분하여 보고해야 한다. 자동 규칙의 통과율은 그 규칙이 어떤 자료를 남겼는지를 나타내며, 문서 장르별 오류율을 직접 제공하지 않는다. 따라서 본 연구는 실제 사례를 제시하여 오문맥과 비발화 문서가 생기는 경로를 설명하되, 표본 크기와 추출 절차가 확인되지 않은 과거 열람 기록을 근거로 전체 수집물의 오류 비율을 추정하지 않는다. 체계적인 수동 코딩이 추가되기 전까지 일반화할 수 있는 결론은 원시 행 수만으로 수용 발화의 규모를 확정할 수 없다는 데 한정된다.

보충표 S-3은 강변호텔, 광장, 라울전, 생활의 발견의 네 파일에서 규칙의 통과·탈락과 점수 조합의 대표행을 대조한 12개 사례이다. 온라인 댓글의 원문을 전부 재배포하지 않고 판정에 필요한 짧은 부분만 제시한다. A는 무관 문서, B는 제목을 포함한 비수용 문서, C는 문맥 통과 미분류, D는 별도 점수 조합에 남은 대표행을 가리킨다. 이들은 배타적 범주가 아니며 같은 행에 복수 특성이 있을 수 있다. 사례 식별자는 보충 감사 기록의 파일·행·해시와 연결된다.

라울전 사례는 중요한 경계를 드러낸다. 영상 제목 끝의 ‘라울’과 댓글 첫머리의 ‘전설’을 붙인 결과가 목표명과 일치해도, 두 필드가 최인훈의 작품을 지시했다는 뜻은 아니다. 또 검색결과에 만족하는지를 묻는 사이트 문구는 사전의 긍정 라벨을 얻지만 독자의 긍정적 감상은 아니다. 여기서 인문학적 문맥 판단은 계산 이후의 장식적 해설이 아니라, 무엇을 작품에 관한 발화로 셀 수 있는지 되묻는 근거가 된다.

www.kci.go.kr

사례별 짧은 인용과 판정 내용은 부록 D의 보충표 S-3에 제시한다.

사례 분석은 보존된 네 파일에서 조건별 선두 후보와 점수 조합의 대표행을 찾아 12행을 목적표집하여 규칙의 작동을 설명하는 방식으로 수행하였다. 이 절차는 사례의 존재와 오류 경로를 확인하기 위한 것이며 전체 347,992행의 오류율을 추정하지 않는다. 별도로 비어 있지 않은 35개 파일에서 시작부터 끝까지 균등한 위치의 10 행씩, 총 350행을 골라 저장된 여덟 라벨과 보존 함수의 재계산값 2,800개를 비교하여 일치를 확인했다. 이 검사는 코드와 저장 출력의 일관성 검사이지, 사람의 정답에 대한 분류 정확도 검사나 코더 간 신뢰도 검사가 아니다. 과거 수동 점검의 표본 크기와 추출법, 독립 코더의 불일치 조정 기록은 확인하지 못했다.

여기서 원시 행은 웹 응답 전체가 아니라 수집기 내부의 전처리와 병합 이후 보존된 분석 CSV의 행이다. 분류 가능은 문맥 조건을 통과한 행에 사전 기반 라벨이 하나 이상 부여된 상태를 뜻한다. 반면 기존 작업에서 잠정 수용 출처 잔여물이라 부른 값은 분류 가능 조건을 적용하지 않은 별도 분기의 점수 조합 수이다. 본고에서는 이 값을 '출처 제한 점수 조합'으로 명명한다. 이는 사람의 발화 수나 후속 검증을 마친 반응 표본의 수가 아니다.

전역 중복 진단에서는 text의 연속 공백을 하나로 줄이고 양끝 공백을 제거한 뒤 대소문자를 유지하여 같은 문자열을 비교했다. 작품·출처·URL·작성일은 이 중복 키에 포함하지 않았다. 따라서 고유 텍스트는 문서나 작성자의 고유성에 관한 판정이 아니다.

제목·문맥 규칙은 목표 작품명과 title, text, url의 결합 문자열을 각각 정규화한다. 이때 숫자·영문·한글 이외의 문자를 제거하고 대소문자를 통일한다. 정규화된 목표명이 결합 문자열 안에 있고, 양끝 공백을 제거한 text의 길이가 15자 이상 5,000자 이하이면 통과한다. 별칭·번역 제목·어절 경계·작성 주체·문서 장르를 판정하는 규칙은 아니다. 필드 경계를 지우는 과정에서 작품명과 같은 문자열이 새로 생길 수도 있다.

www.kci.go.kr

출처 제한 점수 조합은 원시 자료에서 YouTube, ExtremeMovie, WatchaPedia, Goodreads, Letterboxd 등에 해당하는 출처와 제목·길이 조건을 적용한 별도 분기이다. 이 분기는 길이 계산 전 양끝 공백을 제거하지 않는 구현 차이도 있다. 여기서 얻은 8,699행 중 분류 가능 행은 2,477개, 전부 미분류인 행은 6,222개다. 기존 계산은 분류 가능 조건을 적용하지 않고 모호성 3점수와 해석 지속·가설 생성·메타인지·기억 잔류 4점수, source의 조합에 대해 작품 내부 중복을 제거하여 3,428개를 산출했다. 텍스트와 항목 식별자가 중복 키에서 빠져 있으므로, 이 값은 발화 수로 해석할 수 없다.

가동된 구형 수집기는 공식 YouTube API의 JSON 응답과 공개 웹·RSS·검색 경로를 함께 사용했다. 공개 웹 문단 추출은 urllib.request로 응답을 읽고 BeautifulSoup의 html.parser로 script, style, nav, header, footer, form 요소를 제거한 뒤 p, li, blockquote를 추출하는 방식이다. 해당 라이브러리를 사용할 수 없으면 정규식 기반 제거로 대체한다. 수집단계의 25~3,000자 문단 조건과 감사단계의 15~5,000자 조건은 구별한다.

라벨은 한·영 수작업 사전의 문자열 적중으로 부여한다. 소문자화와 문자 정리 후 사전 어휘의 문자열 포함 또는 토큰 일치를 세고, 적중 수를 max(1, sqrt(토큰 수+1))로 나눈다. 가족별 최대 점수가 0보다 크면 해당 라벨을, 모두 0이면 미분류를 부여한다. 동점은 사전 선언 순서를 따른다. 여덟 가족은 감정, 안정, 무시·모호·회피, 모호성 수준, 쾌·혼란·불쾌, 질문·당혹·충격, 발화·침묵·언어상실, 지루함·주의·이탈이다. 이는 부정이나 풍자, 인용 주체를 해석하는 학습모델이 아니다. 세부 사전과 함수는 보충기록에 보존한다.

집계의 일관성은 작품별 수치와 비율의 분모를 대조하여 확인하였다. 36개 작품의 네 집계 항목에서 144개 값의 일치와 분모표 10행의 산술적 정합성을 확인하고, 사례 12행의 식별정보·판정·원문 검사값 및 배포 파일의 무결성을 비교하였다. 이 절차는 결과를 전달하는 표와 파일의 일관성 검사이며, 원시 자료 전체의 재판정이나 분류의 의미적 정확도 측정

www.kci.go.kr

과는 구별된다. 원자료 감사와 집계·배포 파일 검사의 범위는 별도 검증 기록에 각각 제시한다.

(표 2) 감사 집계의 정의와 해석 범위

집계	수	원시 행 대비	운영적 의미	금지되는 해석
원시 분석 행	347,992	100.00%	수집기 전처리·병합 후 보존 행	대중 반응 전체
고유 텍스트	102,189	29.37%	공백 정규화 후 서로 다른 텍스트	고유 문서·이용자 수
제목·문맥 통과	22,831	6.56%	결합 문자열과 길이 규칙 통과 행	의미적으로 검증된 관련 발화
분류 가능	6,089	1.75%	문맥 통과 행 중 라벨 1개 이상 부여	태도 정답·대표 표본
출처 제한 점수 조합	3,428	0.99% (단위 상이)	별도 출처군에서 점수7열+출처의 작품 내 고유 조합	분류 가능 행의 부분집합·발화 수

주: 비율은 표시한 집계 수/원시 행 수의 규모비이다. 점수 조합은 행과 다른 단위이므로 전체 원시 행 중 유효 반응이 차지하는 비율로 해석하지 않는다. 세 진단은 하나의 완전 포함관계가 아니다.

