한국심리학회지 : 상담 및 심리치료
The Korean Journal of Counseling and Psychotherapy
2026, Vol. 38, No. 3, 1123-1144
http://dx.doi.org/10.23844/kjcp.2026.08.38.3.1123

질적 연구에서 인간 연구자와 LLM의 판단 패턴 비교: 사회정의 옹호 역량 분류 과제를 중심으로*

신 나 라†

연세대학교 미래융합연구원 연구교수

본 연구는 상담 및 심리치료 분야의 질적연구에서 동일한 이론적 준거를 적용했을 때 인간 연구자와 생성형 대규모 언어모델(LLM)에서 나타나는 판단 패턴을 비교·분석하였다. 이를 위해 미국상담학회(ACA)의 사회정의 옹호 역량 모형을 이론적 준거로 설정하고, 대학상담사의 사회정의 옹호 상담 경험 인터뷰에서 도출된 32개 진술문을 GPT-SG 모델로 분류하였다. 이를 상담학 질적연구 전문가 3인의 합의 결과(reference standard)와 비교한 결과, 전체 일치율은 53.1%로 나타났으며 유형에 따라 일치와 불일치 양상이 확인되었다. 불일치 사례에서는 LLM이 개입 수준과 옹호 개념을 인간 연구자와 다르게 적용하거나, 다층적인 맥락 단서와 상담 전문직의 역할에 대한 이해를 충분히 반영하지 못하는 판단 패턴이 나타났다. 본 연구의 결과는 인간 연구자와 LLM의 판단 과정 비교를 통해 사회정의 옹호 개념에 포함된 해석적·맥락적 판단 요소를 가시화할 수 있음을 보여주며, 질적연구에서 AI가 연구자의 개념 적용 기준을 점검하고 분석적 성찰을 지원하는 도구로 활용될 가능성을 시사한다.

주요어 : 생성형 인공지능, 대규모 언어모델, 질적 자료 분석, 사회정의 옹호

* 본 논문은 2023년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2023S1A5B5A16075034).

† 교신저자 : 신나라, 연세대학교 미래융합연구원 연구교수, 03722, 서울시 서대문구 연세로50 연세대학교 백양관 N503호, TEL: 02-2123-6381, E-mail:narassyn@yonsei.ac.kr

상담 및 심리치료 분야에서는 인간의 경험과 의미 형성을 개인적·관계적·사회문화적 맥락 속에서 이해하는 것이 중요하다. 질적연구는 인터뷰, 관찰, 문서 등을 통해 인간 경험의 의미와 맥락을 심층적으로 해석하는 방법론으로(Denzin & Lincoln, 2011), 상담심리 분야에서 개인의 경험과 의미 구성, 관계적 맥락을 이해하기 위한 주요 연구 접근으로 활용되어 왔다(김지연, 2018). 특히 내담자의 삶의 맥락, 상담자의 실천 경험, 관계적 상호작용의 의미를 탐색함으로써 상담 이론과 실제를 연결하는 역할을 한다(김지연, 2018). 질적연구에서 코딩은 단순히 자료를 분류하는 절차가 아니라, 자료에 특정 개념과 이론을 적용하여 의미 단위를 판단하고 범주화하는 핵심적 분석 과정이다(신경림 외, 2004). 이 과정에는 연구자의 이론 이해, 맥락에 대한 해석, 가치 판단 등이 복합적으로 작동하므로, 질적 코딩은 본질적으로 해석적 성격을 지닌다(강수정, 2024; Bano et al., 2023).

최근 생성형 인공지능(Generative Artificial Intelligence [AI])의 발전과 함께 상담 및 심리치료 분야에서도 인공지능 활용 가능성에 대한 관심이 증가하고 있다(강수정, 2024; 김도연 외, 2020; 전준, 2024). 그러나 이러한 연구들은 주로 상담 장면에서의 활용 가능성, 상담 보조 도구, 윤리적 고려에 초점을 두고 있으며, 질적연구 방법론 자체에서 인공지능이 어떠한 역할을 수행할 수 있는지에 대한 경험적 검토는 아직 제한적이다. 한편 질적연구 방법론의 맥락에서는 대규모 언어모델(large language models [LLM])을 활용하여 자료 분석을 보조하려는 시도가 점차 확대되고 있다(김윤하, 2025; 박종도 외, 2025; 황윤아, 김영순, 2025). 관련 연구들은 크게 두 흐름으로 전개되어 왔다. 하나는 LLM을 초기 코드 제안, 주제 탐색, 요약 단계에서 연구자의 탐색적 분석을 지원하는 도구로 활용하는 방식이며(김윤하, 2025; Hamilton et al., 2023; Li et al., 2024), 다른 하나는 사전에 정의된 이론적 범주 체계를 적용하여 LLM의 분류 결과를 인간 분석과 비교·검증하는 접근이다(Crocker et al., 2025; Qiao et al., 2024). 이러한 연구들은 특정 조건에서 LLM이 인간 연구자와 일정 수준 유사한 분석 결과를 산출할 수 있음을 보고하였으나(Li et al., 2024), 논의의 초점은 대체로 정확도나 일치도와 같은 수행 지표에 있다.

이러한 맥락에서 인간 연구자와 LLM의 수행 결과를 비교하는 연구가 증가하고 있다. Li 등(2024)은 보건의료 자료 분석에서 GPT-4와 인간 연구자의 질적 분석 결과를 비교하였고, Prescott 등(2024)은 인간과 생성형 AI가 수행한 주제분석의 효율성과 결과 특성을 검토하였다. Hill 등(2026)은 블라인드 방식으로 LLM과 인간 분석자의 주제분석 수행을 비교한 결과, 연역적 코딩에서는 두 주체가 유사한 수행을 보인 반면 귀납적 분석에서는 LLM의 성능이 가변적임을 확인하였으며, 전문가 합의를 비교의 기준점으로(reference standard) 활용하는 방법론적 틀을 제시하였다. 또한 Prinzing 등(2025)은 영성과 같이 단일 지표로 환원되기 어려운 복합적 심리구성 개념에 대해 인간과 GPT의 판단 결과를 비교하여 높은 수준의 일치 가능성을 보고하였다. 그러나 동시에 어떠한 프롬프트와 모델 조건에서 신뢰성과 타당도를 확보한 결과가 산출되는지에 대한 체계적 이해가 충분하지 않음을 지적하며, 인간과 LLM의 결과가 유사하게 나타나는 경우에도 그 판단 과정까지 동일하다고 볼

수 없다는 점을 밝히고 있다(Prinzing, et al., 2025). 신나라와 조한규(2026)의 연구에서 동일한 자료와 프롬프트를 적용하여 4개의 LLM 모델 수행 결과를 비교한 결과, 행동 단서가 명시적인 유형에서는 모델 전반에 걸쳐 높은 일치율이, 맥락 해석이 요구되는 유형에서는 일관되게 낮은 일치율 패턴이 나타났다. 이는 동일한 프롬프트 조건에서도 유형에 따라 수행 양상이 달라질 수 있음을 보여주며, LLM의 분류 결과를 이해하기 위해서는 프롬프트 조건뿐 아니라 과제가 요구하는 판단의 특성을 함께 고려할 필요가 있음을 시사한다.

전준(2024)은 질적연구와 계산사회과학적 접근이 연구 대상, 의미 구성, 해석 방식에 대해 서로 다른 인식론적 전제를 가지고 있다고 보았다. 이러한 관점에서 인간 연구자와 LLM의 비교는 어느 한쪽의 우열을 평가하기 위한 과정이라기보다, 동일한 개념을 적용할 때 어떠한 판단 기준과 해석 초점이 활용되는지를 검토하기 위한 방법론적 시도로 이해될 수 있다. 특히 질적 자료의 분석 코딩에서 중요한 것은 단순한 일치 여부보다, 일치가 나타나거나 차이가 발생하는지를 보다 구체적으로 이해하는 것이다. 동일한 이론적 준거가 주어졌음에도 인간 연구자와 LLM이 서로 다른 분류를 수행한다면, 이는 단순한 오류라기보다 판단 단서와 해석 기준이 다를 가능성을 보여준다. 반대로 일치가 반복적으로 나타나는 패턴을 보인다면 언어적 단서만으로도 비교적 안정적인 판단이 가능한 범주일 수 있다. 그러나 기존 연구의 상당수는 주로 보건의료, 교육, 일반 주제분석 맥락에서 인간과 LLM의 수행 수준을 비교하거나 프롬프트를 최적화하는 데 초점을 두어 왔다. 이에 비해 인간 연구자와 LLM의 판단 차이가 어떠한 유형에서, 어떠한 방식으로 나타나는지에 대한 경험적 검토는 상대적으로 부족하다. 따라서 인간과 LLM의 수행 능력을 비교하고 평가하기보다는 인간 연구자가 암묵적으로 활용하는 판단 특성과 해석의 초점을 간접적으로 드러내기 위한 방법론적 이해를 통해 질적 코딩 과정에서 어떠한 판단이 명시적 단서에 기반하고 어떠한 판단이 맥락적 해석을 요구하는지를 탐색할 필요가 있다.

