대규모 언어모델 기반 사회정의 옹호 역량 분류: 질적 연구에서의 코딩 결과 일치도 및 오류 특성 분석

(Large Language Model - Based Classification of Social-Justice Advocacy Competencies: Concordance with Human Qualitative Coding)

신나라*, 조한규**

(Nara Shin, Han-Gue Jo)

요 약

본 연구는 질적 연구에서 이론적 준거에 근거해 수행되는 코딩 작업에 대규모 언어모델(LLM)을 적용하고, 사회정의 상담 장면에서 LLM이 산출한 분류가 전문가 합의에 기반한 코딩 결과와 어느 정도로 일치하는지를 검증하였다. 분석 준거로는 미국상담학회(ACA)의 사회정의 옹호 역량 모형을 활용하였으며, 선행연구에서 6개의 유형별 예시로 제시된 사회정의 옹호 상담 경험 진술 32개를 입력 자료로 사용하였다. 네 개의 LLM을 비교한 결과, 균형 정확도는 약 42~53%로 무작위 기준선(16.7%)을 상회하였으나, 유형별 성능 편차와 특정 유형 간 체계적 오분류가 확인되었다. 이러한 결과는 LLM이 질적 코딩에서 일정 수준의 예비 분류를 수행할 수 있으나, 맥락과 판단이 핵심이 되는 분석을 인간 연구자와 동일하게 수행하기에는 한계가 있음을 시사한다. 따라서, 본 연구는 LLM을 질적 연구자의 판단을 대체하는 도구로 이해하기보다, 코딩 과정에서 판단의 성격에 따라 역할이 분화되는 협업적 분석 도구로서의 필요성을 제안한다.

■ 중심어 : 대규모 언어모델 ; 질적 코딩 ; 옹호 역량 ; 인간 - AI 협업

Abstract

This study examined the extent to which large language model (LLM) can support qualitative coding grounded in a theoretically defined framework by comparing LLM performance with expert consensus coding in social justice counseling contexts. The analysis employed the American Counseling Association's (ACA) Advocacy Competencies framework (six domains) as the classification criterion and used 32 social justice counseling excerpts previously established as type-specific exemplars. The results showed overall classification accuracies ranging from 42% to 53% across four different LLMs, exceeding the random baseline (16.7%), but also revealed substantial variability across advocacy types and systematic misclassification patterns. These findings suggest that while LLM demonstrates a meaningful capacity for preliminary classification under clearly defined criteria, they face structural limitations in capturing context-dependent and judgment-intensive aspects of qualitative coding. Rather than positioning LLM as substitutes for human qualitative judgment, this study suggests for a collaborative analytic framework in which human and AI roles are differentiated according to the nature of the coding task.

■ keywords : Large Language Model ; Qualitative coding ; Advocacy competencies ; Human - AI collaboration

Ⅰ. 서 론

대규모 언어모델(Large Language Model, LLM)은 디지털 환경에서 축적되는 방대한 텍스트를 신속하게 처리하고 분류할 수 있다는 점에서 최근 다양한 학문 분야에서 분석 도구로 주목받고 있다[1-3]. 질적 자료 분석 맥락에서 LLM은 전사, 요약과 기능적 자동화를 넘어, 분석 과정을 보조하는 도구로 활용될 가능성이 논의되고 있다[3-5]. 일부 연

* 정회원, 연세대학교 미래융합연구원
** 정회원, 국립군산대학교 인공지능융합학과
이 논문은 2023년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2023S1A5B5A16075034)
접수일자 : 2026년 01월 25일 게재확정일 : 2026년 03월 25일
수정일자 : 1차 2026년 02월 20일, 2차 2026년 03월 20일
교신저자 : 조한규 e-mail : hgjo@kunsan.ac.kr

구에서는 LLM이 반복적 패턴 탐색이나 잠재적 범주 제안을 통해 질적 분석을 보조할 수 있음이 보고되었다[6]. 또한 LLM의 분류 결과를 전문가의 수작업 코딩과 일정 수준에서 일치할 수 있다는 결과를 제시한 비교 연구도 있다[7,8].

질적 연구에서 코딩은 텍스트 자료를 해석하여 의미 단위를 범주화하고 이론적 개념과 연결하는 분석 과정이다[9]. 이는 단순한 분류나 라벨링 절차가 아니라, 이론적 준거와 연구자의 판단이 결합된 분석 행위이다[9-11]. 또한 코딩의 신뢰성을 확보하기 위해서는 전문가 훈련과 반복적인 합의 절차가 필요하며, 자료의 양이 증가할수록 일관성과 재현성을 유지하는 데 상당한 시간과 자원이 요구된다[12]. 이로 인해 명확한 분류 준거가 주어진 코딩 과제에 LLM을 보조 도구로 적용하려는 시도가 확산되고 있으며[2,4], 자동화된 분류가 어떤 조건에서 타당한지에 대한 방법론적 질문도 함께 제기되고 있다[1].