고유 텍스트와 문맥 통과는 서로 다른 감사축이며, 출처 제한 점수 조합도 문맥·분류의 순차적 잔존량과 구별된다. 정확 반복 진단은 텍스트의 같은 값을 세고, 문맥 규칙은 작품명과 길이 조건을 검사하며, 점수 조합 진단은 텍스트가 빠진 수치 조합을 센다. 어느 집계도 작성자의 독립성이나 문서의 수용적 성격을 자동으로 복원하지 않는다. 작품 간에 반복된 19,914개 텍

www.kci.go.kr

스트가 265,716행을 차지한다는 사실은 독립성 검증이 필요함을 보여 주지만, 반복 원인별 비율이나 오문맥 비율을 대신하지는 않는다.

실패 상태는 기록의 단위를 나누어 부여한다. 접근 실패는 요청·응답 또는 작품–출처 조합의 상태이며, 문맥 실패와 분류 실패는 수집된 행의 판정 상태이다. 한 작품–출처 조합 안에서도 일부 행은 통과하고 다른 행은 탈락할 수 있으므로 네 범주를 같은 분모의 배타적인 비율로 합산하지 않는다. ‘관찰된 무반응’은 접근과 관찰 범위가 확인된 경우에 한하여 사용할 수 있는 보고 범주이다. 본고의 현재 감사 기록만으로는 그 조건을 충족한 0건 사례를 확정하지 못했으므로, 이 범주는 실제 관찰 결과가 아니라 관찰 실패와 구별하기 위한 보고 기준으로 제시한다.

(표 3) 관찰 상태의 적용 단위와 판단 범위

구분	적용 단위	판정에 필요한 기록	현재 자료에서 허용되는 해석
접근 실패	요청·응답 또는 작품과 수집 경로의 조합	접속 결과, 오류 사유와 관찰 기간을 확인한다.	원장의 검색결과 부재와 제공자 오류가 결합된 상태를 모두 접근 실패로 확정하지 않는다.
문맥 실패	보존된 개별 행	목표 작품명, 제목·본문·주소의 결합, 길이 조건과 탈락 사유를 확인한다.	운영 규칙의 탈락을 뜻하며, 실제로 작품과 무관한 문서라는 최종 판정과 구별한다.
분류 실패	문맥 조건을 통과한 개별 행	어휘 사전과 점수 산식, 여덟 분류 범주의 미분류 여부를 확인한다.	현재 범주를 부여하지 못한 상태이며, 수용자의 무관심이나 인지적 회피로 해석하지 않는다.
한정된 범위에서 관찰된 무반응	관찰 범위가 확인된 작품과 수집 경로의 조합	접근 성공, 관찰 기간과 범위, 발화 판정 기준 및 그 범위에서의 0건 기록을 확인한다.	현재 기록만으로 확정한 사례는 없다. 관찰 실패와 구별하기 위한 보고 기준으로 남긴다.

www.kci.go.kr

IV. 감사 결과

1. 중복·문맥·분류 및 점수 조합의 구별

첫 번째 결과는 규모의 붕괴가 아니라 측정 대상의 교체로 수렴한다. 원시 347,992행에서 정확 정규화 텍스트 초과 중복 245,803행을 제외하면 102,189개의 서로 다른 텍스트가 남는다. 원시 행의 70.63%가 적어도 한 번 앞서 나온 텍스트의 반복이라는 사실은 행 단위 빈도가 발화 빈도나 이용자 빈도를 대신할 수 없음을 보여 준다. 더 나아가 작품 경계를 넘어 재등장한 정확 텍스트 19,914개가 265,716행을 구성했다. 반복은 동일 페이지의 여러 요소, 페이지네이션, 재수집, 공통 문구, 다수 상품 페이지의 템플릿에서 생길 수 있다. 원인별 중복률이 확정되지 않은 상태에서 가중치나 표본 수로 사용할 수 없다.

정확 제목과 문맥 조건을 통과한 자료는 22,831행으로, 보고된 원시 행의 6.56%이다. 325,161행이 이 규칙을 통과하지 못했다. 이는 판정 기준을 바꾸었을 때 관찰 범위가 크게 달라진다는 결과이다. 탈락 행에는 실제 오문맥 문서뿐 아니라 별칭, 번역명, 등장인물이나 장면으로 작품을 지시한 발화도 포함될 수 있다. 따라서 이 비율을 기존 규칙의 오분류율 또는 새 규칙의 정밀도로 보고하지 않는다. 어느 경우가 얼마나 포함되었는지는 통과군과 탈락군을 함께 검토하는 별도의 내용 코딩으로 확인해야 한다.

세 번째 결과는 관련 문맥과 분석 가능성 사이의 추가 간극이다. 문맥 통과 자료 중 6,089건(26.67%)만 설정된 분류 조건을 만족했다. 이는 원시 행의 1.75%이다. 나머지 16,742건을 중립 또는 무반응으로 바꿀 수 없다. 정보 페이지, 뉴스 헤드라인, 상품 설명, 매우 짧은 언급, 라벨 체계 밖의 복합적 발화가 포함될 수 있기 때문이다. 분류기의 출력이 없는 것은 사람의 태도가 없는 것이 아니라 도구와 범주의 관계가 성립하지 않은 상태일 수 있다.

출처 제한 점수 조합은 분류 가능 행에서 이어지는 축소 단계가 아니라 별도의 집계 분기이다. 특정 출처군과 제목·길이 조건을 통과한 8,699

www.kci.go.kr

행에서 일곱 점수와 출처의 작품 내 고유 조합을 세면 3,428개가 된다. 이 분기는 분류 가능 조건을 적용하지 않으며, 입력 8,699행 중 6,222행은 여덟 라벨 모두 미분류이다. 따라서 3,428/6,089는 출처 필터의 잔존율을 나타내지 않는다. 동일한 점수 조합을 지닌 서로 다른 문서가 하나로 합쳐질 수 있으므로 3,428개를 발화 수로 읽을 수도 없다. 계산값의 일치가 그 단위를 수용 발화로 해석할 근거까지 제공하는 것은 아니라는 점이 이 집계에서 드러난다.

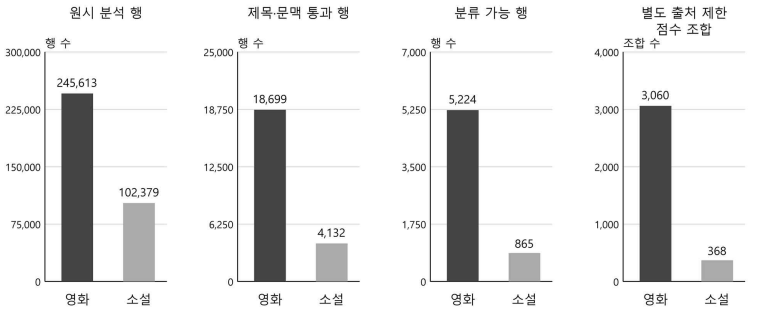
(표 4) 감사 집계의 비율과 적용 범위

항목	분자/분모	비율	의미
고유 텍스트/원시 행	102,189/347,992	29.37%	정확 반복의 전역 진단
문맥 통과/원시 행	22,831/347,992	6.56%	특정 제목·길이 규칙 통과
분류 가능/문맥 통과	6,089/22,831	26.67%	사전 라벨 부여 여부
출처 제한 점수 조합	별도 분기의 8,699행 → 3,428조합	잔존율 미부여	작품별로 일곱 점수와 출처가 같은 조합을 하나로 센 값이며, 발화 수를 뜻하지 않는다.

주: 앞의 세 비율만 동일 단위 또는 텍스트 중복 진단으로 해석한다. 마지막 행은 기존 계산의 정체를 공개하기 위한 별도 집계이며 잔존율을 부여하지 않는다.

www.kci.go.kr

(도판 1) 서로 다른 감사 분기의 매체별 집계



패널별 선형 축의 범위가 다름. 마지막 패널은 후속 단계가 아니며 단위도 다름.
전역 고유 텍스트 102,189개는 별도의 중복 진단으로, 위 패널에 포함하지 않음.

주: 원시 행, 문맥 통과 행, 분류 가능 행은 행 단위로 표시한다. 출처 제한 점수 조합은 별도 패널에 제시하며 앞 패널의 후속 단계로 연결하지 않는다. 세로축은 각 패널의 건수를 표시하는 선형 축이다. 전역 고유 텍스트 수는 표 2의 별도 진단이며 이 도판의 단계에 포함하지 않는다.