본 연구는 이러한 문제의식을 바탕으로 사회정의 옹호(social justice advocacy)를 분석 주제로 선정하였다. 사회정의 옹호는 개인의 심리적 어려움을 개인 내부의 문제에 한정하지 않고, 차별, 배제, 자원 접근의 불평등, 제도적 장벽과 같은 사회·문화·구조적 맥락과 연결하여 이해하며, 상담자가 내담자의 권익 증진과 환경 변화에 전문적으로 개입하는 실천 개념으로 발전해 왔다(Lewis et al., 2002; Toporek & Daniels, 2018). 이는 상담자의 역할을 개인 상담 장면에만 제한하지 않고 학교, 기관, 지역사회, 정책 환경까지 확장하여 이해한다는 점에서 상담 및 심리치료 분야의 중요한 전문직 가치로 논의되고 있다(임은미, 2017; 최가희, 2018). 미국상담학회(American Counseling Association, [ACA])는 개입의 방향(내담자와 함께하는 옹호 vs. 내담자를 대변하는 옹호)과 개입 수준(개인, 체계·지역사회, 사회·공공정책)에 따라 사회정의 옹호 역량 모델을 제시하였으며 이를 6가지 유형(A~F)으로 구분할 수 있다(그림 1). A유형은 내담자의 자기옹호 역량을 강화하는 개입, B유형은 상담자가 내담자를 대신하여 옹호하는 개입, C유형은 지역사회 협력과 자원 연계를 통한 개입, D유형은 조직 및 체계 수준의 변화를 촉진하는 개

그림 1. ACA 사회정의 옹호 역량 모형(신나라, 조한규, 2026)

입, E유형은 공공 인식 증진 활동, F유형은 정책 및 입법 수준의 사회정치적 옹호를 의미한다. 이 모형은 상담자의 개입 대상이나 변화 수준을 포괄적으로 반영하고 있어 상담사의 전문적 역할과 책임이 어떠한 수준에서 수행되는지를 이해할 수 있는 대표적 준거틀로 활용되어 왔다(Lewis et al., 2002; Toporek & Daniels, 2018; 임은미, 구자경 2019). 예를 들어 내담자에게 정보를 제공하는 동일한 행동이라도, 그것이 개인 지원인지, 조직 개입인지, 구조 변화 지향인지에 대한 해석은 개입의 출발점, 변화의 지향점, 상담자의 역할과 책임 위치에 대한 판단에 따라 달라질 수 있다(이재복 외, 2022; 김민정, 최한나, 2021; 김지영, 정은선, 2019). 따라서 사회정의 옹호 유형 분류는 동일한 이론적 준거가 주어졌을 때 인간 연구자와 LLM의 판단 특성을 비교하기에 적절한 과제라 할 수 있다. 이때 이론적 관점과 맥락 해석에 따라 판단이 달라질 수 있는 질적 해석의 구성적 성격을 고려하면, 전문가 합의 분류는 절대적 정답(gold standard)이 아니라 이론적으로 숙련된 판단의 비교 준거(reference standard)로 기능한다(Hill et al., 2026; O'Connor & Joffe, 2020).

이에 본 연구는 상담 및 심리치료 분야의 질적 연구 맥락에서 사회정의 옹호 역량 유형 분류라는 연역적 코딩 과제를 통해 생성형 대규모 언어모델과 인간 연구자의 분류 수행 결과를 비교하고, 일치 및 불일치 사례에서 나타나는 판단 특성을 탐색하고자 한다. 이를 통해 이론적 개념을 적용하는 과정에서 어떠한 판단 기준이 중요하게 작동하는지를 살펴보고, 질적연구에서 LLM 활용이 갖는 방법론적 의미를 검토하고자 한다. 연구 문제는 다음과 같다.

첫째, 생성형 대규모 언어모델은 상담 및 심리치료 분야의 이론 기반 코딩 과제에서 인간 연구자의 합의 분류와 비교할 때 어떠한 수행 양상과 유형별 분류 특성을 보이는가?

둘째, 생성형 대규모 언어모델과 인간 연구자 간 분류 일치 및 불일치 사례에서 나타난 판단 특성의 차이는 상담 및 심리치료 분야 질적연구에서 LLM 활용에 대해 어떠한 방법론적 시사점을 제공하는가?

방 법

연구 설계

본 연구는 단순히 LLM과 인간 연구자의 분류 일치도를 평가하는 것을 넘어, 동일한 이론적 준거를 적용하는 과정에서 나타나는 일치 및 불일치 양상을 비교·분석함으로써 두 주체의 판단 특성과 해석 기준을 탐색하도록 설계하였다. 본 연구에서 활용한 분류 과제는 선행 연구에서 다루어진 이진 분류나 단일 속성 평정과는 구별되는 구조적 특성을 지닌다(이상혁, 김은미, 2025; Prinzing et al., 2025).

ACA 사회정의 옹호 역량 모형의 6개 유형은 개입의 방향(함께하는 옹호 vs. 대변하는 옹호)과 개입의 수준(미시적・체계적・거시적)이라는 두 축이 교차하는 구조이다. 이에 따라 유형 간 개념적 경계가 중첩될 수 있고, 분류 과정에서 진술문의 맥락과 관계적 의미에 대한 종합적 해석이 요구된다. 이러한 과제 특성으로 인해 동일한 이론적 준거가 주어지더라도 유형에 따라 판단 양상이 달라질 수 있으며, 이는 LLM과 인간 연구자가 이론적 범주를 적용하는 과정에서 나타나는 판단 차이를 살펴보기에 적합한 조건을 제공한다.

질적 연구에서 코딩은 객관적 사실의 단순한 분류가 아니라 연구자의 이론적 관점과 해석이 개입된 의미 구성 과정이다(O'Connor & Joffe, 2020). 따라서 본 연구에서는 상담학 질적연구 전문가 3인의 합의 분류를 절대적 정답(gold standard)이 아니라, 특정 이론적 틀에 근거하여 형성된 해석적 판단 기준의 비교 준거(reference standard)로 설정하였다.

본 연구는 신나라(2025)의 대학상담사의 사회정의 옹호 경험 연구에서 도출된 32개의 진술문과 이에 대한 후속 LLM 분류 및 오류 특성 분석 결과(신나라, 조한규, 2026)를 활용한 재분석(re-analysis) 연구이다. 선행 연구에서 비교한 4개의 생성형 언어모델(GPT-OSS, GPT-SG, Qwen3, Gemma3) 가운데 인간 연구자와 가장 높은 일치율을 보인 GPT-SG 모델을 분석 대상으로 선정하고, 인간 연구자의 합의 분류와 비교하여 일치 및 불일치 사례에서 나타나는 판단 특성과 해석적 함의를 심층적으로 분석하였다.

자료 수집 및 분석

원자료 및 진술문 도출 과정

본 연구에서 분류 과제에 활용한 자료는 신나라(2025)의 대학상담사의 사회정의 옹호 상담 경험 연구에서 제시된 사회정의 옹호 경험 진술문 32개이다. 원자료 수집을 위해 대학상담센터에서 5년 이상 근무한 상담사 4명을 대상으로 1:1 심층면담을 실시하였으며, 인터뷰는 참여자당 1~2회, 회기당 평균 90분 동안 진행되었다. 원 연구에서는 전사된 축어록을 토대로 질적 연구 전문가 3인이 반복적 검토와 논의를 통해 공통 주제와 범주를 도출하였다. 이 과정에서 인터뷰 자료는 분석에 적합한 문단 단위로 범주화되었고, ACA 사회정의 옹호 역량 모형에 따라 분류되었다. 최종적으로 15개의 공통 주제와 48개의 개별 주제가 도출되었으며, 유형별 특징을 드러내는 대표 진술 32개가 예시로 제시되었다.

본 연구는 원자료 전체가 아닌 32개 진술문을 분석 단위로 활용하였다. 이 진술문은 ACA 사회정의 옹호 역량 유형의 내용을 보여주는 사례로 제시되어, 유형별 핵심 특성과 범주 간 차이를 비교하기에 적절한 자료이다. 또한 특히 장문의 원자료를 그대로 입력할 경우 LLM 내부에서 수행되는 요약・압축 과정이 비가시적이므로, 텍스트 해석의 차이를 최소화하고 동일 자료 안에서 드러나는 판단의 차이를 드러내기 용이하다는 장점이 있다. 수행 결과를 비교하기 위한 분석 조건을 확보하기 위해 동일한 진술문을 인간 연구자와 LLM 모두에게 제공하였다.

인간 합의 분류 구성 과정

본 연구에서 인간 합의 분류는 다음 절차를

통해 구성되었다(신나라, 2025). 상담학 질적 연구 경험을 지닌 전문가 3인은 모두 10년 이상의 대학 상담 경력과 한국상담심리학회 또는 한국상담학회 1급 자격을 보유하였다. 사회정의 상담 관련 교육과 수퍼비전을 10회 이상 이수하였으며, 대학 상담자를 대상으로 5년 이상 강의 및 수퍼비전을 수행하였다. 분석자들은 먼저 ACA 6유형 정의와 판별 규칙에 근거하여 진술문으로 독립적으로 분류하였다. 이후 1차 분류에서 불일치한 항목에 대해 각자의 판단 근거를 제시하고 논의하였으며, 합의 과정을 거쳐 최종 분류를 확정하였다. 합의 과정에서 연구자들은 (1) 개입의 주된 목적(변화의 지향점), (2) 개입이 작동하는 수준(개인 - 체계/지역사회 - 정책), (3) 상담자의 역할 위치(내담자와 함께/내담자를 대변)의 세 가지 기준을 중심으로 선정하였다. 예를 들어, 내담자가 속한 대학 조직 내에서 이루어진 개입은 '체계'로 분류하였고, 대학 외부 기관과의 연계는 지역사회 협력으로 구분하였다. 또한 상담센터 내부 개입과 상담센터 외 조직 개입을 구분하고, 이에 따른 개별 옹호 활동을 분류하였다. 이러한 논의를 통해 상담자가 프로그램을 연계하거나 외부 자원을 연결한 사례에서 이를 지역사회 협력(C)과 체계 옹호(D)로 구분하는 기준을 명확화하였다.

연구자들은 선입견을 배제하고 해석 왜곡을 최소화하기 위해 수차례에 걸쳐 의견을 교환하였으며, ACA 유형 정의를 반복적으로 검토하면서 합의에 도달하였다. 이러한 절차는 질적 코딩이 해석적 판단을 수반하는 수행이라는 점을 전제로 하여(Bano et al., 2023), 판단 근거의 명료화와 일관성 확보를 위한 과정으로 수행되었다(Denzin & Lincoln, 2011). 본 연구에서는 이 합의 결과를 LLM 분석의 비교 준거로 설정하고, LLM의 분류 결과와 대응시켜 일치 및 불일치 양상을 분석하였다.