그림 1 ACA 옹호 역량 모형[21]

질적 연구의 코딩 과정과 LLM 적용에 관한 연구는 두 개의 흐름으로 전개되고 있다. 먼저 LLM을 주제 제안이나 초기 코드 생성 등 탐색적 분석을 지원하는 인간 - AI 협업 중심의 접근이다[8, 13-15]. 다음은 사전에 정의된 이론이나 범주 체계를 고정한 상태에서, LLM이 산출한 분류 결과를 인간 코딩과 비교·검증하는 접근이다[7, 16-18]. 후자의 접근은 명확한 분류 기준을 갖춘 이론 모형에 기반한 텍스트 분류 과제를 통해 LLM의 분류 결과를 정량적으로 평가할 수 있다는 장점이 있다. 그러나 개념 경계가 중첩되고 맥락적 판단이 요구되는 질적 코딩 과제에서 적용 가능성은 충분히 검증되지 않았다[16-18]. 따라서 질적 연구 맥락에서 중요한 질문은 LLM이 인간 판단을 어느 정도로 대체할 수 있는가가 아니라, 어떤 과업 조건에서 어떤 종류의 판단을 수행할 수 있는가를 규명하는데 있다[4,6]. 이를 위해서는 명확한 분류 준거를 갖추면서도 해석적 판단을 요구하는 코딩 과제가 필요하다.

사회정의 옹호 상담은 내담자의 문제를 개인 수준에만 한정하지 않고, 제도적·사회문화적 맥락과 연결하여 이해하려는 실천으로[1,20], 이러한 과업 조건을 포함하는 사례 영역이다. 미국상담학회(American Counseling Association, ACA)는 사회정의 옹호 실천을 구조화된 범주 체계로 제시하며, 각 유형에 대해 비교적 명시적인 정의와 판별 기준을 제공한다[25]. 사회정의 옹호 역량 모형(Social Justice Advocacy Counseling Competencies)은 개입의 방향(내담자와 함께하는 옹호 vs. 내담자를 대변하는 옹호)과 개입 수준(개인, 체계·지역사회, 사회·공공정책)에 따라 여섯 가지 유형으로 나타난다(그림 1).

표 1. ACA 옹호 역량 모형 정의[21].

유형	정의
A	상담자가 내담자 자신의 목소리를 찾고 스스로를 옹호할 수 있도록 강점과 자원을 강화하는 개입.
B	내담자의 직접적 참여 없이 상담자가 내담자를 대신하여 문제를 인식하고 변화를 촉진하기 위해 수행하는 옹호 활동.
C	학교 및 지역사회 기관과의 협력을 통해 환경적 요인을 확인하고 자원을 연계하여 내담자와 공동체가 문제 해결 역량을 갖추도록 지원하는 활동을 의미.
D	조직이나 제도 내의 구조적 장벽을 확인하고 변화를 촉진하는 실천
E	상담자와 공동체가 사회적 이슈를 대중에게 알리고 인식을 증진하는 활동.
F	정책 및 입법 수준에서의 변화를 촉진하기 위해 다양한 이해관계자와 협력하는 개입.

특히 ACA 옹호 역량 모형은 명시적 정의와 판별 기준이 제시되어 있다는 점에서(표1) 이론에 근거해 텍스트를 코딩해야 하는 질적 연구의 준거기반 분류 과제로서 LLM 적용 가능성을 검증하기에 적합하다. 다만 일부 유형은 상담자의 역할과 행동 기준이 비교적 명확하게 드러나는 반면(예: A, B 유형), 일부 유형(예: C 유형)은 개입의 방향과 수준이

동시에 고려되어 범주 간 경계가 중첩될 수 있다는 점에서 상대적으로 해석적 판단의 비중이 커진다[20,21]. 이러한 특성은 동일한 준거가 주어지더라도 판단 과정에서 요구되는 해석 수준이 달라질 수 있으며, 이는 LLM이 질적 코딩 과제에서 보이는 과업 민감성을 검증할 수 있는 조건을 제공한다. 따라서 본 연구에서는 ACA 사회정의 옹호 역량을 명확한 범주 정의와 구조를 갖춘 준거 기반 분류 체계로 설정하였다.

본 연구는 선행연구에서 도출된 상담 경험 진술 32개를 입력 자료로 활용하여[21], 준거가 명시된 조건에서 동일한 텍스트에 대해 인간 연구자와 LLM이 어떠한 분류 결과를 산출하는지를 비교하고 분류 정합성과 오류 양상을 분석하고자 한다(그림 2). 구체적인 연구 질문은 다음과 같다.