매체별로 영화 질의의 원시 행은 245,613개, 문맥 통과는 18,699행, 분류 가능은 5,224행이며, 소설 질의는 각각 102,379개, 4,132행, 865행이었다. 별도 출처 제한 점수 조합은 영화 3,060개와 소설 368개다. 이 마지막 값은 앞의 행 집계와 같은 단위로 비교하지 않는다. 작품 수도 22편과 14편으로 다르고, 그뿐만 아니라 각 경로의 지원 범위와 제목의 중의성, 반환되는 문서의 장르 역시 다르다. 따라서 이 도판은 두 매체의 수용량이나 감정 분포를 비교하는 것이 아니라, 같은 형식의 수집표가 서로 다른 자료 구성을 남겼음을 보여 주는 진단이다.

온라인 자료 자체에서도 파일의 보존과 내용의 타당성은 별개의 질문으로 남는다. 과거 원장의 36개 번들에 포함된 108개 CSV는 기존 체크섬과 일치했고, 원시 행·고유 텍스트·문맥 통과·분류 가능·출처 제한 점수 조합의 다섯 집계도 재현되었다. 그러나 집계의 재현은 그 이름으로 부른

www.kci.go.kr

대상이 실제 무엇이었는지까지 보증하지 않는다. 점수 조합 3,428개를 발화의 수로 읽거나 서로 다른 분기를 하나의 잔존 과정으로 읽으면, 계산은 같아도 해석은 잘못될 수 있다. Lazer 등이 논의한 자료 접근·공유·프라이버시·연구윤리의 과제에 더하여,[11] 본 사례는 기록의 단위와 주장 사이의 대응을 별도로 감사해야 할 필요를 보여 준다.

2. 접근 기록과 무반응 판단의 경계

52개 출처 중 분석 행이 확인된 출처는 일부에 집중되었다. 알라딘 공개 웹, YES24 공개 웹, 유튜브, 리디 공개 웹, 워드프레스 공개 웹, 익스트림무비 공개 웹, Bing News RSS, 씨네21 공개 웹, TMDb 공개 웹, 교보 공개 웹, Letterboxd 관련 경로, 클리앙 공개 웹, LibraryThing 공개 웹, 왓챠피디아 등이 원시 행을 제공했다. 반면 다수 출처는 36개 작품 전부 또는 대부분에서 0행을 반환했다. 이 0에는 제공자 오류와 접근 제한이 포함되어 있으므로 플랫폼 이용자의 침묵으로 바꿀 수 없다.

생산적 작품–출처 셀은 영화 265/1,144, 소설 112/728로 전체 377/1,872였다. 분모는 계획된 경로 배정 수이며 실제 요청 수나 이용자 노출 수가 아니다. 영화 중심 경로의 소설 비적용과 제공자 오류, 검색결과 부재가 같은 0 행으로 나타날 수 있고, 도서 판매 사이트의 많은 행도 독서 경험이 아니라 상품·메뉴 텍스트일 수 있다. 같은 수의 경로를 배정했다는 사실만으로 두 매체에 같은 관찰 기회를 제공했다고 판단할 수 없다.

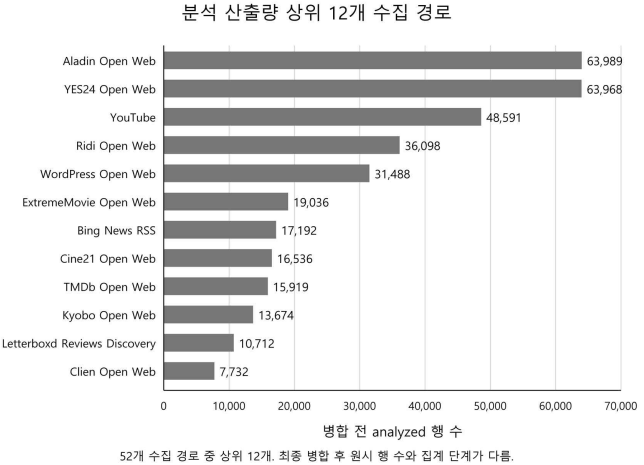
출처별 반환량의 집중은 어떤 경로가 전체 자료의 구성을 크게 좌우하는지 보여 준다. 다만 이 횡단면 집계만으로 한 플랫폼의 폐쇄가 다른 경로로의 이동을 일으켰다고 판정할 수는 없다. 그러한 인과관계에는 시점별 접근 기록과 대체 수집 경로가 필요하다. 여기에서 직접 확인할 수 있는 것은 출처별 산출량의 불균형이며, 그 불균형이 어떤 문서 장르와 반복 구조를 동반하는지는 행 사례와 파서 기록을 함께 읽어야 한다. 이런

11) David Lazer et al. (2009), “Computational Social Science,” *Science*, 323(5915), p. 722. 이 문헌에서는 데이터 접근·공유·프라이버시 및 연구윤리의 과제를 참조한다.

www.kci.go.kr

의미에서 실패의 생태는 관찰된 집중과 그 발생 경로에 관한 검증을 연결하는 분석 과제이다.

(도판 2) 분석 산출량 상위 12개 수집 경로



52개 수집 경로 중 상위 12개. 최종 병합 후 원시 행 수와 집계 단계가 다름.

주: 출처별 병합 전 analyzed 행 수를 기준으로 정렬했다. 서로 독립된 12개 플랫폼을 뜻하지 않는다. 전체 52개 경로와 0행 경로는 보충표 S-2에 제시한다. 출처별 합계는 전출처 병합 이후의 원시 347,992행과 집계 단계가 다르다.

문맥 규칙의 통과 여부는 작품을 지시하는 표현 방식과 문서의 장르에 따라 달라질 수 있다. 정확한 제목이 포함된 상품 정보는 남는 반면, 제목 없이 장면이나 등장인물을 언급한 실제 감상은 탈락할 수 있다. 현재 감사는 이러한 가능성을 구별할 필요를 보여 주지만 각 유형의 빈도나 오류율을 추정하지는 않는다. 따라서 규칙을 엄격하게 만들었다는 사실과 자료가 더 타당해졌다는 판단 사이에는 통과군과 탈락군을 함께 읽는 내용 검증이 놓여야 한다. 정밀도와 재현율 역시 그러한 표본의 선정 방식과 분모를 밝힌 뒤에 평가할 수 있다.

본 자료의 분류는 학습된 언어모델이 아니라 한·영 어휘 사전의 문자열

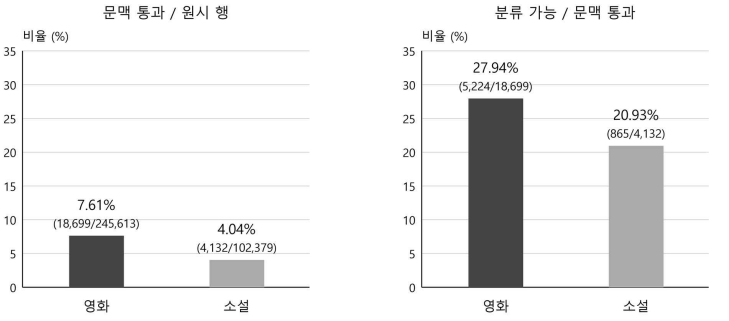
www.kci.go.kr

적중 규칙에 따른다. 문맥 통과 22,831행 가운데 여덟 라벨 가족 중 하나라도 미분류가 아닌 값이 부여된 행은 6,089개였다. 이 수치는 표현과 사전의 적중 관계를 보여 주며, 실제 태도를 올바르게 이해한 비율을 뜻하지 않는다. 가령 사이트가 검색결과에 만족하는지를 묻는 문구에도 '만족'이 있어 긍정 라벨이 붙을 수 있다. 반대로 작품에 관한 해석 가능성이 있는 짧은 말이 모든 라벨에서 미분류로 남을 수도 있다. 분류 가능 여부와 문화수용의 증거 여부는 이렇게 실제 사례에서도 갈라진다. 긍정 라벨이나 미분류 상태만으로는 감상자가 미결정성을 유지하면서 이해를 재구성했는지 알 수 없다. 최적 모호성에 관한 해석은 이처럼 출처와 문맥이 확인된 발화를 읽는 일에서 출발해야 하며, 라벨의 존재나 부재가 그 경험의 성립을 대신 증명하지는 않는다.

실제 비반응에 가장 가까운 주장을 하려면 최소한 출처 접근이 성공했고, 검색 범위와 기간이 정의되었고, 작품 식별 규칙이 검증되었으며, 수용 주체와 문서 장르가 확인되어야 한다. 그 뒤에도 표현되지 않은 감상, 비공개 계정, 삭제된 게시물, 비디지털 독자·관객은 관찰되지 않는다. 그러므로 가능한 최종 문장은 "정의한 플랫폼·기간·질의·필터 범위에서 공개 수용 발화를 발견하지 못했다"이다. "대중이 반응하지 않았다"는 문장은 이 자료로 지지되지 않는다.

www.kci.go.kr

(도판 3) 매체별 두 가지 규칙 통과 비율



패널별 분모가 다르며 누적 손실률이 아님.
점수 조합은 분류 가능 행의 부분집합이 아니므로 그 비율은 제외함.