LLM 분류와 인간 합의 분류 간의 일치도는 전체 정확도와 함께 Cohen's κ를 산출하였다. Cohen's κ는 관찰된 일치율에서 우연에 의해 발생할 수 있는 일치 가능성을 보정한 지표로, 다음 식에 따라 계산된다. $\kappa=(po-pe)/(1-pe)$ 여기서 po는 관찰된 일치율, pe는 우연적 일치율을 의미한다(신나라, 조한규, 2026; Landis & Koch, 1977). 본 연구에서는 6개 범주 분류 과제에서 GPT-SG의 분류 결과와 인간 합의 분류 간 일치도를 평가하기 위해 Cohen's κ를 사용하였으며, Landis와 Koch(1977)의 기준에 따라 0.41~0.60을 중간 수준의 일치(moderate agreement)로 해석하였다.

LLM 분석 및 실행 조건

본 연구에서는 LLM의 분류 수행을 비교하기 위해 모델 실행 조건의 일관성을 통제하였다. 분석에는 구조화된 지시문 준수와 출력 일관성이 강화된 공개 가중치 기반 모델인 gpt-oss-safeguard-20b(이하 GPT-SG)를 활용하였다(OpenAI, 2025). 선행 연구(신나라, 조한규, 2026)에서는 동일 자료와 프롬프트를 적용하여 4개 LLM 모델(GPT-OSS, GPT-SG, Gemma3, Qwen3)의 분류 수행을 비교하였다. GTP-SG 모델이 가장 높은 일치율을 보였으며 출력 형식 준수와 재현성이 상대적으로 높아 본 연구의 비교 분석에 활용하였다.

모델 추론은 공개 가중치를 활용하여 로컬 실행 환경(LM Studio)에서 수행하였다. 외부 서버를 경유하지 않는 로컬 추론 방식은 질적 연구 자료에 포함될 수 있는 민감 정보를 위한 보호 조치이다. 또한 동일 프롬프트를

반복 실행했을 때 같은 결과가 나오도록 temperature 값을 0으로 고정하여 확률적 변동을 최소화하였다. 이는 성능 극대화를 위한 설정이 아니라, 일치 및 불일치가 우연적 생성 변동이 아니라 판단 기준의 차이에서 비롯되었는지를 분석하기 위한 방법론적 조건이다.

모든 진술문에는 동일한 프롬프트 구조와 출력 형식을 적용하였다. 프롬프트는 시스템 프롬프트(system prompt)와 사용자 프롬프트(chat prompt)의 두 부분으로 구성되었다(표 1). 시스템 프롬프트에는 모델의 역할과 분석 원칙, ACA 사회정의 옹호 역량 모형의 유형 정의를 포함하였다. 구체적으로 모델이 상담심리 및 사회정의 옹호 분야의 전문가 역할을 수행하도록 설정하고, 개입의 방향(함께하는 옹호・대변하는 옹호)과 개입 수준(개인・체계・사회)을 고려하여 단일 범주를 선택하도록 지시하였다. 사용자 프롬프트에는 분석 대상 진술문을 제시하고, 해당 진술문을 ACA 사회정의 옹호 역량 모형의 여섯 유형(A – F) 중 하나로 분류한 후 그 근거를 기술하도록 요청하였다. 모든 진술문은 동일한 시스템 프롬프트와 출력 형식을 유지한 상태에서 독립적으로 입력되었으며, 각 진술문에 대해 단일 ACA 유형 또는 '해당 없음'과 그 판단 근거를 산출하도록 하였다.

표 1. ACA 사회정의 옹호 역량 분류를 위한 프롬프트 구조(신나라, 조한규, 2026)

구분	내용
시스템 프롬프트	# 시스템 역할 # 당신은 상담심리 및 사회정의 옹호(Social Justice Advocacy) 분야의 권위 있는 전문가입니다. 당신의 임무는 상담자의 발화(인용구)를 분석하여, '미국상담학회(ACA) 옹호 역량 모형(Advocacy Competencies)'의 6개 영역 중 가장 적절한 하나의 코드로 분류하는 것입니다. # 분석 원칙 # - 맥락적 해석: 단어의 사전적 의미보다 개입의 방향(함께/대변)과 범위(개인/시스템/사회적 수준)를 종합적으로 고려하여 영역을 도출합니다. - 단일 분류: 가장 비중이 높은 영역 코드 1개(A~F)만 부여합니다. - 예외 처리: 발화에 사회정의 옹호 의도나 기능이 없으면 "해당사항 없음"으로 분류합니다. # 옹호 역량 상세 정의 # [각 6가지 옹호 역량에 대한 영문 정의 입력] # 출력 형식 # | 분석 영역 | 분석 근거 | | (A~F 또는 해당없음) | (상세 설명) |
사용자 프롬프트	# 진술문 제시 # "다음 상담자 발화를 ACA 사회정의 옹호 역량 모형의 여섯 유형(A – F) 중 가장 적절한 하나의 범주로 분류하고, 판단 근거를 제시합니다." [진술문을 하나씩 입력]

번역 과정에서 발생할 수 있는 개념적 왜곡을 방지하고 분류 기준의 이론적 일관성을 유지하기 위해, ACA 모형의 영어 원문을 시스템 프롬프트에 사용해 분류 기준의 개념적 일관성을 유지하였으며, 사용자 프롬프트에 입력한 진술문은 번역 과정에서 발생할 수 있는 의미 손실과 편향을 최소화하기 위해 한국어 원문을 그대로 사용하였다. 프롬프트의 내용은 인간 전문가 3인의 합의 분류에 사용된 ACA 사회정의 옹호 역량 모형의 정의와 판별 기준(Toporek & Daniels, 2018)과 동일하다. 인간 연구자와 LLM이 동일한 이론적 준거를 적용하도록 설계함으로써 비교 조건의 개념적 일관성을 확보하고자 하였다.

본 연구에서 프롬프트는 인간 연구자가 활용한 ACA 사회정의 옹호 역량 모형의 개념적 준거를 생성형 언어모델에 제공하기 위한 목적으로 사용되었다. 사용자 프롬프트에는 분석 대상 진술문을 제시하고 단일 범주와 판단 근거를 산출하도록 요청하였다. 이러한 구성은 특정 프롬프트의 우수성을 검증하거나 최적화하기보다는 LLM에 일관된 분석 조건을 제공하고, 인간 연구자와 동일한 이론적 준거에 기반한 분류 결과를 비교하기 위한 것이다. 이에 본 연구에서는 이론적 준거와 분류에 필요한 지시를 제공하는 수준에서 프롬프트를 구성하였으며, 결과는 이러한 프롬프트 조건 내에서 해석할 필요가 있다.

윤리적 고려

상담학 질적연구 자료는 참여자의 경험과 가치 판단을 포함하는 민감 텍스트이다. 생성형 언어모델 활용 과정에서는 개인정보 보호와 자료의 2차 이용, 해석 책임의 윤리적 문제가 중요하게 제기될 수 있다(Hitch, 2024). 본 연구는 인터뷰 원자료 전체를 분석 대상으로 사용하지 않고, 선행연구에 제시된 익명화된 진술문을 분석 자료로 활용하였다. 해당 진술문은 개인 식별 정보가 제거된 상태로 구성되어 있으며, 이론적 범주와의 대응 관계를 검토하는 데 필요한 최소한의 정보만을 포함한다. 연구 참여자들에게는 자료의 2차 연구 활용에 대한 동의를 받은 범위 내에서 자료를 사용하였으며, 참여자 개인정보 및 비밀보호 원칙을 침해하지 않는 범위에서 연구를 수행하였다.

또한 본 연구에서는 생성형 언어모델을 외부 서버에 자료를 입력하는 방식이 아니라 LM Studio의 로컬 환경에서 추론을 수행하는 방식을 채택하였다. 로컬 환경에서의 분석은 외부로 공유되지 않으므로 질적 연구 자료의 기밀성과 연구 참여자 보호 원칙을 보다 엄격하게 준수할 수 있다. 또한 윤리적 정당성을 위해 사용된 LLM 모델 버전, 실행 조건, 프롬프트 구조는 분석 과정의 일부로 투명하게 보고하였다(Denzin & Lincoln, 2011; Hitch, 2024).

본 연구에서 LLM의 유형 분류는 그 자체로 인간 전문가 합의한 분류를 대체하는 최종 코딩의 결과물로 사용되지 않았다. LLM의 판단을 무비판적으로 수용하여 데이터에 적용한 것이 아니라 인간 연구자 분류와의 비교를 위한 분석 자료로 삼고, LLM의 분류가 이 기준과 어떻게 일치하고 달라지는지에 대해 인간 연구자의 판단 하에 해석하였다. 즉 LLM은 본 연구에서 해석의 주체가 아니라 인간 연구자의 판단을 비교하는 보조적 분석 도구로 활용되었다.