1. LLM은 ACA 옹호 역량 유형 분류 과제에서 인간 코딩과 얼마나 일치하는 정합성을 보이는가?

2. LLM의 오분류는 특정 옹호 유형 간에 어떠한 경향성이나 패턴을 나타내는가?

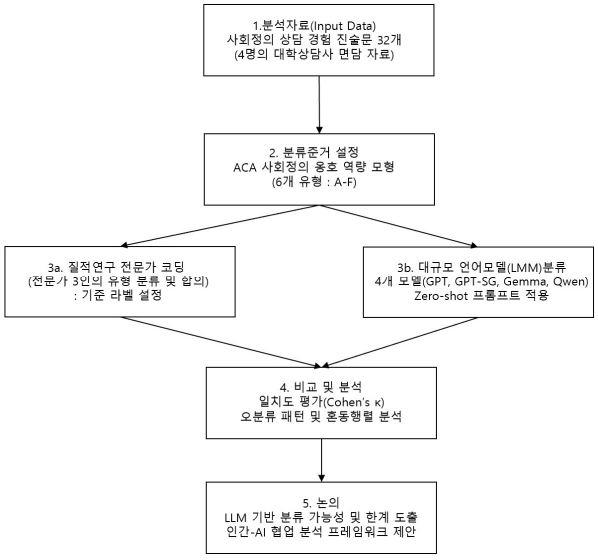

그림 2 연구 방법 개요

Ⅱ. 방 법

1. 분석 자료

본 연구의 분석 자료는 [21]의 연구 결과에 사회정의 옹호 유형별 예시로 제시된 사회정의 상담 경험 진술 32개이다. 해당 진술은 사회정의 상담 경험이 있는 대학 상담사 4명을 대상으로 실시한 1:1 심층 면담 자료를 바탕으로 구성되었다.

분석 자료는 수집된 인터뷰를 축어록으로 작성한 후, 질적 연구 전문가 3인이 반복적으로 검토하여 각 사례의 공통 주제를 도출하는 과정을 통해 구성되었다. 분석자는 질적연구를 수행하는 상담 전문가로 10년 이상의 대학 상담 경력을 보유하고, 한국상담학회 또는 한국상담심리학회 1급 자격을 취득하였으며, 사회정의 상담 관련 교육과 수퍼비전을 10회 이상 이수하였다. 개인 판단의 편향을 최소화하고 분류의 일관성을 확보하기 위해 전문가들은 독립적으로 1차 분류를 진행하였다. 이후 불일치 항목에 대해 토론하였으며, 개입의 목적, 개입 수준(개인 - 조직/지역사회 - 정책), 상담자의 역할 위치(함께/대변)을 고려하여 최종 라벨을 합의하였다. 분석자 간 합의를 통해 자료는 사회정의 옹호 유형에 따라 분류되었고, 그 결과 최종적으로 15개의 공통주제와 48개의 개별 주제가 도출되었다[21].

연구 결과는 각 옹호 유형의 핵심적 특징을 드러내기 위해, 분석 과정에서 도출된 주제와 유형을 대표하는 연구 참여자의 진술을 선별하여 총 32개의 진술문을 ACA 옹호 역량 예시로 제시하였다. 이러한 사례 기반 진술문은 각 유형의 핵심적 특징을 비교적 명확하게 드러내어 LLM 간 분류 결과를 직접적으로 비교하기에 적합하다. 이에 본 연구는 분류 준거와 입력 텍스트 간 대응 관계를 보다 일관되게 검토하기 위해, 32개의 진술문을 LLM의 분류 입력 자료로 활용하였다. 독자의 이해를 돕기 위해, 각 옹호 유형별로 핵심 내용을 담고 있는 진술문 일부를 1개씩 선정하여 표 2에 제시하였다.

2. 분류 준거

본 연구의 분류 준거는 ACA의 사회정의 상담 옹호 역량 모형을 토대로 하였으며[19-21], 여섯 가지 유형(A - F)은 상담자의 역할과 개입 방향에 관한 비교적 명확한 범주와 구조에 따라 구분된다(표1).

본 연구에서는 이 모형을 사회정의 상담의 이론적

의미를 해석하기 위한 틀로 적용하기보다, 질적 연구에서의 코딩 과제를 구성하는 분류 체계로 활용하였다. 본 연구에서 질적 코딩은 프로그래밍 코드 작성이 아니라, 진술문을 범주 체계(A - F)에 따라 분류·라벨링하는 분석 절차를 의미한다. 이에 따라 선행연구에서 3명의 전문가가 합의 과정을 통해 분류한 32개 진술문의 유형을 기준 라벨로 설정하고, 동일 진술문에 대해 LLM이 산출한 분류 결과와의 일치 여부를 비교하였다. 여기서 기준 라벨은 절대적 정답을 의미하기보다, 동일한 분류 과업에서 LLM의 결과를 평가·비교하기 위한 전문가 합의 기반의 준거로 기능한다.

표 2. ACA 옹호 역량 유형 진술문 예시[21].