주: 첫 패널은 문맥 통과 행/원시 행(영화 7.61%, 소설 4.04%), 둘째 패널은 분류 가능 행/문맥 통과 행(영화 27.94%, 소설 20.93%)이다. 분모가 다르므로 누적 손실이나 서로 같은 표본의 성능 비교로 해석하지 않는다. 점수 조합/분류 가능 행의 비율은 포함관계가 성립하지 않아 이 도판에서 제외한다.

V. 수용 가시성의 인프라적 선택성

1. 플랫폼 생태가 만드는 관찰 가능성

수용 가시성의 인프라적 선택성은 자료 바깥의 플랫폼 비판으로만 설명되지 않는다. 접근 가능한 경로를 정하는 플랫폼, 페이지의 일부를 텍스트로 옮기는 수집기, 그 텍스트에 작품명과 라벨을 요구하는 연구자의 규칙이 함께 관찰 범위를 만든다. II장에서 구분한 접근·표현·판정의 세 층은 IV장의 출처별 반환량, 반복 텍스트, 문맥·분류 조건의 통과 수에 각각 연결된다. 다만 수치 하나가 어느 층의 원인을 전부 밝히지는 않는다. 수치와 처리 기록, 실제 행의 문맥을 함께 제시해야 선택이 일어난 지점

www.kci.go.kr

을 구체적으로 설명할 수 있다.

보충표 S-3의 「라울전」 사례는 이 관계가 추상적인 배경에 머물지 않음을 보여 준다. 영상 제목 끝의 '라울'과 댓글 첫머리의 '전설'은 각각의 필드에서는 목표 작품명을 이루지 않는다. 수집된 필드를 결합하고 경계를 지우는 정규화가 개입한 뒤에야 '라울전'이라는 일치가 생긴다. 따라서 이 판정값을 소설에 관한 발화의 존재로 읽기 전에, 어떤 연결이 작품명처럼 보이는 문자열을 만들었는지 되짚어야 한다. 관계적 인프라의 분석은 바로 이 지점에서, 이미 주어진 대상을 세는 일과 처리 과정에서 대상을 구성하는 일을 구별하게 한다.

'인프라적'이라는 수식어는 원인을 단일 플랫폼의 편향으로 좁히지 않는다. 수집 코드, 네트워크, 검색 제공자, 데이터 스키마, 파일 형식, 분류 모델, 연구자의 임계값이 연결된 전체 배열이 관찰면을 만든다. '선택성'은 의도적 검열만을 뜻하지 않는다. 기술적 표준, 비용, 속도 제한, 언어 자원, 제목 규칙, 분석 편의도 선택 효과를 낸다. Crawford가 AI를 광물·노동·에너지·분류 체계에 의존하는 물질적·정치적 시스템으로 파악한 관점은 문화 데이터 역시 비물질적 텍스트의 구름이 아니라 계산 자원과 제도적 접근의 조건 위에서 제한적으로 복무하는 자리에 놓여 있음을 상기시킨다.[12]

계산 과정에서 사라질 수 있는 것은 문서의 장르와 발화의 상황이다. 보충표 S-3에서 검색결과에 만족하는지를 묻는 사이트 문구는 '만족'이라는 어휘를 포함하여 긍정 라벨을 얻었다. 그러나 사이트가 답변을 요청하는 행위와 독자가 작품에 대한 만족을 표현하는 행위는 같은 수용 발화로 묶이지 않는다. 이 차이를 읽는 순간 라벨의 계산 결과는 유지되더라도, 그 값을 독자의 태도에 관한 증거로 사용할 수 있는지는 다시 판단해야 한다. 인문학적 독해는 이처럼 계산이 끝난 뒤 의미를 덧붙이는 데 그치지 않고, 무엇을 같은 자료로 세고 무엇을 다른 상태로 남길지 결정하는

12) Kate Crawford (2021), *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale University Press, p. 8.

www.kci.go.kr

과정에 참여한다.

이 점에서 본 연구의 탈경계적 기여는 도구와 해석의 일방향적인 결합을 넘어서는 데 있다. 데이터 감사는 인문학적 해석이 사용하는 증거의 범위를 점검하고, 문맥에 대한 해석은 다시 수집과 분류의 단위를 수정하도록 요구한다. 행·문서·발화·작성 주체·작품·출처·관찰 시점을 구분하는 일은 그 왕복을 가능하게 한다. 이러한 관계를 보존할 때 연구 장치의 실패를 사람의 침묵으로 전가하지 않는 비판적 실천이 가능해진다.

자료 구조에서는 반응 수와 관찰 상태를 별도 필드로 보존해야 한다. 접근 실패라면 반응 수를 미관찰 값으로 두고 접근 상태와 오류 사유를 기록한다. 관찰 범위를 확인한 뒤 발화를 발견하지 못한 경우에만 그 범위의 관찰 건수를 0으로 기록한다. 작품에 해당하지 않는 경로는 적용 불가로, 문맥이나 분류를 판정하지 못한 행은 각 탈락 사유로 남긴다. 이렇게 하면 연구자는 0이라는 결과를 보기 전에 어떤 관찰이 가능했는지 되짚을 수 있다. 자료 감사의 실천적 효과는 빈칸을 없애는 데보다 빈칸에 잘못된 의미를 부여하지 않는 데 있다. 이와 같은 기록과 이유 제시의 중요성은 생성형 인공지능 플랫폼의 거버넌스 문서를 대상으로 절차적 신뢰와 감사 가능성을 논의한 유성희와 서효중의 연구를 자료 처리의 문제에 연결하여 살펴볼 수 있다.[13]

2. AI 판독의 인식론적·윤리적 경계

자료의 계보는 모델 출력의 해석 범위를 정한다. 작품과 무관하거나 반복된 입력을 사용했다면, 출력의 일관성이 높아도 수용의 분포를 설명하는 데에는 제약이 남는다. 그러므로 모델이나 규칙에 관한 정보와 함께 접근 기록, 텍스트 결합 방식, 중복 키, 문맥 조건, 제외 사유를 공개해야 한다.

공개된 자료의 이용에도 별도의 정당화가 필요하다. 공개 게시물이라도

13) 유성희, 서효중 (2026), 「생성형 AI 플랫폼 거버넌스의 절차적 신뢰: 사적 규범, 정책 드리프트, 시민사회형 알고리즘 감사」, 『시민사회와 NGO』, 24(1), 24쪽, 51쪽. 이 연구의 대상은 생성형 AI 제공자의 거버넌스 문서이며, 여기서는 기록·이유제시·이의제기 및 감사 가능성의 논의를 참조한다.

www.kci.go.kr

작성자는 학술 데이터셋 편입이나 자동 분류를 예상하지 않았을 수 있다. 본고의 사례표는 판정에 필요한 최소 구간만 제시하며 계정명·연락처·플랫폼의 이용자 식별자를 제외한다. 희소하거나 민감한 개인 발화의 공개는 피하고, 원문 전체와 URL은 제한된 감사 기록에 보존한다. 다만 짧은 인용이나 해시만으로 재식별 위험이 사라지는 것은 아니므로 저자가 사례별로 검토해야 한다. 자료에 접근할 수 있다는 사실과 그 이용을 정당화하는 판단은 구별된다.

분류 실패는 텍스트와 연구자의 범주 사이에서 검토해야 한다. 반어·양가성·인용처럼 복합적인 표현을 남기기 어려운 범주라면, 분류 체계가 무엇을 놓쳤는지부터 질문할 필요가 있다. 이를 곧바로 수용자의 무관심이나 인지적 회피로 설명하지 않는다. 최적 모호성에 관한 후속 해석 역시 유효한 수용 자료를 확보한 다음의 과제이다.

마지막으로 지켜야 할 원칙은 자동화 사용을 연구 기록에 포함하는 것이다. 생성형 AI가 코드 작성, 분류 규칙 초안, 문장 윤문에 사용되었다면 도구명, 사용일, 역할, 사람이 수행한 검증을 공개해야 한다. 단, 원자료에 개인정보나 서비스 약관상 제한된 내용이 있다면 이를 외부 모델에 전송하지는 않아야 한다. 원고 작성에 사용한 도구의 역할은 연구윤리 및 활용 고지에 남기고, 자료와 인용 근거의 확인 범위는 각각의 검사 기록에서 구별한다. 사용 사실의 공개와 산출물의 검증은 서로를 대신할 수 없으므로, 수행한 점검과 남아 있는 확인 과제를 함께 제시해야 한다.