결 과

LLM 분류 과제 전체 수행 양상

본 연구는 상담학 질적연구 전문가 3인의 합의 분류를 비교 준거로 설정하고, 동일한 32개 진술문에 대해 GPT-SG 모델이 수행한 분류 결과와의 일치 여부를 비교하였다. GPT-SG 모델과 인간 연구자의 전체 일치율은 53.1%(17/32)로 나타났으며, Cohen's κ=.458, 95% CI [.25, .67]로 산출되었다(해당없음 1건 제외). 이는 중간 수준의 일치로, 우연 일치를 고려하더라도 두 분류 결과 간에 일정 수준의 대응 관계가 존재함을 보여준다. 다만 표본 크기가 작아 신뢰구간이 비교적 넓다는 점에서 이 값은 신중하게 해석될 필요가 있다. 본 연구의 분석 단위는 32개 진술문으로, 유형별 사례 수가 적다(표 2 참조). 따라서 유형별 일치율은 한두 사례의 변동에도 수치가 크게 달라질 수 있으므로, 이를 확정적 비율이 아니라 경향성을 보여주는 지표로 해석하였다. 이에 본 연구는 개별 수치의 크기보다 어떤 유형에서 일치와 차이가 나타나는지, 그 차이가 어떠한 방향으로 반복되는지에 초점을 두어 결과를 제시한다. 본 연구의 분류 과제는 ACA 사회정의 옹호 역량의 6개 범주(A-F) 중 하나를 선택하는 단일 선택 구조이므로, 별도의 기준 없이 무작위로 범주를 선택할 경우 기대 일치율은 1/6(16.7%)이다. 따라서 GPT-SG 모델의 53.1% 일치율은 단순한 우연적 선택이나 임의 추정보다 높은 수준의 분류 수행을 보였으며, 제시된 이론적 범주 정의를 바탕으로 진술문을 일정 수준 일관되게 분류할 수 있음을 시사한다.

그러나 동시에 전체 사례의 약 절반에서는 인간 전문가와 LLM의 분류가 다르게 나타났다(표 2). 이는 동일한 이론적 준거틀이 주어지더라도 인간 연구자와 LLM이 서로 다른 판단 기준이나 해석 초점을 적용할 수 있음을 보여준다. 따라서 본 연구에서 이러한 판단 차이는 단순한 오류나 실패로 보기보다, 질적 연구에서 개념 적용이 어떠한 해석 과정을 거쳐 이루어지는지를 드러내는 분석 자료로 이해하였다(Bano et al., 2023; Hitch, 2024). 이에 따라 연구 결과는 단순 일치율 자체보다는 어떤 유형에서 일치와 차이가 나타났는지, 그리고 차이가 어떠한 방식으로 반복되는지에 초점을 두고 결과를 제시하였다.

먼저 유형별 결과를 살펴보면, A(내담자/학생 역량 강화), B(내담자/학생 옹호), D(체계 옹호) 유형에서는 비교적 높은 일치율이 나타난 반면, C(지역사회 협력), E(공공정보 제공), F(사회·정치적 옹호) 유형에서는 낮은 일치율이 확인되었다(표 2). 특히 C와 F 유형은 무작위 기준선(16.7%) 수준에 머물렀다. 이는 GPT-SG 모델이 유형에 따라 상이한 수행을 보임을 의미하며 행동 목표나 개입 대상이 비교적 명시적인 유형에서는 안정적인 분류를 보였으나, 개입 맥락과 변화 수준에 대한 해석이 요구되는 유형에서는 어려움을 보였음을 의미한다.

다음으로 판단 차이가 나타난 사례들을 혼

표 2. GPT-SG 모델의 인간 합의 분류 유형별 일치율(괄호 안은 사례 수)

비교 준거	A	B	C	D	E	F	총
GPT-SG	71.4%(5/7)	83.3%(5/6)	16.7%(1/6)	100%(4/4)	33.3%(1/3)	16.7%(1/6)	53.1%(17/32)

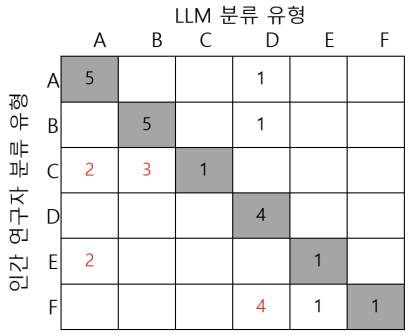

그림 2. GPT-SG 모델의 혼동 행렬

동행렬로 검토한 결과, 이러한 차이는 무작위적으로 분포하지 않고 일정한 경향성을 보였다(그림 2). 혼동행렬을 분석하면 오분류의 방향성이 체계적임을 확인할 수 있다. C유형(6개) 중 GPT-SG가 정확히 분류한 것은 1개(16.7%)이며, 나머지 5개 중 2개는 A유형으로, 3개는 B유형으로 분류되었다. F유형(6개) 중 정확히 분류된 것은 1개(16.7%)이며, 나머지 중 4개는 D유형으로 분류되었다. E유형(3개)은 1개만 정확히 분류되었으며, 나머지 2개는 A유형으로 분류되었다. 반면 D유형은 4개 모두 정확히 분류되어 100%의 일치율을 보였고, A유형(7개) 중 5개(71.4%), B유형(6개) 중 5개(83.3%)가 일치하였다. 해당없음 분류는 GPT-SG에서 1건 발생하였으며, 해당 진술문의 전문가 합의 라벨은 A유형이었다. 이 사례는 혼동행렬의 A~F 범주 구성에 반영되지 않았으며, 전체 32개 진술문 기반 정확도 산출 시 오분류로 처리하였다. 이 결과는 LLM이 사회정의 옹호 의도가 명확하지 않다고 판단하는 경우가 발생할 수 있음을 보여주며, 이는 LLM이 텍스트에서 옹호 관련 언어 단서가 명시적으로 드러나지 않을 때 분류를 유보하는 경향과 관련이 있을 수 있다.

GPT-SG의 근거 설명을 검토한 결과, 정확히 분류된 사례에서는 명시적인 행동 단서와 범주 정의가 직접 연결되는 경향이 나타났다. 반면 오분류 사례에서는 기관 연계, 협력, 지원과 같은 표면적 행위 단서가 주요 판단 근거로 제시되었으나, 해당 행위가 수행되는 관계적 맥락이나 변화의 지향점은 충분히 반영되지 않는 경향이 관찰되었다.

연구 결과를 종합하면, 생성형 언어모델은 특정 조건에서 이론적 범주를 일정 수준 적용할 수 있었으나, 판단 차이가 발생한 지점에서는 인간 연구자와 LLM이 선택하는 의미 단서와 해석 수준이 구조적으로 달랐다. 이러한 결과는 상담 및 심리치료 분야 질적연구에서 이론 적용이 단순한 범주 선택이 아니라, 맥락 이해와 가치 판단을 포함한 해석적 과정임을 보여주고 있다.

유형별 분류 일치 패턴 분석

유형별 분석 결과, A(내담자/학생 역량강화), B(내담자/학생 옹호), D(체계 옹호) 유형에서는 상대적으로 높은 일치율이 나타났다. 인간 연구자와 GPT-SG 모델의 분류가 일치한 사례들을 검토한 결과, 몇 가지 공통적 특성을 살펴볼 수 있었다. 첫째, 상담자의 개입 행동이 구체적으로 서술된 진술문에서 높은 일치율이 나타났다. 예컨대 자원 제공, 외부 기관 연계, 제도 개선 요청 등과 같이 개입의 대상과 행동이 명확히 드러난 사례에서는 모델과 인간 연구자의 판단이 비교적 일관되게 수렴하였다. 둘째, 진술문에 개입의 결과가 직접적으로 제시된 경우에도 일치율이 높게 나타났다. 개입이 특정 제도 변화, 정책 참여, 환경 개선으로 이어졌음이 명시적으로 기술된 사례에서는 모델이 이론적 범주를 비교적 안정적으로 적용

표 3. 일치된 유형별 대표 진술문과 GPT-SG 모델의 판단 근거 예시

유형	진술문	판단 근거(요약)
A	외국인 유학생들의 경우에는 정보를 준다거나 ... 어떤 정보를 같이 찾아주고... 유학생끼리 교류할 수 있는 장을 만들어줘서 같은 나라에서 온 사람들끼리 모일 수 있는 그런 장을 안내해 주거나, 이미 있는 커뮤니티에 연결 이런 것만으로도 ...	정보 제공, 자원 연결, 사회적 지지망 형성이 명시되어 개인 적응과 자기 해결 역량을 강화하는 활동으로 판단. 개인 차원의 지원과 자원 공유는 A유형의 핵심 전략과 일치
B	가정폭력 상황 학생을 위해 ... (중략) 학생에게 정보를 전달하기 위해서 아동보호전문기관, 지역가족센터, 가정폭력상담소 등에 전화를 해서 학생이 어떤 도움을 받을 수 있는지 물어봤던거 같아요. ...학생이 직접 전화할 수도 있지만 미리 상담자가 전화해서 내담자 상황에 맞게 정보를 좀 대신 물어봐주고 학생에게 전달해서 그 학생이 여러 가지 중에 하나를 선택할 수 있었죠.	학생이 학교에서 제공할 수 있는 지원을 넘어 외부 기관에 직접 연락해 필요한 도움과 서비스를 파악하고, 그 정보를 학생에게 전달함으로써 학생이 적절한 자원을 선택하도록 돕는 행위를 수행. 특정 내담자를 위해 외부 시스템에 개입해 장애를 해소하고 필요한 서비스를 연결해 준 것으로 B유형으로 판단.
C	학교가 있는 지역에 자살 예방센터랑 00구 정신건강복지센터가 있는데 이런 지역사회센터들이 관내에 있는 대학교들이랑 협력하기를 원해요. 지역사회센터랑 협업 사업들을 수행하면 위기 학생들을 의뢰할 때 좀 더 신속하게 할 수가 있거든요. 그래서 MOU를 맺거나 하는 업무도 거시적인 면에서는 사회정의 옹호가 될 수 있을거라고 생각해요. 예를 들면 정기적인 행사 중 생명 존중 문화를 조성하는 자살 예방 캠페인이라든가 정신 건강 관련 프로그램을 할 때 협력해서 사업을 하는거죠.	대학과 지역사회 정신건강·자살예방센터가 협력해 위기 학생을 신속히 지원하고, 생명 존중 문화 조성 캠페인 등 공동 사업을 추진, MOU 체결과 같은 협업 과정이 명시됨. 이는 지역사회와 학교가 함께 행동한 공동체 협력에 해당하여 C 유형으로 판단함.
D	청각 장애인 학생 같은 경우에는 상담자가 천천히 말해도 장애 학생은 그 말을 다 빨리 못 알아듣는 경우가 있어요. 근데 이제 이 사람의 말을 AI가 입력해서 글씨로 해 주는 기구가 있어요. 장애학생 센터에서 학생들이 서비스를 요청하지 않으면 도구를 구비해 놓지 않는 경우가 많죠. 그래서 제가 장애지원센터 선생님이 얘기했더니 선생님이 기안을 올려서 결국 그 기구를 다섯 대를 샀어요.	특정 학생을 넘어 접근성 개선을 위한 조직 환경 변화(인프라 도입) 시도라는 구조적 단서가 분명함. 이는 조직(또는 학교) 수준에서 환경적 요인을 인식하고, 시스템 차원의 변화를 추진한 행위이므로 D유형 시스템 옹호에 해당.
E	단과대별로 심리 상담 특강을 요청하거든요. (중략) 학업 스트레스가 크다 보니 단과대에서 정신 건강을 어떻게 하면 잘 돌볼 수 있는지 의사소통을 어떻게 하면 잘 할 수 있는지 이런 것들을 일회성으로 해주기를 바라는 게 있는데 심리 상담과 관련한 공공정보를 제공하는 것이라고 볼 수 있을 것 같아요.	위기 예방과 인식 개선을 지향하는 집단적 공공 메시지 전달로 해석됨. “심리 상담과 관련한 공공정보를 제공하는 것”은 특정 집단(학생들)에게 교육·정보를 전달해 인식 개선 및 자원 활용을 돕는 E 유형에 해당함.
F	심리상담사 법도 자기 옹호 부분에서 중요한 주제라고 생각해요. 사회정의 상담 공부를 하고 나서는 학회에서 오는 공청회 관련된 거, 뉴스나 이런 것들은 스크랩해서 보는 편이에요.	특정 내담자가 아니라 사회·정치적 이슈에 대한 정보를 수집하고, 대중 인식이나 정책 변화를 도모하려는 행위로 해석. 정책/입법 수준의 공적 이슈에 대한 지속적 관찰과 참여가 명시되어 F유형 사회·정치적 옹호로 판단.