유형	진술문
A	*"저는 내담자들이 자기가 속한 체계 안에서 자원을 찾는 역할을 중요하게 접근하는 것 같아요. 상담자가 그 사람 인생에 들어가서 뭘 바꿔주는데 한계가 있잖아요. 그러면 예를 들면 가족 안에서 갈등을 크게 겪는 내담자의 경우에는 어떻게 자기가 속한 환경에서 좀 소통해 볼 수 있는지 자원을 찾게 하고…"*
B	*"자살 위험성을 호소하면서 방문하는 내담자들을 즉시 만나서 이야기하는 역할을 했던 것 같고… 가족하고 협력하거나 단과대나 교수님하고도 공유하고 협력하면서 학생들 안전을 보호하는 역할을 했던 것 같아요."*
C	*" 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."*
D	*"저는 외국인 유학생들을 전담하는 상담사가 이제는 있어야 된다고 생각하고 국제교류팀과 상의했어요. 그나마 중국어나 영어로 상담을 할 수 있는 분은 있어도 모집하기에는 처우가 좀 좋지도 않고 예산도 확보해야하고. 보고서도 작성해야 하지만….."*
E	*"단과대별로 심리 상담 특강을 요청하거든요. … 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."*
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."*

3. 대규모 언어모델(LLM)

본 연구에서는 광범위한 언어 과제에서 높은 성능을 보이는 GPT-OSS 20B(GPT)를 주 모델로 사용하였다. 비교 분석을 위해 유사한 규모의 Gemma 3(27B), Qwen 3(30B-A3B-2507), 그리고 사용자 정책 준수를 강화하도록 파인튜닝된 GPT-OSS-Safeguard 20B(GPT-SG)를 함께 포함하였다. 상대적으로 가장 최근에 공개된 GPT는 일반 지식·추론(Measuring Massive Multitask Language Understanding, MMLU), 수학(American Invitational Mathematics Examination, AIME 2025), 과학(graduate-level Google-proof Q&A, GPQA Diamond) 등 zero-shot 벤치마크 환경에서 Gemma3와 Qwen3 대비 전반적으로 우수한 성능 지표를 보이는 것으로 보고되었다(표 3)[22-24]. 모든 모델은 Hugging Face 라이브러리를 통해 공개적으로 제공되는 가중치를 기반으로 LM Studio에서 로컬 환경에 구축하여 추론을 수행하였다. 추론 환경은 모델의 기본 설정을 유지하되, 재현성을 확보하기 위해 생성 설정에서 temperature를 0으로 고정하였다.

표 3. LLM 모델별 주요 성능 지표[22-24]

벤치마크 지표	Gemma3	Qwen3	GPT
MMLU	78.6	81.3	85.3
AIME 2025	24.0	21.6	98.7
GPQA-Diamond	42.4	54.8	71.5

4. 프롬프트 설계 및 분류 절차

본 연구는 시스템 프롬프트를 통해 모델의 즉시 활용 가능한 추론 능력을 활용하였다. ACA의 6가지 유형을 분류하기 위해 사용한 시스템 프롬프트는 다음과 같다.

```
### 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 원문 정의 입력]

### 출력 형식 ###
| 분석 영역 | 분석 근거 |
| (A~F 또는 해당없음) | (상세 설명) |
```

기준의 명확성과 질적 자료의 맥락 보존을 동시에 확보하기 위해, ACA 모형의 영어 원문을 시스템 프롬프트에 사용해 분류 기준의 개념적 일관성을 유지하였으며, 입력 진술문은 번역 과정에서 발생할 수 있는 의미 손실과 편향을 최소화하기 위해 한국어 원문을 그대로 사용하였다[25]. 그 결과 시스템 프롬프트의 길이는 2,591 토큰이었다. 프롬프트가 과도하게 길어질 경우 모델 성능이 저하될 가능성이 있으나, 선행연구에 따르면 구조화된 지시와 적절한 정보량을 포함한 프롬프트는 LLM 성능을 향상시킬 수 있으며, 이러한 효과는 복잡하고 전문성이 요구되는 과제에서 더욱 두드러진다[26, 27].

각 진술문(표 1의 예시를 포함한 총 32개)은 독립된 세션에서 사용자 프롬프트로 실행하였으며, 결과변수는 단일 ACA 유형(A~F) 또는 '해당 없음'과 해당 분류의 근거 설명으로 구성하였다. 전체 실행 결과, '해당 없음'은 GPT-SG 모델에서 1건 산출되었으며, 해당 진술문의 전문가 라벨은 A 유형이었다.

5. 분석 및 평가 방법

전문가 3인이 ACA 옹호 역량 모형의 6가지 유형(A~F)으로 분류한 32개 진술문을 기준 라벨로 설정하고, 동일 진술문에 대해 LLM이 산출한 분류 결과와의 일치 여부를 비교하여 각 LLM의 전체 정확도와 ACA 유형별 평균 정확도를 산출하였다. 또한, 클래스 불균형의 영향을 통제하기 위해 6개 ACA 유형에 대한 재현율의 평균인 균형 정확도(balanced accuracy)를 산출하였다. 6가지 유형을 무작위로 분류할 경우 기대되는 정확도의 기준선(우연 일치)은 1/6(16.7%)이므로, 본 연구의 정확도는 이 기준선과 비교하여 해석하였다. 더불어 클래스 불균형으로 인해 정확도가 과대평가될 가능성을 보완하기 위해, 우연 일치를 보정한 Cohen's κ를 추가로 산출하였다[28]. Cohen's κ는 관찰된 일치율(po = 정확도)과 라벨 분포에 기반한 우연 일치율(pe)을 이용해 $\kappa = (po - pe)/(1 - pe)$로 계산하였으며, 다중 분류(A~F) 과제에서 모델 예측과 기준 라벨 간의 일치도를 평가하는 지표로 활용하였다.