(표 5) 주장 수준과 필요한 증거

주장	현재 자료의 지지 여부	추가로 필요한 증거
원시 수집물에 대규모 정확 중복이 있다	지지	중복 원인별 추적은 별도 필요
허용적 수집물에는 작품과 무관하거나 작품에 대한 수용 발화가 아닌 문서가 포함될 수 있으므로 내용	확인한 사례는 이러한 포함 가능성을 지지한다. 다만 전체 수집물의 오류 비율은 현재 점검만으로	판정 상태별로 실제 행을 표집하여 내용을 검토하고, 각 표본의 분모와 선정 절차를 공개해야 한다.

www.kci.go.kr

검증이 필요하다.	확정하지 못하였다.	
접근·필터가 관찰 가능한 자료를 선택한다	지지	시간대별·수집기별 반복 감사로 강화 가능
영화와 소설의 대중 반응량이 다르다	지지하지 않음	매체별 표집틀과 노출분모 필요
작품 형식이 온라인 반응을 유발한다	지지하지 않음	타당한 반응 표본과 인과 설계 필요
작품이 신경적으로 최적 모호성을 만든다	지지하지 않음	통제 실험과 행동·생리·신경 측정 필요

Ⅵ. 결론

1. 연구의 기여와 실천적 함의

본 연구는 36개 작품과 52개 출처의 기록을 통해 온라인 문화수용 자료가 수집되고 판정되는 조건을 검토했다. 보고된 원시 347,992행에 대한 정확 반복 진단과 문맥·분류 조건별 집계는 서로 다른 검사를 나타낸다. 자료의 크기는 그대로 수용의 규모가 되지 않으며, 통과 수가 줄었다는 사실만으로 남은 자료의 의미적 타당성이 완성되지도 않는다. 이 결과는 감정 분포나 작품 간 수용 차이를 제시하기에 앞서, 행이 어떤 문서와 발화에서 비롯되었는지 확인해야 한다는 요구로 이어진다.

첫째, 본 연구는 관찰 실패와 문화적 부재 사이에서 판단이 성립하는 조건을 구체화했다. 접근·문맥·분류의 실패를 나누어 기록하면 연구 장치가 알 수 없었던 것을 수용자가 하지 않은 일로 단정하는 오류를 피할 수 있다. 이는 결측을 세분하는 기술적 조치인 동시에, 기록되지 않은 경험에 대한 해석의 책임을 정하는 문제이다.

둘째, 플랫폼과 연구자의 처리 규칙을 같은 감사 범위 안에 두었다. 플랫폼은 자료의 공개 범위를 정하고, 연구자는 그중 무엇을 행과 범주로 남길지 결정한다. 수용 가시성의 인프라적 선택성은 이 공동의 구성을 설명

www.kci.go.kr

하는 개념이다. 인문학적 문맥 판단을 분석이 끝난 뒤의 해설에 가두지 않고, 자료 단위를 구성하는 단계로 되돌린다는 데 본 연구의 의의가 있다.

셋째, 재현성과 타당성이 서로 다른 검증 과제임을 자료의 수준에서 분리했다. 같은 파일과 코드를 다시 읽는 절차는 중요하지만, 그 절차가 보존하는 행이 어떤 경험을 지시하는지는 별도로 확인해야 한다. 본 연구가 제안하는 감사는 더 강한 모델을 선택하는 문제에 앞서, 모델에 어떤 증거를 제공하고 그 결과로 무엇을 주장할지 결정하는 기초 작업이다.

실천적으로, 본 연구는 네 가지 보고 원칙을 제안한다. 첫째, 원시 행과 고유 문서, 이용자, 작품–출처 시도를 구분한다. 둘째, 각 출처의 요청 수, 성공 수, 오류 유형, 관찰 기간을 공개한다. 셋째, 관련성·분류 규칙의 통과 수뿐 아니라 수동 검증의 정밀도와 재현율을 보고한다. 넷째, '발견되지 않음'을 반드시 관찰 창과 함께 문장화한다. 이 원칙은 더 작은 숫자를 만드는 규율이 아니라, 숫자가 말할 수 있는 범위를 정확히 만드는 규율이다. 이는 후속 연구에서도 지켜야 할 보고 원칙이며, 본 연구가 모든 요청 로그와 의미적 정답에 근거한 성능평가를 확보했다는 뜻은 아니다.

2. 한계와 후속 과제

첫째, 본 연구는 전체 수집물에서 정규화된 텍스트의 정확 일치 여부를 중복 진단의 중심으로 사용했다. 근접 중복, 템플릿 일부 반복, 번역·축약·재게시, 동일 사용자의 반복 발화는 완전히 식별하지 못했다. 반대로 서로 다른 작성자가 우연히 같은 짧은 문장을 작성한 경우를 초과 중복으로 셀 가능성이 있다. 후속 연구는 문서 길이와 출처를 층화한 뒤 문자·의미 기반 근접 중복을 함께 비교해야만 한다.

둘째, 정확 제목·문맥 조건은 판정 범위를 좁히는 운영 규칙이다. 그 결과만으로 정밀도가 높아졌다고 확정할 수는 없다. 별칭과 번역 제목, 인물명이나 장면을 통한 지시를 놓칠 수 있으므로, 통과군과 탈락군 모두에서 내용 검증이 필요하다. 층화 표본은 작품·출처·판정 상태별로 나누어 추출하고, 독립적인 사람 코더의 판단과 불일치를 함께 기록해야 한다.

www.kci.go.kr

통과군만 확인해서는 전체 관련 문서 중 얼마나 찾았는지를 뜻하는 재현율을 추정할 수 없다.

셋째, 52개 출처는 가능한 온라인 공간의 전체 목록이 아니다. 접근 가능한 공개 웹에 편중되며, 비공개 커뮤니티, 메신저, 오프라인 감상, 삭제된 콘텐츠를 포함하지 않는다. 출처별 이용자 구성과 추천 알고리즘의 노출분모도 알 수 없다. 그러므로 본 연구의 수치는 특정 수집 시점과 구현에 묶인 인프라 감사 결과일 따름이다.

넷째, 출처 제한 점수 조합 3,428개는 수용 발화의 표본틀로 사용할 수 없다. 서로 다른 텍스트가 점수의 동일성 때문에 합쳐질 수 있기 때문이다. 후속 검증은 원시 행의 문서 식별자·텍스트·출처·작품 연결을 보존한 자료로 돌아가야 한다. 그 자료에서 작품·출처·판정 상태별 표본을 추출하고 독립적인 코더가 문서 장르, 작품 관련성, 작성 주체와 발화의 독립성을 판단할 필요가 있다. 합의된 발화를 확보해도 플랫폼 내부의 표집 확률과 노출분모가 없으면 모집단 대표성이 자동으로 성립하지 않는다.

다섯째, 본 연구가 점검한 것은 수용 해석의 자료 조건이며, 감상자의 인지적 변화를 측정한 것은 아니다. 최적 모호성을 이해의 재구성이 가능한 관계적 조건으로 탐구하려면, 검증된 발화 자료에 더하여 그 해석이 요구하는 경험적 증거를 설계해야 한다. 본 연구의 역할은 그 후속 질문을 대신 해결하는 데 있지 않고, 자료 식별의 불확실성을 미학적 효과로 오인하지 않도록 출발점을 분명히 하는 데 있다.