하였다. 이러한 결과는 생성형 언어모델이 텍스트 내에서 기능적·행동적 단서를 중심으로 범주를 구성하는 경향이 있으며(Hamilton et al., 2023; Pattyn, 2024), 이론 기반 코딩에서 초기 범주 정렬 단계에서는 일정 수준의 보조적 가능성을 가질 수 있음을 확인하게 한다. GPT-SG 모델이 활용한 판단 근거를 살펴보기 위해 각 유형별로 옹호 행위의 대상과 수준, 상담자의 역할이 비교적 명시적으로 드러난 진술문과 GPT-SG의 판단 근거 예시를 표 3에 제시하였다.

분류 불일치 패턴 분석

C(지역사회 협력), E(공공정보 제공), F(사회·정치적 옹호) 유형에서는 인간 연구자와 LLM의 분류 결과가 불일치하는 비율이 높게 관찰되었다. 특히 일정한 패턴으로 분류하는 경향이 존재하는 것으로 나타났는데, 이는 단순한 기술적 오류라기보다 동일한 진술을 해석하는 과정에서 인간 연구자와 LLM이 서로 다른 판단 수준(level of judgment)을 적용하고 있음을 보여준다. 이러한 분류 불일치를 패턴을 이해하기 위해 옹호 개입이 지니는 다층적 성격이 GPT-SG 모델의 판단 과정에서 어떻게 해석·환원되는가를 중심으로 분석하였다.

인간 연구자는 개입의 표면적 행동뿐 아니라 변화의 지향 수준, 전문직 책임의 확장성, 구조적 맥락에 대한 인식을 통합하여 범주를 판단하였다. 반면 GPT-SG 모델은 진술문에서 가장 명시적으로 드러난 행동 단서와 직접적 개입 대상을 중심으로 범주를 적용하는 경향을 보였다. 그 결과, 분류 불일치는 주로 (가) 개인 - 조직 - 체계 수준의 모호성, (나) 가치·의미 중심 옹호의 행동 단서 환원, (다) 맥락 축소에 따른 옹호 비분류, (라) 전문직 자기옹호의 체계·제도 개입 환원이라는 네 가지 양상으로 정리될 수 있었다. 이러한 차이는 사회정의 옹호 개념이 지니는 다층적 구조와 관련된다. 사회정의 옹호를 개인, 조직, 지역사회, 정책 수준이 교차하는 해석적 구성물로 이해해 왔으며(김민정, 최한나, 2021; 이재복 외, 2022; 임은미, 2017), 개입의 의미는 단일 행동이 아니라 맥락과 책임의 위치 속에서 판단된다고 보았다. 본 연구에서 나타난 반복적 오분류 패턴은 이러한 연구자의 해석적 판단 특성이 LLM의 분류 수행에서는 충분히 반영되지 않음을 드러낸다. 즉, 이러한 결과는 LLM은 언어적 단서에 기반한 확률적 판단을 수행하는 반면(Bano et al., 2023), 인간 연구자의 판단에 가치·맥락·전문적 책임과 관련된 요소가 반영될 가능성을 보여준다.

개입 수준에 대한 판단의 차이

개입의 출발점은 개인 수준이지만 동시에 조직 또는 제도적 요소가 포함된 사례에서 혼동이 반복적으로 나타났다. 이 양상은 두 가지 방향으로 나타났다(표 4 참조).

하나는 공동체 협력 개입(C)이 개인 대변 옹호(B)로 축소 해석되는 경우이다. C-1, C-2, C-6 사례에서 인간 연구자는 상담자가 특정 학생을 돕는 과정에서 기관 간 네트워크를 형성하고 협력 구조를 활용·구축했다는 점에 주목하여 C 유형으로 분류하였다. 반면 GPT-SG 모델은 “학생을 위해 외부 기관에 연락함”, “기관에 연계함”이라는 행동 단서에 판단을 고정하여, 특정 학생을 대신해 체계에 접근한 대리적 개입으로 해석하고 B 유형으로 분류하였다. 다른 하나는 개인 또는 대변 옹호(A, B)가 체계 옹호(D)로 확대 해석되는 경

표 4. 개입 수준에 대한 판단의 차이

유형-번호	진술문 요약	인간 판단 근거	LLM 판단 유형과 근거
C-1	성폭력 피해 학생을 위해 의료·경찰 기관 연계 및 동행 지원	학생이 위기 상황에서 필요한 지원을 선택·활용하도록 돕는 기관 간 협력 행위로 해석	B, 학생을 대신해 외부 체계에 접근한 대리 개입으로 판단
C-2	중독·가정폭력 관련 외부 기관 연계	학생이 스스로 도움을 받을 수 있도록 기관과 협력하여 정보 확대 및 연계/협력	B, 기관 연계 및 의뢰 행위를 대리 옹호로 해석
C-6	법률 프로그램 및 외부 기관과의 네트워크 구축	학생의 문제 해결 역량을 확장하기 위한 기관 간 협력 지원으로 해석	B, 외부 체계 접근을 대신 수행하고 연결한 행위에 초점
A-6	학교 차원의 캠페인·프로그램 예산 지원	학생 참여와 정서·관계 증진이라는 목적에서 출발한 개입으로, 개인의 역량 강화를 위한 프로그램 지원	D, 학교 차원의 예산·운영 구조를 변화시키는 체계 개입으로 판단
B-5	성희롱·성폭력 사건 처리 후 규정 개정	개별 피해 학생의 권익 보호에서 출발한 대변 옹호로 해석	D, 규정 개정이라는 조직 차원의 변화 단서에 근거

우이다. A-6과 B-5 사례에서 인간 연구자는 개입의 출발점과 목적에 주목하여 각각 A·B 유형으로 분류하였으나, GPT-SG 모델은 진술문에 드러난 조직·제도 관련 결과 단서를 중심으로 D 유형으로 분류하였다.

이러한 차이는 인간 연구자가 개입의 출발점과 변화의 지향 수준을 위계적으로 고려한 반면, 모델은 진술문에 가장 명시적으로 드러난 조직·제도 관련 단서를 중심으로 판단을 구성하는 경향을 나타낸다.

옹호 개념 적용의 차이

예방 교육, 인식 개선, 사회정의 실천과 같이 가치 지향적 의미를 포함하는 진술에서는 옹호 개념의 적용에 따라 유형 판단의 차이가 두드러지게 나타났다. 인간 연구자는 해당 진

표 5. 옹호 개념 적용의 차이

유형-번호	진술문 요약	인간 판단 근거	LLM 판단 유형과 근거
E-1	생명 존중·도박 예방 교육	위기 발생 이전에 구조적 위험 요인을 낮추기 위한 예방적 옹호로 해석	A, 교육을 통해 직접적으로 개인 인식·대처 역량 강화로 판단
E-2	성평등·인권 감수성 증진 교육	사회적 불평등과 차별에 대한 집단 차원의 인식 전환을 지향	A, 개인에게 직접적 정보 전달 및 교육 제공이라는 명시적 행동 단서에 근거

술을 사회적 인식 확장이나 구조적 예방 차원의 옹호로 해석하였으나, GPT-SG 모델은 진술문에 명시적으로 드러난 교육, 특강, 정보 제공이라는 행위 단위에 초점을 맞추어 개인 역량 강화 범주로 적용하는 경향을 보였다(표 5).