Ⅲ. 결 론

표 4에 제시된 바와 같이 전체 32개 진술문 기준 모델별 균형 정확도는 GPT-SG(53.6%), Qwen3(51.6%), GPT(48.6%), Gemma3(42.1%) 순으로 나타났다. 즉, 네 모델 모두 전체적으로 약 42~53% 수준의 정확도를 보였으며, 이 중 GPT-SG 모델이 가장 높은 전체 정확도를 보였다.

표 4. LLM 모델의 정확도(괄호 안은 진술문 수).

구분		Gemma3	Qwen3	GPT	GPT-SG	LLM 평균
ACA 유형	A(7)	85.7%(6)	42.9%(3)	100.0%(7)	71.4%(5)	75.5%
	B(6)	50.0%(3)	66.7%(4)	83.3%(5)	83.3%(5)	70.8%
	C(6)	33.3%(2)	0.0%(0)	0.0%(0)	16.7%(1)	12.5%
	D(4)	50.0%(2)	100.0%(4)	75.0%(3)	100.0%(4)	81.3%
	E(3)	0.0%(0)	66.7%(2)	33.3%(1)	33.3%(1)	33.3%
	F(6)	33.3%(2)	33.3%(2)	0.0%(0)	16.7%(1)	20.8%
정확도		46.9%(15)	46.9%(15)	50.0%(16)	53.1%(17)	49.2%
균형 정확도		42.1%	51.6%	48.6%	53.6%	

ACA 유형별 결과에서는 유형에 따라 모델 성능의 변동 폭이 크게 나타났다. 일부 유형에서는 특정 LLM이 높은 정확도를 보였으나(Qwen3의 D 유형, GPT의 A 유형, GPT-SG의 D 유형), 다른 유형에서는 동일 LLM의 일치율이 0%에 근접하는 유형도 관찰되어(C, E, F 유형) 모델 간 성능 변동성이 크게 나타났다.

그림 3 GPT-SG의 혼동행렬. 정답 라벨이 A 유형인 진술문 1건은 모델이 '해당사항 없음'으로 분류하여, A-F 범주로 구성된 혼동행렬에는 반영되지 않았다.

특히 C 유형은 다수 모델에서 일치율이 매우 낮아 무작위 분류 기준선(16.7%)을 하회하였다. 이는 해당 유형의 개념적 경계가 텍스트 단서만으로는 충분히 포착되기 어렵거나, 모델이 C 유형을 다른 유형으로 체계적으로 흡수(오분류)하는 경향이 존재할 가능성을 시사한다. 반면 A, B, D 유형은 비교적 모든 모델에서 높은 정확도를 보여, 전문가 판단과 LLM 판단이 공유하는 언어적 단서가 상대적으로 명료할 가능성을 나타낸다.

그림 3는 가장 높은 성능을 보인 GPT-SG 모델(균형 정확도 53.6%)의 혼동행렬을 제시한다. GPT-SG의 Cohen's κ는 0.458로 나타났으며, 이는 클래스 분포에 의해 발생할 수 있는 우연 일치를 보정하더라도 정답 라벨과 중간 수준의 분류 일치도를 보였음을 의미한다. 오분류 양상을 살펴보면, C 및 E 유형이 A 또는 B 유형으로 오분류되는 진술문이 상대적으로 많았고, F 유형이 D 유형으로 오분류되는 경향도 관찰되었다. 이러한 결과는 GPT-SG가 무작위 추정보다 유의미한 분류 신호를 포착하고 있음을 보여주는 한편, 특정 유형 간 경계가 모호한 구간에서 체계적인 혼동이 발생할 수 있음을 보여준다.

Ⅳ. 논 의

본 연구는 ACA의 사회정의 상담 옹호 역량 모형(6유형)을 이론적 준거로 삼아, 사회정의 옹호 상담 경험 진술에 대해 LLM이 산출한 유형 분류가 질적 연구자들의 합의에 기반한 코딩 결과와 어느 정도 일치하는지를 검증하였다. 연구 결과, 모든 모델이 무작위 기준선(16.7%)을 상회하는 균형 정확도를 보였으며, 이는 사전에 정의된 이론적 준거가 명시적으로 제시된 조건에서 LLM이 일정 수준의 분류 수행 능력을 보일 수 있음을 시사한다. 다만 전체 정확도가 절반 내외에 머물렀고, 유형별 성능 편차 및 특정 유형 간 체계적 오분류가 확인되었다. 특히 최고 성능을 보인 GPT-SG의 경우에도 Cohen's κ가 0.458로 나타나, 우연 일치를 보정하더라도 분류 일치도가 중간 수준으로 나타났다. 이는 선행 연구에서도 인간 합의 결과와 LLM 분류 간 합의도가 Cohen's κ≈.39 - .40 수준으로 보고된 결과와도 유사한 범위이다[29-30]. 따라서 현 단계에서 LLM의 분류 결과를 질적 연구에서의 전문가 코딩과 동일한 수준의 판단으로 간주하기에는 한계가 존재하며, 인간과 LLM이 서로 다른 판단 단서를 선택하는 과정적 차이를 반영한 결과로 해석할 수 있다.