수집 실패를 무반응과 구별하는 일은 문화수용 연구를 포기하는 일이 아니다. 오히려 관찰한 것과 아직 알 수 없는 것을 나눔으로써, 자료에 남지 않은 사람의 경험을 성급하게 재단하지 않으려는 작업이다. 플랫폼과 수집기, 연구자의 범주가 함께 만든 기록을 그 자체로 수용의 전부라 부를 수는 없다. 그 기록이 생겨난 관계를 되짚을 때 비로소 계산은 해석의 근거가 되고, 해석은 다시 계산이 보존해야 할 차이를 요구할 수 있다.

www.kci.go.kr

부록 A. 감사 변수와 판정 규칙

(부록 표 A-1) 핵심 변수 사전

변수	단위	정의	해석 주의
`work_id`	작품	36개 목표 작품의 비식별 내부 키	작품명과 일대일 대응표 별도 보존
`source`	출처	수집 경로 또는 제공자 이름	동일 플랫폼의 여러 경로를 구분
`source_attempt`	작품-출처	36개 목표 작품과 52개 수집 경로를 조합한 계획상의 셀을 뜻한다. 실제 인터넷 요청 횟수와는 구별한다.	반응 존재 여부가 아님
`productive_instance`	작품-출처	분석 행이 1건 이상 반환된 조합	관련성·타당성을 보장하지 않음
`raw_row`	행	허용적 기존 규칙으로 저장된 항목	문서·사람·반응 수와 다름
`exact_text_duplicate_excess`	행	같은 텍스트의 첫 출현을 초과한 반복	중복 원인의 동일성을 뜻하지 않음
`distinct_text`	텍스트	전역 정확 정규화 텍스트 기준 고유값	고유 이용자 수가 아님
`cross_target_recurrent_text`	텍스트	둘 이상의 목표 작품 번들에 재등장한 정확 텍스트	여러 작품에서 같은 텍스트가 나타났다는 사실만으로 각 행이 해당 작품의 반응이 아니라고 확정할 수는 없다. 작품별 원문 문맥과 발화의

www.kci.go.kr

			독립성을 추가로 확인해야 한다.
`exact_context_pass`	행	정규화 목표 제목과 최소 문맥 조건 통과	수용 발화 확정이 아님
`classifiable`	행	설정 라벨에서 하나 이상이 미분류가 아닌 항목	정답 라벨 또는 대표표본이 아님
`provisional_reception_residue`	작품별로 구별한 점수와 출처의 고유 조합을 센다.	별도의 출처·제목·길이 조건을 통과한 행에서 일곱 점수 열과 출처 열만 남긴 뒤, 각 작품 안에서 같은 조합을 하나로 묶어 센 값이다.	이 값은 분류 가능 행의 부분집합이 아니다. 서로 다른 텍스트나 문서 또는 수용 발화의 수를 나타내지도 않는다. 기존 변수명은 과거 결과를 추적하기 위해 보존하되, 본문에서는 출처 제한 점수 조합이라는 명칭을 사용한다.

(부록 표 A-2) 집계 대조식

대조 항목	식	결과
계획된 작품-출처 셀	36 작품 × 52 출처	1,872
매체별 계획 셀 합	1,144 + 728	1,872
매체별 원시 행 합	245,613 + 102,379	347,992
중복 진단 합	245,803 + 102,189	347,992
작품 간 재등장 텍스트	19,914개 텍스트가 구성한 행	265,716
생산적 인스턴스 합	265 + 112	377

www.kci.go.kr

(부록 표 A-3) 무결성 검사와 타당성 검사의 분리

검사 층위	대상	결과	결론의 범위
원본 대응	영화 프레임 284,400개	내부 영화 코퍼스의 처리 기록에 해당한다.	이 기록을 온라인 수용 자료의 내용 타당성을 입증하는 근거로 사용하지 않는다.
원본 대응	소설 문장 22,209개	내부 소설 코퍼스의 처리 기록에 해당한다.	이 기록을 온라인 수용 자료의 내용 타당성을 입증하는 근거로 사용하지 않는다.
번들 대응	온라인 번들 36개	36개 번들의 쉼표로 구분된 자료 파일 108개를 기존 검사값과 대조하여 모두 일치함을 확인하였다.	이 결과는 지정된 온라인 파일의 보존 상태를 확인한 범위에 한정된다.
정규화 중복	온라인 행 347,992개	고유 102,189개, 초과 중복 245,803개	행 수를 독립된 관찰치의 수로 사용하려면 발화와 문서의 독립성을 추가로 검토해야 한다.
작품 간 재등장	정확 텍스트 19,914개	265,716행을 구성	반복만으로 작품 관련성을 확정할 수 없으며, 문맥과 발화의 독립성을 추가로 검토해야 한다.
내용 관련성	허용적 기존 통과	확인한 사례에는	확인한 사례에는

www.kci.go.kr

	행	작품과 무관한 문서, 여러 문서에서 반복되는 공통 문구와 상품 안내가 포함되어 있다.	작품과 무관하거나 작품에 대한 수용 발화가 아닌 내용이 포함되어 있다. 전체 수집물의 오류율은 현재 점검만으로 확정하지 못하였다.

부록 B. 연구윤리 점검표

(부록 표 B-1) 윤리와 위험 완화

항목	본 원고의 조치	공개·보존 원칙
공개 게시물의 맥락	공개 접근과 연구 동의를 구분	플랫폼 약관과 기관 지침 재검토
개인정보	개인 계정명과 민감한 정보를 제외하고, 판정의 이유를 설명하는 데 필요한 최소한의 구간만 제시한다.	집계치와 사례를 일관되게 식별하는 번호 및 판정 근거를 공개하되 원문 전체는 재배포하지 않는다.
재식별 위험	희소한 조합에 속하거나 민감한 내용을 담은 개인 발화는 공개 예시에서 제외한다.	원래의 인터넷 주소와 전체 원문은 접근이 제한된 감사 기록에 보존하고, 공개 예시는 비식별 사례 번호를 통해 해당 기록과 연결한다.
삭제권과 맥락 붕괴	삭제된 원문을 복원·재배포하지 않음	재수집보다 감사 로그를 우선 보존
자동 분류	라벨을 인간의 진실한 태도로 간주하지 않음	오류표본과 판정 불일치를 함께 기록
플랫폼 부담	요청 제한과 로봇 정책 준수 여부 확인	요청률·재시도 정책 기록
취약 발화	개인 식별 가능한 민감 발화 분석 제외	필요 시 기관생명윤리 검토 선행

www.kci.go.kr

해석 위해	작품·매체·집단의 인기나 감정으로 일반화하지 않음	주장-증거 대응표 유지

부록 C. 재현성 점검표

(부록 표 C-1) 재현을 위한 최소 공개 항목

영역	공개할 항목	현재 상태/추가 조치
환경	운영체제, 언어·패키지 버전, 실행일	Python 3.14.3 등 일부 작품에서 확인된 세부 실행 기록을 전체 작품의 실행 환경으로 일반화하지 않는다.
질의	작품명, 별칭, 정규화 규칙	목표 작품 36편의 이름과 수집 경로 52개의 목록, 감사에 사용한 정규화 규칙을 보충표와 방법 절에 제시한다. 실제 검색에서 사용된 모든 질의 변형, 결과 순위와 페이지별 응답 기록은 복원하지 못하였으므로 확인된 목록과 구별하여 표시한다.
접근	출처별 엔드포인트 유형, 기간, 요청·오류 수	보존된 수집 결과와 출처별 집계는 대조하였다. 전체 요청·응답 및 페이지별 접근 로그가 모두 복원된 것은 아니다.
변환	HTML/JSON 필드 매핑, 텍스트 결합 규칙	필드 매핑과 결합 규칙을 버전별 보존
중복	전역/작품별 범위, 해시 함수,	정확 텍스트 기준을

www.kci.go.kr

	정규화 전후	기본값으로 명시
문맥	제목 일치, 길이, URL 사용 규칙	보존된 사전 규칙과 표본 재계산 결과를 제시한다. 모든 버전의 단위시험 기록까지 확보하였다는 뜻은 아니다.
분류	라벨 정의, 미분류 처리, 실제 사전 규칙을 제시한다. 이 분류에는 학습모델의 프롬프트를 적용하지 않았다.	내용 검증 전에는 이 감사 집계를 수용자의 태도나 감정 분포로 해석하지 않는다.
감사	단계별 건수, 제외 사유, 수동 검증표	본문의 합계와 자동 대조
버전	원장 체크섬, 코드 커밋, 산출물 생성일	기존 자료 수집 기록표의 원시 347,992행과 대조
생성형 AI	도구명·사용일·역할·인간 검증	코드 점검·오류 추적·형식 교정의 보조 활용 공개

(부록 표 C-2) 분석 중단 기준

점검	중단 조건	본 연구의 판단
중복 독립성	행 수를 독립 반응 수로 볼 근거 없음	반응 빈도 분석 중단
관련성	확인한 사례에 작품과 무관하거나 수용 발화가 아닌 내용이 포함되어 있으며, 전체 행의 관련성을 확인하는 내용 검증은 완료되지 않았다.	감정 분포 분석 중단
작성 주체	이용자 발화와 상업·정보 문서 미분리	관객·독자 비교 중단
대표성	플랫폼 노출분모·표집확률 부재	모집단 추론 중단

www.kci.go.kr

인과성	형식 지표와 반응 사이 식별 전략 부재	형식-수용 상관 인과 중단
신경 타당성	행동·생리·신경 측정 부재	최적 모호성의 신경적 주장 중단

보충표 S-2의 출처는 52개의 독립 플랫폼이 아니라 수집 경로 이름이다. 같은 플랫폼의 직접 경로와 발견 경로가 함께 있을 수 있다. 각 경로의 계획 작품 수는 36이며, 산출량은 전출처 병합 전의 analyzed 합계이다. 합계 354,775행은 병합 후 원시 347,992행과 단계가 다르다. 이 목록의 0행을 플랫폼의 무반응으로 해석하지 않는다. 전체 경로 목록과 산출량은 부록 D의 보충표 S-2에, 추가 기록 열은 별도 데이터 보충자료의 table_s2_sources.csv에 제시한다.