예를 들어, E-1 사례에서 인간 연구자는 상담자가 심리적 위기 사건이 발생하지 않도록 예방 교육 사업을 주도적으로 수행했다는 점에 주목하여 이를 E 유형으로 분류하였다. 이 사례에서 옹호의 핵심은 특정 개인의 변화가 아니라, 위기 발생 가능성을 낮추는 환경 조성이었다. 그러나 GPT-SG 모델은 해당 진술을 “교육을 제공함”이라는 직접적 개입 행동으로 해석하여 개인의 인식과 대처 역량을 강화하는 A 유형으로 분류하였다. 이는 인간 연구자가 옹호 개념의 규범적·가치적 층위를 함께 고려한 반면, 모델은 기능적 행위 중심으로 범주를 적용하는 차이를 반영한다.

다층적 맥락 단서 해석의 차이

하나의 진술문 안에 개인·조직·사회 수준이 복합적으로 포함된 경우, 인간 연구자는 맥락적 층위를 종합적으로 고려하여 범주를 적용하였다. 반면 GPT-SG 모델은 가장 두드러진 행동 단서 하나에 판단을 고정하는 경향을 보였다(표 6).

F-1 사례의 경우 성소수자 관련 이슈나 F-2 사례에서 청년주거 문제와 같은 사회·정치적 사안에 대한 지속적 관심과 정보 수집이 나타나고 있다. GPT-SG 모델은 ‘모니터링’, ‘참여’, ‘활동’과 같은 행동 단서를 중심으로 분류를 수행한 반면, 인간 연구자는 해당 행동이 놓인 사회·정치적 맥락과 옹호의 의도를 중심으로 해석하였다.

특히 A-7의 집단상담 프로그램의 경우, 인간 연구자는 학생과 학과의 요청에 의해 개인의 역량 강화를 목적으로 개설되었다는 점에서 A유형(내담자/학생 역량강화)로 해석하였다. 반면 GPT-SG 모델은 진술문에 옹호의 필요성이나 환경적 장벽, 상담자의 옹호적 역할이 명시적으로 드러나지 않는다는 이유로 이를 옹호 범주로 분류하지 않았다. 이러한 차이는 다양한 맥락을 병렬적으로 유지하는 인간 연구자의 해석 방식과 단일 단서를 중심으로 범주를 선택하는 모델의 판단 방식 간의 구조적 차이를 드러낸다.

상담 전문직에 대한 이해의 차이

상담 전문직의 처우 개선, 정책 제안, 제도적 발언과 같이 전문 분야의 지식을 요구하거나, 사회적 맥락에서 전문직 정체성과 책임의

표 6. 맥락 축소 및 단일 단서 중심의 판단

유형-번호	진술문 요약	인간 판단 근거	LLM 판단 유형과 근거
F-1	성소수자 관련 제도·사회 이슈 모니터링	사회·정치적 담론과 제도 변화에 개입하려는 옹호로 해석	D, 제도 환경에 대한 인식 활동으로 판단
F-2	청년 주거 시민단체 활동 추적	정책·사회 구조에 대한 공적 옹호 지향	E, 집단 이슈 참여 및 연대 활동으로 해석
A-7	학생·학과 요청에 따른 집단상담 프로그램 개설	개인의 역량 강화를 목적으로 한 역량강화 옹호로 해석	옹호로 분류하지 않음, 장벽·옹호 역할이 명시되지 않음

표 7. 전문직 자기 옹호 해석의 차이

유형-번호	진술문 요약	인간 판단 근거	LLM 판단 유형과 근거
F-4	상담센터 연구직 처우 개선 보고	전문직의 지속 가능성과 사회적 책임 수행을 위한 자기 옹호로 해석	D, 조직 내부 인사·운영 구조 개선으로 판단
F-5	상담 교육비 예산 확보 필요성 제기	전문성 강화를 통한 내담자·학생 옹호의 기반 마련을 위한 제도, 정책으로 해석	D, 자원 및 예산 구조 개선에 초점

확장을 포함하고 있는 진술에서 해석 차이가 나타났다(표 7). 인간 연구자는 이를 사회·정치적 옹호의 범주로 해석하였으나, 모델은 조직 운영 개선이나 체계 정비와 같은 내부적 기능 차원으로 분류하는 경향을 보였다. 이는 인간 연구자가 전문직 책임의 공공적 확장이라는 의미를 해석한 반면, 모델은 조직 단위의 제도 개선이라는 기능적 차원에 초점을 두었다고 해석할 수 있다.

F-4와 F-5 사례에서 인간 연구자는 상담자의 연구직 처우 개선 요구와 상담 교육비 예산 확보 노력을 전문직의 지속 가능성과 사회적 책임 수행을 위한 자기옹호로 해석하였다. 즉, 전문성 강화가 궁극적으로 내담자와 학생에 대한 옹호의 기반을 마련한다는 점에서 이를 F 유형으로 분류하였다. 반면 GPT-SG 모델은 이러한 개입을 조직 내부의 인사·예산 구조 개선으로 해석하여 D 유형으로 분류하였다. 이는 전문직 자기옹호가 갖는 윤리적·정치적 의미보다는, 진술문에 드러난 조직 운영 및 제도 변화 요소에 판단을 고정한 결과로 볼 수 있다.

종합하면, GPT-SG 모델은 무작위 수준을 넘어서는 수행을 보였으나, 분류 불일치가 발생한 지점에서는 인간 연구자와 LLM이 선택하는 판단 단서와 해석 수준이 구조적으로 달랐다. 행동 단서가 명확하고 개입 수준이 직접적으로 기술된 진술에서는 비교적 높은 일치율이 나타났지만, 개입의 의도·가치 지향·책임의 위치·변화의 지향 수준이 복합적으로 얽힌 사례에서는 해석 차이가 두드러졌다. 이러한 결과는 상담 및 심리치료 분야 질적연구에서 이론 적용이 단순한 범주 선택이 아니라, 맥락 이해와 가치 판단을 포함한 해석적 과정임을 시사하며, LLM이 이론적 정의를 무작위로 적용하는 것은 아니지만 상담학 질적연구에서 전제되는 해석적 판단 구조와는 다른 방식으로 범주를 구성하고 있음을 보여준다. 인간 연구자와 동일한 결과를 산출하더라도 그 과정에서 활용되는 판단 단서와 해석 초점에는 차이가 있을 가능성을 시사한다.

논 의

본 연구는 상담 및 심리치료 분야의 질적 코딩에서 인간 연구자와 생성형 대규모 언어모델이 동일한 이론적 준거를 어떻게 다르게 적용하는지를 탐색하였다. 이를 위해 GPT-SG 모델의 분류 결과를 인간 연구자 3인의 합의 분류(reference standard)와 비교한 결과, 전체 일치율은 53.1%, Cohen's κ는 .458로 나타났다. 이는 6개 범주를 임의로 선택할 때 기대되는 무작위 기준선(16.7%)을 상회하는 수준이며,

우연 일치를 고려하더라도 중간 수준의 일치를 의미한다. 이러한 결과는 선행 연구에서 보고된 LLM과 인간 연구자의 일치도 범위(κ ≈.39~.40; Crocker et al., 2025; Li et al., 2024)와 대체로 유사하지만, 본 연구의 신뢰구간이 넓어 직접적인 비교에는 신중한 해석이 필요하다. 또한 본 연구는 개입의 방향과 수준, 관계적 맥락을 동시에 고려해야 하는 ACA 사회정의 옹호 역량의 6개 범주를 분류 대상으로 하였다는 점에서, 이진 분류나 단일 속성 평정 과제를 다룬 선행 연구(이상혁, 김은미, 2025; Prinzing et al., 2025)보다 구조적으로 복잡한 과제를 다루고 있다. 따라서 본 연구의 결과는 LLM의 성능을 평가하기보다, 동일한 이론적 준거가 주어졌을 때 인간 연구자와 LLM의 분류 과정에서 나타나는 판단 특성과 차이를 탐색한 결과로 이해될 필요가 있다.

본 연구에서 나타난 유형별 수행 차이가 프롬프트 설계의 영향인지, 혹은 과제 자체의 특성에서 비롯된 것인지는 중요하게 검토할 필요가 있다. 신나라와 조한규(2026)는 본 연구와 동일한 자료와 프롬프트를 서로 다른 4개의 LLM 모델에 적용하여 도출된 결과를 비교・분석하였다. 그 결과 C유형의 낮은 일치율과 D유형의 높은 일치율 패턴이 4개 모델 전반에 걸쳐 일관되게 나타났다(신나라, 조한규, 2026). 특히 C유형은 33%를 넘는 모델이 하나도 없었으며, D유형은 두 모델에서 100%의 일치율을 나타냈다. 이러한 결과는 유형별 수행 차이가 단순히 특정 모델이나 단일 프롬프트의 영향만으로 설명되기 어렵고, 사회정의 옹호 유형이 요구하는 판단의 특성과 관련될 가능성을 시사한다. 또한 Nguyen-Trung(2025)은 LLM의 연역적 코딩 오류가 의미론적으로 구분 가능한 범주에서도 반복적으로 나타나고 있음을 보고하며, 이를 단순한 프롬프트 설계의 문제가 아니라 이론적 기준을 적용하는 과정에서 나타나는 구조적 한계로 해석하였다. 그러나 이것이 프롬프트 효과를 완전히 배제하는 것은 아니다. Zhuo 등(2024)은 복잡한 추론이 요구되는 과제일수록 프롬프트 민감성이 증가한다는 점을 확인하였으며, 이는 본 연구에서 행동 단서가 비교적 명시적인 A・B・D 유형보다 맥락적 추론이 요구되는 C・E・F 유형에서 반복적으로 오분류가 발생한 결과에 대한 설명이 될 수 있다. 따라서 향후 연구에서 프롬프트 조건과 모델을 비교하여 프롬프트 효과와 과제 특성의 영향을 체계적으로 분리할 필요가 있다.