GPT-SG 모델은 일반 모델 대비 안전, 정책 준수 및 지시 이행을 강화하는 방향으로 설계된 변형 모델이며[31], 이러한 설계가 프롬프트에서 요구한 절차(단일 분류, 예외 처리, 준거 정의)의 준수를 높여 상대적으로 안정적인 분류를 보였을 가능성이 있다. 즉, 지시 이행 능력이 강화된 GPT-SG 모델이 프롬프트에 제시된 6가지 유형의 세부 기준을 충실히 반영했을 것으로 해석된다. 실제로 일반 GPT 모델은 C, F 유형에서 0%의 일치율을 기록했으며, GPT-SG는 비록 낮은 수준이지만(16.7%) 해당 유형을 부분적으로 선별하여, 최소한의 구분 특성을 포착했음을 시사한다.

유형별 결과에서 가장 두드러진 특징은 A, B, D 유형에서의 상대적으로 높은 일치율과 C 유형에서의 일관된 저성능 및 체계적 오분류이다. 이는 동일한 분류 준거가 주어졌더라도 질적 코딩 과제에서 요구되는 판단의 성격이 유형별로 상이함을 의미한다.

혼동 행렬 분석 결과, C 유형(지역사회 협력)에 해당하는 진술문은 주로 B 유형(내담자/학생 옹호) 또는 D 유형(체계 옹호)으로 분류되는 경향을 보였다. 이는 LLM이 "기관 연계", "의뢰", "협력"과 같은 표면적 언어 단서를 포착하더라도, 해당 개입이 내담자와 함께 수행된 협력인지, 내담자를 대변하는 옹호인지, 혹은 체계 변화를 지향하는 실천인지를 구분하는 데 필요한 관계적 맥락을 충분히 반영하지 못했음을 시사한다.

이와 유사하게 E (공공정보 제공)와 F(사회·정치적 옹호) 유형에서도 낮은 일치율이 관찰되었는데, 이는 해당 유형들이 단일 행동 단서보다는 사회문화적 맥락, 공적 담론의 대상, 그리고 다층적인 관계

맥락을 전제로 판단되어야 하기 때문으로 해석된다[21]. 이러한 경향은 한국어 질적 코딩에서 LLM이 명시적 개념이나 행동 분류에는 비교적 안정적인 성능을 보이나, 추상적 의미 통합이나 거시적 맥락 해석이 요구되는 경우 성능이 저하된다는 선행연구와 일치한다[16,17].

이러한 결과는 LLM이 이론적 정의를 고려하는 과정에서의 오류라기보다, 분류 판단에 활용하는 단서의 우선순위가 인간 연구자와 다르게 구성되어 있음을 보여준다[6]. 인간 연구자는 개입의 주체, 협력의 상호성, 그리고 개입이 지향하는 변화의 수준을 이론적 준거에 비추어 종합적으로 조정하는 반면, LLM은 텍스트에 명시적으로 드러난 행동 중심의 단서를 기반으로 판단했을 가능성이 있다[8,12]. 따라서 C 유형뿐 아니라 E, F 유형에서도, 사회문화적·거시적 맥락 및 관계적 경계를 해석적으로 구성해야 하는 경우 판단의 초점이 특정 수준으로 이동하며 다른 유형으로 재귀속되었을 것으로 예상된다.

종합하면, 이러한 오분류는 단순한 오류나 분석 실패로 보기 보다는, 질적 코딩 과정에서 요구되는 판단의 층위 차이를 드러내는 지점으로 이해될 수 있다. 이는 선행연구에서 논의된 바와 같이 LLM이 기능적 자동화를 넘어 분류·해석을 보조할 수 있다는 결과와 맥락을 같이한다[5,15]. 다만, LLM은 명확한 행동 기준과 역할 정의가 제시된 분류 과제에서는 예비 코딩 또는 초기 분류 단계에서 효율성을 제공할 수 있으나, 관계적 맥락과 개입의 의미를 종합적으로 조정해야 하는 경우에는 인간 연구자의 해석과 합의가 필수적이다[1,13]. 이는 질적 코딩에서의 판단이 특정 주체의 고정된 속성이라기보다, 입력 텍스트와 이론적 준거, 분석 도구의 상호작용 속에서 구성되는 과정임을 보여준다[3]. 이러한 관점에서 LLM은 질적 연구자의 판단을 대체하는 도구라기보다, 질적 코딩 과정에서 특정 유형의 판단을 수행하는 기술적 보조 행위자로 이해될 필요가 있다[4,10,14].