보존된 출처별 요약 파일의 작성 시각은 2026년 8월 9일부터 12일까지에 걸쳐 있다. 이 시각은 개별 게시물의 작성일이나 최초 웹 응답 시각과 구별한다. 8월 12일의 네 작품 실행 기록에는 Python 3.14.3과 pandas 3.0.2, beautifulsoup4 4.14.3, lxml 6.1.0의 버전 및 실행 파일의 검사값이 남아 있으며, 해당 검사값은 보존된 구형 실행 파일과 일치하였다. 다만 이 부분 기록을 36편 전체의 단일 실행 환경으로 확대하지 않는다.

분석의 재현 범위는 보존된 자료와 그 자료에 적용된 규칙을 기준으로 정한다. 8월 12일 이후 오류 수정·보완과 질적 개선을 반영한 연구자의 통합 연구 지침 5.2판(md)과 Python 검증 코드(py)는 다언어 검색, 번역 자료의 격리 및 언어별 민감도 분석을 지원한다. 다만 원자료를 보존하기 위해 기존 연구자료 전체에 이 갱신된 절차를 적용하여 실행·검증하지는 않았다. 따라서 본 연구의 집계 대상은 연구자의 로컬 환경에 보존된 36개 작품의 온라인 수집물이다. 12개 언어의 표본 확보, 번역의 의미 검증, 독립적인 사람 코더 간의 일치도 측정은 수행 범위에 포함하지 않았다. 공개 재현자료의 집계·도판 재계산 절차는 원시 행의 의미 판정

www.kci.go.kr

과 구별되며, 후자에는 비공개 원자료의 열람과 별도의 검증이 필요하다.

부록 D. 작품별 집계와 수집 경로 목록

보충표 S-1과 S-2는 본문에서 사용한 작품 및 경로 집계의 전체 목록이다. S-1의 작품명에서는 파일 식별용 밑줄만 띄어쓰기로 표시하였다. 별도 데이터 보충자료에는 파일명, 검사값과 추가 진단 열을 함께 제공한다. S-3의 12개 사례표는 S-2 다음에 수록하였다. 세 표는 S-1, S-2, S-3의 순서로 배치하였다.

(보충표 S-1) 36개 작품의 집계

수집 목표 작품명	매체	원시 행	문맥 통과 행	분류 가능 행	별도 점수 조합
7월의 아이들	소설	11,086	3	3	0
9월의 다알리아	소설	9,997	10	3	5
가면고	소설	0	0	0	0
강변호텔	영화	10,211	1,149	250	167
강원도의 힘	영화	11,693	1,016	241	165
광장	소설	769	768	140	4
구운몽	소설	10	0	0	0
그 후	영화	13,801	2,103	563	218
그레이 구락부 전말기	소설	7,462	22	7	1
극장전	영화	10,395	602	231	100
금오신화	소설	9,927	910	255	120
누구의 딸도 아닌 해원	영화	9,652	267	83	68

www.kci.go.kr

수집 목표 작품명	매체	원시 행	문맥 통과 행	분류 가능 행	별도 점수 조합
다른 나라에서	영화	12,174	502	124	35
달과 소년병	소설	9,830	23	11	13
당신 자신과 당신의 것	영화	12,398	523	141	77
대숲 소리	소설	9,743	190	26	20
돼지가 우물에 빠진 날	영화	9,066	145	62	37
라울전	소설	8,989	23	5	15

(보충표 S-1 계속) 36개 작품의 집계

수집 목표 작품명	매체	원시 행	문맥 통과 행	분류 가능 행	별도 점수 조합
바다의 편지	소설	11,987	204	37	0
밤의 해변에서 혼자	영화	12,776	1,302	469	289
북촌방향	영화	10,149	1,113	321	132
생활의 발견	영화	12,030	1,816	386	172
소설가의 영화	영화	12,402	729	226	150
수유천	영화	9,669	929	195	51
여자는 남자의 미래다	영화	11,372	407	177	34
여행자의 필요	영화	12,354	896	193	91
오! 수정	영화	11,010	298	80	109
옥희의 영화	영화	11,465	560	162	71

www.kci.go.kr

수집 목표 작품명	매체	원시 행	문맥 통과 행	분류 가능 행	별도 점수 조합
우상의 집	소설	11,297	14	4	0
웃음소리	소설	10,835	1,755	339	176
인트로덕션	영화	9,587	821	201	77
잘 알지도 못하면서	영화	10,740	1,573	495	646
지금은 맞고 그때는 틀리다	영화	10,543	483	185	151
클레어의 카메라	영화	11,131	668	186	96
풀잎들	영화	10,995	797	253	124
회색인	소설	447	210	35	14

주: 전체 합계는 원시 347,992행, 문맥 통과 22,831행, 분류 가능 6,089행, 별도 점수 조합 3,428개이다. 마지막 열은 앞선 분류 가능 행의 부분집합이나 검증된 발화 수가 아니다. 작품명은 수집 목표명이며 판본 정보나 대표성을 뜻하지 않는다.

(보충표 S-2) 52개 수집 경로의 산출량

수집 경로	계획 작품 수	병합 전 analyzed 행
Aladin Open Web	36	63,989
YES24 Open Web	36	63,968
YouTube	36	48,591
Ridi Open Web	36	36,098
WordPress Open Web	36	31,488

www.kci.go.kr

수집 경로	계획 작품 수	병합 전 analyzed 행
ExtremeMovie Open Web	36	19,036
Bing News RSS	36	17,192
Cine21 Open Web	36	16,536
TMDb Open Web	36	15,919
Kyobo Open Web	36	13,674
Letterboxd Reviews Discovery	36	10,712
Clien Open Web	36	7,732
Letterboxd	36	7,140
LibraryThing Open Web	36	1,908
WatchaPedia	36	792
Blogspot Open Web	36	0
Cineaste Open Web	36	0
DCInside Open Web	36	0

(보충표 S-2 계속) 52개 수집 경로의 산출량

수집 경로	계획 작품 수	병합 전 analyzed 행
Broad Open Web Seeds	36	0
Brunch/Tistory Open Web	36	0
IMDb	36	0
Google Search	36	0

www.kci.go.kr

수집 경로	계획 작품 수	병합 전 analyzed 행
Google News RSS	36	0
Goodreads	36	0
FMKorea Open Web	36	0
DuckDuckGo HTML Discovery	36	0
Daum/Tistory RSS/Seed	36	0
Daum	36	0
Brunch Open Web	36	0
DVDPrime Open Web	36	0
MUBI Open Web	36	0
Medium Open Web	36	0
Instiz Open Web	36	0
KMDb Open Web	36	0
MLBPark Open Web	36	0
MSN	36	0

(보충표 S-2 계속) 52개 수집 경로의 산출량

수집 경로	계획 작품 수	병합 전 analyzed 행
Naver Blog RSS/Seed	36	0
Naver	36	0
Nate Pann Open Web	36	0

www.kci.go.kr

수집 경로	계획 작품 수	병합 전 analyzed 행
NamuWiki Open Web	36	0
Movist Open Web	36	0
Metacritic Open Web	36	0
Reddit	36	0
Ppomppu Open Web	36	0
Substack Open Web	36	0
StoryGraph Open Web	36	0
Ruliweb Open Web	36	0
RottenTomatoes Open Web	36	0
Trakt Open Web	36	0
TodayHumor Open Web	36	0
TheQoo Open Web	36	0
X/Twitter	36	0

주: 계획 작품 수는 각 경로 36개, 전체 1,872개 작품-경로 셀이다. 병합 전 산출량 합계는 354,775행으로 병합 후 원시 347,992행과 구별된다. 0행은 해당 수집 경로의 산출 기록이며 독자나 관객의 무반응을 입증하지 않는다.