분석 결과에서 드러난 유형별 수행 패턴은 ACA 사회정의 옹호 역량 모형의 구조적 특성과도 일관된다. A유형(내담자 역량강화)과 B유형(내담자 대변), D유형(체계 옹호)에서 상대적으로 높은 일치율이 나타난 것은 상담자의 역할과 행동 기준이 텍스트에 비교적 명시적으로 드러나는 특성과 관련된다. 반면 C유형(지역사회 협력), E유형(공공정보 제공), F유형(사회・정치적 옹호)은 단순한 행동 단서만으로는 분류가 어려워 개입의 방향, 관계의 성격, 사회문화적 맥락을 함께 고려해야 하는 유형으로 상대적으로 낮은 일치율이 나타났다. 특히 C유형이 주로 B유형으로, F유형이 주로 D유형으로 분류된 패턴은 오분류가 무작위로 발생한 것이 아니라 특정 방향성을 가지고 나타났음을 보여준다. 이는 인간 연구자와 LLM이 동일한 진술문에서 서로 다른 판단 단서를 우선적으로 활용했을 가능성을 시사한다. 이러한 결과는 사회정의 옹호가 단순한 행동의 존재 여부보다 개입의 목적, 책임의 위치, 구조적 맥락에 대한 해석을 요구하는 다층적 개

념(김지영, 정은선, 2019; 임은미, 2017; 최가희, 2018)이라는 점을 잘 보여주고 있다.

특히 ACA 옹호 개념은 상담자가 개인 상담을 넘어 지역 사회와 공공 정책 수준에서 체계 변화를 촉진하는 역할을 수행해야 한다는 관점에서 전문직의 사회적 책임을 강조하는 방향으로 제시되었다(Lewis et al., 2002; Toporek & Daniels, 2018). 반면 한국의 대학 상담사는 대학 조직 내에서 비교적 주변적인 위치에 놓인 채 다양한 행정적·조직적 업무를 수행하며 내담자를 위한 지원 체계를 구축하는 역할을 수행하고 있다는 점에서 그 맥락이 다르게 나타날 수 있다(신나라, 2025). 본 연구에서 인간 연구자와 생성형 언어모델의 판단을 비교한 결과, 한국 상담 현장의 맥락 속에서 상담자의 옹호가 정책 개입의 형태뿐 아니라 상담 실천과 조직 협력 속에서 관계적으로 수행되고 있음을 드러내는 단서를 제공하였다.

본 연구에서 관찰된 오분류 패턴은 인간 연구자와 LLM이 동일한 진술문을 해석할 때 서로 다른 수준의 판단을 활용할 가능성을 보여준다. 인간 연구자는 개입의 변화 지향점, 상담자의 역할 위치, 구조적 맥락을 종합적으로 고려하여 범주를 판단하였으나, GPT-SG는 진술문에 가장 명시적으로 드러난 행동 단서와 직접적 개입 대상을 중심으로 범주를 선택하는 경향을 보였다. 이러한 경향은 LLM이 표면적 언어 단서와 통계적 연관성에 상대적으로 의존한다는 선행 연구 결과와도 일치한다(Bano et al., 2023). 다만 이러한 해석은 LLM의 출력 근거와 오분류 패턴에 대한 탐색적 분석에 기반한 것이며, 인간 연구자의 실제 판단 과정을 직접 분석한 결과는 아니라는 점에서 신중하게 해석될 필요가 있다. 따라서 LLM이 명확한 행동 기준과 역할 정의가 제시된 분류 과제에서는 예비 코딩 보조 도구로 활용될 가능성이 있으나(Hamilton et al., 2023; Mazeikiene & Kasperiuniene, 2024; Pattyn, 2024), 관계적 맥락과 개입의 의미를 종합적으로 조정해야 하는 경우에는 인간 연구자의 해석이 필수적임을 시사한다.

본 연구의 의의는 LLM의 정확도나 오류 자체보다는 인간 연구자가 암묵적으로 적용해 온 판단 기준을 가시화하였다는 것이다. 질적 연구에서 이론을 적용할 때는 연구자의 전문성에 기반한 해석 판단이 요구되지만, 실제로 어떠한 단서와 기준에 의해 판단이 이루어지는지는 충분히 논의되지 않았다. 본 연구에서 드러난 인간–LLM 간 판단 차이는 범주 경계가 어디에서 형성되는지, 연구자가 어떠한 맥락 정보를 중요하게 고려하는지를 보여주는 분석 단위로 기능하였다. 이러한 점에서 인간과 LLM의 비교는 LLM의 수행 능력을 평가하기 위한 절차라기보다는 인간 연구자의 해석 과정을 성찰적으로 검토하기 위한 방법론적 장치로 이해될 수 있다. LLM은 인간 연구자를 대체하는 자동 코더라기보다, 연구자가 개념을 적용하는 과정에서 어떤 판단 기준을 활용하는지를 드러내는 성찰적 비교장치(reflexive comparator)로 기능할 수 있다(김영순 외, 2025; 박종도 외, 2025; Hitch, 2024). 따라서 생성형 언어모델은 질적연구에서 연구자의 해석을 대체하기보다, 분석 과정의 투명성과 성찰성을 높이는 보조 도구로 활용될 가능성이 있다.

이러한 가능성에도 불구하고 질적 연구 과정에서 AI의 활용은 윤리적 검토를 전제로 이뤄져야 한다. 상담 및 심리치료 분야의 질적 자료는 개인의 심리적 경험, 관계적 갈등, 사회적 취약성 등 민감한 내용을 포함하는 경우가 많아 일반 텍스트 자료와 동일한 방식으로

처리되기 어렵다. 따라서 LLM 활용은 단순한 효율성의 문제가 아니라, 연구자의 해석 책임과 참여자 보호 원칙 속에서 검토되어야 한다(강수정, 2024; Hitch, 2024). 본 연구에서도 AI 산출물을 최종 판단으로 사용하지 않고 인간 연구자의 합의 분류를 비교 준거로 설정하였으며, 모델 조건과 프롬프트를 가능한 범위 내에서 통제하여 분석 절차의 투명성을 확보하고자 하였다. 향후 관련 연구에서는 모델 버전, 프롬프트 구조, 자료 비식별화 절차, AI 개입 범위, 결과 해석의 한계를 명시적으로 보고할 필요가 있으며, 기관윤리심의(IRB) 검토 절차 또한 보다 구체적으로 정비될 필요가 있다. 이는 기술 활용의 부가 조건이 아니라 상담 및 심리치료 연구의 윤리적 정당성을 유지하기 위한 핵심 기준이라 할 수 있다.

본 연구는 몇 가지 한계를 지닌다. 첫째, 분석 자료는 유형별 특성을 드러내는 32개 진술문으로 제한되어 있으며, 이 진술문은 연구참여자 4인의 경험에 기초하고 있다. 특히 유형별 사례 수가 적어(예, E유형 3개) 유형별 일치율은 한두 사례의 변동에도 민감하므로, 본 연구의 유형별 수치는 확정적 비율이 아니라 경향성으로 해석되어야 한다. 따라서 실제 인터뷰 원자료를 입력하거나, 다른 경험을 포함하고 있는 진술문을 대상으로 분석할 경우 다른 결과가 나타날 가능성이 있다. 둘째, 본 연구는 ACA 원문과 전문가 합의 결과를 반영하여 프롬프트의 내용 타당성을 고려하였으나, 개발-테스트 분리 설계 및 프롬프트 반복 수정과 같은 절차적 타당도 확보 과정을 거치지 않았다. 프롬프트의 정의 방식, 예시 제공 여부, 역할 지시 수준 등에 따라 분류 결과가 달라질 수 있으므로, 본 연구의 결과는 특정 프롬프트 설계 조건에서 관찰된 판단 패턴으로 이해될 필요가 있다. 셋째, 본 연구는 ACA 모형의 정의를 영어 원문으로 제시하되 진술문은 한국어 원문을 그대로 입력하였는데, 이러한 교차 언어적 제시 조건이 분류 결과에 미치는 영향은 별도로 통제·검증하지 않았다. 지시문과 분석 자료가 서로 다른 언어로 제시될 경우 지시문 언어에 따라 모델 수행이 달라질 수 있다는 보고(Kim et al., 2025)를 고려하면, 향후 연구에서는 지시문 및 자료의 제시 언어 조건을 체계적으로 비교하여 언어 조건이 판단 패턴에 미치는 영향을 검토할 필요가 있다. 넷째, 해석의 한계이다. 본 연구는 인간 합의 분류를 절대적 정답(gold standard)이 아닌 특정 이론 틀에 기반한 해석적 기준점(reference standard)으로 활용하였다. 인간 연구자의 분류와 근거에 대한 분석과 LLM의 분류 근거 설명에 대한 분석은 귀납적·탐색적 성격을 가지며 체계적 코딩 절차를 거친 것이 아니다. 또한 LLM의 분류 근거와 인간 연구자의 실제 판단 과정이나 판단 특성을 직접적으로 설명하는 데에는 한계가 있다. 따라서 본 연구는 수행 결과에서 나타난 판단 차이의 패턴을 기술하고 해석하는 데 초점을 두었으며, 이를 근거로 인간 연구자와 LLM의 판단 특성 차이를 일반화하는 데에는 한계가 있다.

이러한 한계를 바탕으로 향후 연구에 대해 다음과 같이 제언한다. 첫째, 다양한 질적자료와 복수 모델 비교, 프롬프트 조건 변형, 복수 코드 허용 설계 등을 통해 LLM의 판단 패턴이 어떠한 조건에서 달라지는지 검토할 필요가 있다. 둘째, 인간과 LLM 수행 결과의 불일치를 분석하고 그 결과를 프롬프트 수정에 활용할 수 있다. 예를 들어 LLM 결과와 인간 코딩의 불일치를 분석하여 그 원인이 프롬프트의 모호성에서 비롯된 것인지 개념 자체의 모

호성에서 비롯된 것인지를 구분하고, 이를 토대로 프롬프트를 반복 수정하는 절차를 제시한 바 있다(Bunt et al., 2025). 이러한 절차는 보다 타당한 프롬프트를 구성하는 데에도 도움이 될 수 있다. 셋째, 본 연구는 이론적 기준이 명시된 연역적 코딩을 활용하였으므로, 향후 귀납적 질적 분석 맥락에서 LLM이 개념 생성과 범주 정교화 과정에 미치는 영향을 탐색함으로써 인간–AI 협업의 가능성과 한계를 보다 확장된 수준에서 논의할 수 있을 것이다.