본 연구 결과를 해석할 때 몇 가지 고려사항이 있다. 본 연구는 로컬 환경에서 동일한 프롬프트를 모든 모델에 일괄 적용함으로써 비교의 공정성과 재현성을 확보했다는 강점을 지닌다. 그러나 이러한 설계는 모델별로 최적화된 프롬프트 구성(예: 정의 길이, 제시 방식, 예시 제공 여부, 단계적 질문 구조)의 효과까지 함께 비교하기에는 한계가 있다. 따라서 후속 연구에서는 프롬프트 입력 조건 변화에 따른 성능 변화 및 본 연구에서 확인된 유형별 오분류 양상에 관한 정밀한 검증이 필요하다. 또한 본 연구는 예시를 제공하지 않는 zero-shot 프롬프트 기반으로 분류를 수행하였으며, 적절한 예시를 포함하는 few-shot 설정을 적용할 경우 일치도가 향상될 가능성이 있다. 다만 few-shot 예시를 선정하는 과정에서 연구자의 주관적 판단이 개입되어 편향이 발생할 우려가 있으므로, 유형별 대표성, 난이도, 길이 등 예시 선정 기준을 객관화하고 그 절차를 명확히 제시할 필요가 있다.

그리고 C, E, F 유형과 같이 개념 경계가 중첩되거나 맥락 의존성이 큰 범주에 대해서는 판단 기준을 단계적 의사결정 구조로 명시화할 필요가 있다(예: 개입 주체 → 관계/협력 방식 → 변화의 수준 또는 개입 목표). 이러한 구조를 프롬프트 설계에 반영했을 때 분류 정확도와 Cohen's κ, 그리고 오분류 패턴(예: C, E→A, B, F→D)이 어떻게 변화하는지를 체계적으로 분석하는 것이 요구된다[26-27]. 마지막으로, 입력 데이터의 단위와 정보 밀도가 분류 성능에 미치는 영향을 검증하기 위해, 문장 단위 진술과 확장된 맥락 서술(사례 배경, 상호작용, 개입 과정 등)을 비교가 요구된다[9,13,18]. 이를 통해 맥락 정보의 증가가 LLM의 관계적, 의도적 판단을 통합하는 과정에 어떤 영향을 미치는지, 그리고 모호한 경계 범주에서 오분류를 감소시키는 데 기여하는지를 평가할 수 있다. 특히 본 연구가 영문 분류 준거와 한글 진술문을 동시에 다루는 만큼, 한국어 이해에 강점을 갖는 국내 LLM(한국어 특화 상용·공개 모델 등)을 함께 포함하여 교차 검증하거나, 국내 LLM과 글로벌 LLM의 합의/불일치 사례를 분석하는 방식이 일치도 향상 및 오류 유형 규명에 도움이 될 수 있다.

V. 결 론

본 연구는 LLM이 질적 연구 코딩을 자동화할 수 있는지를 단정적으로 판단하기보다, 이론적 준거가 명시된 조건에서 인간 연구자와 LLM의 판단 차이를 실증적으로 확인하고자 하였다. 연구 결과는 LLM이 특정 조건에서는 분류 보조 도구로 활용될 가능성을 보이지만, 관계적 맥락 판단에서는 인간 연구자의 해석이 여전히 중요함을 시사한다. 이러한 결과는 AI와 인간 연구자가 수행하는 협업적 코딩 체계를 설계하는 데 필요한 경험적, 이론적 초석으로 기능할 것으로 기대된다.

REFERENCES

[1] 강수정, “상담 및 심리치료에서 대규모 언어 모델(LLM)의 적용 기회와 한계 및 윤리적 고려 사항”, *디지털콘텐츠학회논문지*, 제25권, 제12호, 3751-3759쪽, 2024년

[2] 이상혁, 김은미, “대규모 언어 모델은 분석 도구가 될 수 있는가? GPT를 활용한 내용 분석의 신뢰도와 타당도를 중심으로”, *한국언론학보*, 제69권, 제1호, 5-38쪽, 2025년

[3] 황윤아, 김영순, “인문사회과학 분야의 AI 활용 질적 연구에 대한 질적메타분석: 해외연구를 중심으로”, *인문사회과학연구*, 제26권, 제4호, 295-332쪽, 2025년

[4] 전준, “질적연구는 인공지능으로 인해 진보할 것인가?: 계산사회과학과 질적연구의 관계에 대한 소고”, *한국사회학*, 제58권, 제4호, 123-150쪽, 2024년

[5] Hamilton, L., Elliott, D., Quick, A., Smith, S., & Choplin, V, “Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data”, *International journal of qualitative methods*, 22, 16094069231201504, 2023.

[6] Hitch, D., “Artificial intelligence augmented qualitative analysis: the way of the future?”, *Qualitative Health Research*, vol. 34, no. 7, pp. 595-606, 2024.

[7] Jeong, Y., Smith, M., Gallo, R. J., Knowlton, L. M., Lin, S., & Shieh, L., “Leveraging ChatGPT for thematic analysis of medical best practice advisory data”, *JAMIA open*, vol. 8, no. 5, ooaf126, 2025.

[8] Prescott, M. R., Yeager, S., Ham, L., Rivera Saldana, C. D., Serrano, V., Narez, J., et al., “Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses”, *JMIR AI*, 3, Article e54482, 2024.