(보충표 S-3) 판정 규칙과 원행을 대조한 사례 12개

ID·구분	작품·출처·행	최소 발췌	판정 근거
E01 A 무관 문서	강변호텔 / Aladin Open Web / L7	100자평 작성하면 추가 마일리지 2% 적립	목표 제목이 어느 필드에도 없고, 본문은 북펀드 상거래 안내다.
E02 A	강변호텔 /	100자평 작성하면	다른 북펀드 항목도 동일한

www.kci.go.kr

ID·구분	작품·출처·행	최소 발췌	판정 근거
무관 문서	Aladin Open Web / L8	추가 마일리지 4% 적립	비수용 문구를 포함한다.
E03 D 잔여 오염 반례	강변호텔 / Clien Open Web / L3480	70% 이상의 투과율 필름	부모 제목에 강변호텔이 있지만 본문은 차량 틴팅이다. 전부 미분류인 실제 잔여 대표행.
E04 B 제목 포함 비수용	강변호텔 / Kyobo Open Web / L4656	검색 결과에 대한 의견을 남겨주세요	본문에 목표 제목이 있으나 이용자의 의견이 아니라 사이트의 입력 요청이다.
E05 B 제목 포함 비수용	강변호텔 / Kyobo Open Web / L4686	상품선택 품절 무료배송 [BLU-RAY] 강변호텔	상품 판매 UI다. 제목과 길이 조건을 통과해도 발화 주체는 확인되지 않는다.
E06 B 제목 포함 비수용	강변호텔 / Kyobo Open Web / L4736	검색결과에 만족하셨나요?	sentiment=positive, pleasure=pleasure. '만족' 문자열 때문에 사이트 질문 자체가 긍정 반응으로 라벨링된다.
E07 C 문맥 통과 미분류	광장 / Bing News RSS / L6	1년도 안 돼 짐싸는 청년들	부모 제목은 소설 광장의 개정판 기사인데 본문은 고용 관련 뉴스다. 8개 라벨 모두 unclassified.
E08 D 잠정 잔여 대표행	광장 / YouTube / L766	우리나라는 광장인것 같다	작품 관련 해석 가능성이 있는 짧은 발화지만 8개 라벨 모두 미분류. 실제 3,428 점수 조합의 첫 대표행으로 남는다.
E09 D 잠정 잔여 대표행	광장 / YouTube / L767	남녁x 남녘o 자막 수정 부탁드림다	작품 자체의 수용보다 교육영상 자막 정정에 가깝다. 전부 미분류인데 실제 잔여 대표행이다.
E10 C 문맥 통과	라울전 / Kyobo Open	검색 결과에 대한 의견을 남겨주세요	라벨 사전 미적중. B와 C가 서로 배타적인 문서 장르가

www.kci.go.kr

ID·구분	작품·출처·행	최소 발췌	판정 근거
미분류	Web / L2939		아님도 보여 준다.
E11 D 경계 결합 반례	라울전 / YouTube / L7406	전설카드 이리모으기도힘듦	축구영상 제목 끝 '라울'과 댓글 시작 '전설'을 연결하면 '라울전'이 생긴다. 개별 title/text/url에는 라울전이 없지만 결합 문맥은 통과한다.
E12 D 잠정 잔여 대표행	생활의 발견 / ExtremeMovie Open Web / L3146	기존에 잘 알려진 책이나 곡명 혹은 경구	여러 영화의 제목 짓기에 관한 논평. question 라벨, 실제 잔여 대표행이다. 단일 작품에 대한 독립 수용으로 단정할 수 없다.

www.kci.go.kr

연구윤리 및 생성형 AI 활용 고지

생성형 AI는 원고의 교정·교열, 비학술적인 비평적 문체의 축소·완화, 표현 방식의 정형화와 표준화 및 균질화, 문장 수준의 오류 탐지에 제한적으로 사용되었다. 특히 영어를 모국어로 삼지 않는 비영어권 국가에 거주 중인 연구자가 작성한 어색한 영어 문장을 고급스럽고 세련된 학술 영어 문장으로 재번역하는 데 공헌하였다. 연구 질문과 가설 설정, 연구 대상 선정, 원자료의 확보, 분석 코드와 변수 설계, 통계분석과 시각화, 작품 분석 및 해석, 결론 도출과 최종 판단은 연구자가 책임지고 수행하였다. 코드 작성 및 파이프라인 구축은 연구자의 로컬 컴퓨터에 설치된 Python 3.14 및 PyCharm 환경에 의해 수행되었으며, AI는 결과의 정확성과 신뢰도를 점검하고 문제 요소를 보강하는 데 국소 활용되었다. 논문 설계와 구조 정립, 문제의식의 주제화, 방법론의 설정과 적용에 이르는 전 과정을 숙고하여 스스로 논문화하였다.

www.kci.go.kr

참고문헌

김홍중 (2022), 「플랫폼의 사회이론: 플랫폼 자본주의와 알고리즘 통치성을 중심으로」, 『사회와 이론』, 41, 7~48쪽.

유성희·서효중 (2026), 「생성형 AI 플랫폼 거버넌스의 절차적 신뢰: 사적 규범, 정책 드리프트, 시민사회형 알고리즘 감사」, 『시민사회와 NGO』, 24(1), 23~61쪽.

이재연 (2018), 「디지털 시대의 인문학에서 디지털 인문학 시대로 - 한국문학에서 본 디지털 인문학 연구」, 『역사학보』, 240, 157~190쪽.

한상우 (2020), 「인문학 기반 데이터 리터러시 모형 설계에 관한 연구」, 『정보관리학회지』, 37(1), 179~195쪽.

Bowker, Geoffrey C., Susan Leigh Star (1999), *Sorting Things Out: Classification and Its Consequences*, MIT Press.

boyd, danah, Kate Crawford (2012), "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society*, 15(5), pp. 662~679.

Crawford, Kate (2021), *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale University Press.

Frascaroli, Jacopo, Helmut Leder, Elvira Brattico, Sander Van de Cruys (2024), "Aesthetics and Predictive Processing: Grounds and Prospects of a Fruitful Encounter," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1895), Article 20220410.

Gillespie, Tarleton (2010), "The Politics of 'Platforms'," *New Media & Society*, 12(3), pp. 347~364.

Lazer, David et al. (2009), "Computational Social Science," *Science*, 323(5915), pp. 721~723.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, Emre Kıcıman (2019), "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *Frontiers in Big Data*, 2, Article 13.

Plantin, Jean-Christophe, Carl Lagoze, Paul N. Edwards, Christian Sandvig (2018), "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook," *New Media & Society*, 20(1), pp. 293~310.

Ruths, Derek, Jürgen Pfeffer (2014), "Social Media for Large Studies of Behavior," *Science*, 346(6213), pp. 1063~1064.

www.kci.go.kr

Star, Susan Leigh, Karen Ruhleder (1996), "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces," *Information Systems Research*, 7(1), pp. 111~134.

Tufekci, Zeynep (2014), "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls," *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pp. 505~514.

van Dijck, José, Thomas Poell, Martijn de Waal (2018), *The Platform Society: Public Values in a Connective World*, Oxford University Press.

www.kci.go.kr

Abstract

Collection Failure Is Not Nonresponse – The Infrastructural Selectivity of Reception Visibility in AI-Based Cultural Reception Research

Byongsoo Nam (Center for Comparative Literature, Institute of Humanities, Yonsei University, Research Fellow)

This article examines the risk of interpreting failed online data collection as public nonresponse. It audits records concerning twenty-two films by Hong Sang-soo and fourteen novels by Choi In-hun across fifty-two collection routes. The resulting 1,872 work-source records are treated as units of procedural inspection, not as counts of independent responses. The analysis distinguishes three diagnostic branches applied to 347,992 stored rows: exact-text repetition, contextual and classification checks, and source-score combinations. These checks address different questions and do not establish a representative audience sample. The article distinguishes access failure, unresolved work context, classification failure, and non-observation within a specified observation window. It proposes the infrastructural selectivity of reception visibility to describe how platform access and researchers' processing rules jointly determine what becomes available as reception evidence. Its contribution is to locate the point at which technical non-observation can become an unwarranted judgment about cultural silence. Preserving document context and recording the conditions of exclusion make this judgment open to scrutiny. The argument connects critical infrastructure studies with data auditing while keeping file preservation distinct from semantic and interpretive validation. It thereby identifies evidential conditions that should precede claims about audience attitudes or cognitive experience.

www.kci.go.kr

Keywords: AI-based cultural research, platform infrastructure, reception visibility, data audit, failure ecology, construct validity, online reception

논문 투고일: 2026년 08월 17일
심사 완료일: 2026년 09월 17일
게재 확정일: 2026년 10월 15일

www.kci.go.kr