상담 및 심리치료 분야 질적연구에서 AI의 활용은 인간 연구자의 역할을 대체하기보다, 인간 연구자의 해석과 판단에 작용하는 기준을 더욱 선명하게 드러내고, 이를 성찰적으로 검토하는 방법론적 도구로 활용될 수 있다. 이는 AI의 활용이 질적연구의 본질을 약화시키는 것이 아니라, 인간 경험의 의미를 해석하는 연구자의 전문성이 무엇인지 학문적으로 재조명하는 계기가 될 수 있음을 시사한다.

참고문헌

강수정 (2024). 상담 및 심리치료에서 대규모 언어 모델(LLM)의 적용 기회와 한계 및 윤리적 고려 사항. **디지털콘텐츠학회논문지, 25**(12), 3751-3759.

김도연, 조민기, 신희천. (2020). 상담 및 심리치료에서 인공지능 기술의 활용: 국외 사례를 중심으로. **한국심리학회지: 상담 및 심리치료, 32**(2), 821-847.

김민정, 최한나 (2021). 상담자의 옹호활동에 관한 개념도 연구. **한국심리학회지: 상담 및 심리치료, 33**(3), 853-881.

김영순, 황윤아, 윤혜림, 우지연 (2025). 질적연구물 연구방법 영역의 인간-AI 협업 과정에 관한 문헌사례연구. **다문화와 교육, 10**(4), 1-25.

김윤하 (2025). 대규모 언어 모델은 한국어 질적 코딩을 수행할 수 있는가? - GPT-4 기반 귀납적 코딩 자동화 가능성 평가. **인공지능인문학연구, 21**, 107-135.

김지연 (2018). 국내 상담학 분야의 질적연구 동향 분석. **질적탐구, 4**(2), 131-168.

김지영, 정은선 (2019). 사회정의 옹호상담을 적용한 상담자 자기옹호에 관한 고찰. **상담학연구, 20**(4), 1-17.

박종도, 김영순, 최은하, 황윤아 (2025). 질적연구에서 자료처리 과정의 인간-AI 협업 경험에 관한 질적 사례연구. **다문화와 교육, 10**(4), 27-51.

신나라 (2025). 대학상담사의 사회정의 옹호 상담 경험 연구. **학습자중심교과교육연구, 25**(1), 761-782.

신나라, 조한규 (2026). 대규모 언어모델 기반 사회정의 옹호 역량 분류: 질적 연구에서의 코딩 결과 일치도 및 오류 특성 분석. **스마트미디어저널, 15**(3), 127-135.

신경림, 조명옥, 양진향, 고명숙, 공병혜, 김강미자, 김경선, 김남선, 김남초, 김미영, 김분한, 김순이, 김애경, 김애정, 김영경, 김영혜, 김영희, 김옥현, 김은하, … 하주영 (2004). **질적연구방법론**. 이화여자대학교출판부.

이상혁, 김은미 (2025). 대규모 언어 모델은 분석 도구가 될 수 있는가? GPT를 활용한 내용 분석의 신뢰도와 타당도를 중심으로. **한국언론학보, 69**(1), 5-38.

이재복, 장소정, 김영신, 남목민, 조훈제, 이수정, 연규진 (2022). 상담자들의 사회정의

옹호역량 발달 과정에서의 경험에 대한 합의적 질적 연구. 한국심리학회지: 상담 및 심리치료, 34(1), 1-34.

임은미 (2017). 한국 상담자를 위한 사회정의 옹호역량 척도(SJACS-K)의 개발 및 타당화. 상담학연구, 18(6), 17-36.

임은미, 구자경 (2019). 다문화 사회정의 상담. 학지사.

전준 (2024). 질적연구는 인공지능으로 인해 진보할 것인가?: 계산사회과학과 질적연구의 관계에 대한 소고. 한국사회학, 58(4), 123-150.

최가희 (2018). 사회정의와 상담심리의 역할. 한국심리학회지: 상담 및 심리치료, 30(2), 249-271.

황윤아, 김영순 (2025). 인문사회과학 분야의 AI 활용 질적연구에 대한 질적메타분석: 해외연구를 중심으로. 인문사회과학연구, 26(4), 295-332.

Bano, M., Zowghi, D., & Whittle, J. (2023). AI and human reasoning: Qualitative research in the age of large language models. *AI Ethics, 3*(1), 1-15.

Bunt, H. L., Goddard, A., Reader, T. W., & Gillespie, A. (2025). Validating the use of large language models for psychological text classification. Frontiers in Social Psychology, 3, 1460277.

Crocker, A. B., Schmidt, M., Tejeda, J. D., & Rodriguez-Mori, H. (2025). Proof of concept: Leveraging large language models for qualitative analysis of participant feedback. *The Journal of Extension, 63*(1), 16.

Denzin, N. K., & Lincoln, Y. S. (Eds.) (2011). The SAGE *handbook of qualitative research* (4th ed.). SAGE.

Hamilton, L., Elliott, D., Quick, A., Smith, S., & Choplin, V. (2023). Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. *International Journal of Qualitative Methods, 22*, 16094069231201504.

Hill, C., Dahil, A., Simpson, G., Hardisty, D., Keast, J., Pinn, C. K., & Dambha-Miller, H. (2026). Large language models for thematic analysis in healthcare research: A blinded mixed-methods comparison with human analysts. PLOS *Digital Health, 5*(4), e0001189.

Hitch, D. (2024). Artificial intelligence augmented qualitative analysis: The way of the future? *Qualitative Health Research, 34*(7), 595-606.

Kim, Y., Russell, J., Karpinska, M., & Iyyer, M. (2025. 10. 7～10.). *One ruler to measure them all: Benchmarking multilingual long-context language models* [Poster presentation]. Conference on Language Modeling(COLM), Montreal, Canada.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data, *Biometrics, 33*(1) 159-174.

Lewis, J. A., Arnold, M. S., House, R., & Toporek, R. L. (2002). *ACA Advocacy competencies.* Alexandria, VA: American Counseling Association

Li, K. D., Fernandez, A. M., Schwartz, R., Rios, N., Carlisle, M. N., Amend, G. M., Patel, H. V., & Breyer, B. N. (2024). Comparing GPT-4 and human researchers in health care data analysis: Qualitative description study. *Journal of Medical Internet Research, 26*, e56500.

Mazeikiene, N., & Kasperiuniene, J. (2024).

AI-enhanced qualitative research: Insights from Adele Clarke's situational analysis of TED Talks. *The Qualitative Report, 29*(9), 2502-2526.

Nguyen-Trung, K. (2025). ChatGPT in thematic analysis: Can AI become a research assistant in qualitative research? *Quality* & *Quantity, 59,* 4945-4978.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*, 1609406919899220.

OpenAI. (2025. 10. 29). *Introducing gpt-oss-safeguard.* https://openai.com/index/introducing-gpt-oss-safeguard/

Pattyn, F. (2024). The value of generative AI for qualitative research: A pilot study. *Journal of Data Science and Intelligent Systems, 3*(3), 184-191. h

Prescott, M. R., Yeager, S., Ham, L., Rivera Saldana, C. D., Serrano, V., Narez, J., Paltin, D., Delgado, J., Moore, D. J., & Montoya, J. (2024). Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI, 3,* e54482.

Prinzing, M., Bounds, E., Melton, K., Glanzer, P., Fredrickson, B., & Schnitker, S. (2025). Can an algorithm tell how spiritual you are? Using generative pretrained transformers for sophisticated forms of text analysis. *Journal of Personality, 93*(6), 1258-1270.

Qiao, S., Fang, X., Garrett, C., Zhang, R., Li, X., & Kang, Y. (2024). *Generative AI for qualitative analysis in a maternal health study: Coding in-depth interviews using large language models (LLMs).* medRxiv.

Toporek, R. L., & Daniels, J. (2018). *American Counseling Association Advocacy Competencies.* American Counseling Association.

Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., & Chen, K. (2024). ProSA: Assessing and understanding the prompt sensitivity of LLMs. *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 1950-1976). Association for Computational Linguistics.

원 고 접 수 일 : 2026. 03. 16
수정원고접수일 : 2026. 05. 06
게 재 결 정 일 : 2026. 06. 16

한국심리학회지: 상담 및 심리치료
The Korean Journal of Counseling and Psychotherapy
2026, Vol. 38, No. 3, 1123-1144

Comparing Judgment Patterns Between Human Researchers and an LLM in Qualitative Research: Focusing on the Classification of Social Justice Advocacy Competencies

Nara Shin

Research Professor, Institute of Convergence Science, Yonsei University

This study explored judgment patterns between human researchers and a large language model (LLM) when provided with identical theoretical criteria in qualitative counseling research. Using the American Counseling Association (ACA) Social Justice Advocacy Competencies, 32 statements from interviews with university counselors were classified using GPT-SG. The model's classifications were compared with the consensus judgments of three qualitative research experts (reference standard). Overall agreement was 53.1%, with variation in agreement across categories. In mismatched cases, the LLM differed from human researchers in judging intervention levels and applying advocacy concepts, and showed limited consideration of multilayered contextual cues and professional counseling roles. These findings suggest that comparing human and LLM judgments can illuminate the interpretive and contextual nuances inherent in social justice advocacy, and that an LLM may serve as a reflective tool for qualitative researchers to examine conceptual criteria and facilitate analytical reflection.

Key words : Generative Artificial Intelligence, Large Language Model, Qualitative Data Analysis, Social Justice Advocacy