[9] 김영순, 황윤아, 윤혜림, 우지연, “질적연구물 연구방법 영역의 인간-AI 협업 과정에 관한 문헌사례연구”, *다문화와 교육*, 제10권, 제4호, 1-25쪽, 2025년

[10] 김지연, “국내 상담학 분야의 질적연구 동향 분석”, *질적탐구*, 제4권, 제2호, 131-168쪽, 2018년

[11] Hill, C. E., Knox, S., Thompson, B. J., Williams, E. N., Hess, S. A., & Ladany, N., “Consensual qualitative research: an update. *Journal of couseling psychology*”, vol. 52, no. 2, 196, 2005.

[12] O’Connor, C., & Joffe, H., “Intercoder reliability in qualitative research: Debates and practical guidelines”, *International journal of qualitative methods*, 19, 1609406919899220, 2020.

[13] 박종도, 김영순, 최은하, 황윤아, “질적연구에서 자료처리 과정의 인간-AI 협업 경험에 관한 질적 사례연구”, *다문화와 교육*, 제10권, 제4호, 27-51쪽, 2025년

[14] 정여주, 이명준, 강슬기, 김효정, “인공지능(AI) 상담 연구 동향과 열린 교육에서의 실천적 과제”, *열린교육연구*, 제33권, 제2호, 27-48쪽, 2025년

[15] Nguyen-Trung, K., “ChatGPT in Thematic Analysis: Can AI become a research assistant in qualitative research?”, *Quality & Quantity*, pp. 1-34, 2025.

[16] 김윤하, “대규모 언어 모델은 한국어 질적 코딩을 수행할 수 있는가? - GPT-4 기반 귀납적 코딩 자동화 가능성 평가”, *인공지능인문학연구*, 21, 107-135쪽, 2025년

[17] Pattyn, F., “The value of generative AI for qualitative research: A pilot study”, *Journal of Data Science and Intelligent Systems*, vol. 3, np. 3, pp. 184 - 191, 2024.

[18] Qiao, S., Fang, X., Garrett, C., Zhang, R., Li, X., & Kang, Y., “Generative AI for qualitative analysis in a maternal health study: Coding in-depth interviews using large language models (LLMs)”, *medRxiv*, https://doi.org/10.1101/2024.09.16.24313707, 2024.

[19] “ACA Advocacy Competencies”, http://www.counseling.org/Publications/, accessed Feb. 3, 2009.

[20] Toporek, R. L., Lewis, J. A., & Crethar, H. C., “Promoting systemic change through the ACA advocacy competencies”, *Journal of Counseling & Development*, vol. 87, no. 3, pp. 260-268, 2009.

[21] 신나라, “대학상담사의 사회정의 옹호 상담 경험 연구”, *학습자중심교과교육연구*, 제25권, 제1호, 761-782쪽, 2025년

[22] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al., “Qwen3 technical report”, arXiv, arXiv:2505.09388, 2025.

[23] Kamath, G.T., et al., “Gemma 3 Technical Report”, arXiv, abs/2503.19786, 2025.

[24] “OpenAI. Introducing gpt-oss”, https://openai.com/index/introducing-gpt-oss/,

(accessed Jan., 5. 2026).
[25] “American Counseling Association Advocacy Competencies”, https://www.counseling.org/, (accessed Jan., 5. 2026).
[26] Liu, Y. Y., Zheng, Z., Zhang, F., Feng, J. C., Fu, Y. Y., Zhai, J. D., et al., “A comprehensive taxonomy of prompt engineering techniques for large language models”, *Frontiers of Computer Science*, vol. 20, no. 3, 2003601, 2026.
[27] Do, X. L., Dinh, D., Nguyen, N. H., Kawaguchi, K., Chen, N., Joty, S., & Kan, M. Y., “What makes a good natural language prompt?”, *In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 1, pp. 5835-5873, 2025.
[28] Landis, J. R., & Koch, G. G., “The measurement of observer agreement for categorical data”, *Biometrics*, pp. 159-174, 1977.
[29] Crocker, A. B., Schmidt, M., Tejeda, J. D., & Rodriguez-Mori, H., “Proof of concept: Leveraging large language models for qualitative analysis of participant feedback”, *The Journal of Extension*, vol. 63, no. 1, 16, 2025.
[30] Li, K. D., Fernandez, A. M., Schwartz, R., Rios, N., Carlisle, M. N., Amend, G. M., et al., “Comparing GPT-4 and human researchers in health care data analysis: Qualitative description study. *Journal of Medical Internet Research*, 26, e56500, 2024.
[31] “OpenAI. Introducing gpt-oss-safeguard”, https://openai.com/index/introducing-gpt-oss-safeguard/, (accessed Jan., 5. 2026).

저 자 소 개

신나라(정회원)

2010년 연세대학교 신학/신문방송학 학사 졸업
2012년 연세대학교 상담학 석사 졸업
2018년 연세대학교 상담학 박사 졸업

<주관심분야 : 사회정의상담, AI기반 심리상담, 디지털 헬스, 질적연구 등>

조한규(정회원)

2007년 국립공주대학교 컴퓨터정보통신공학부 학사 졸업.
2009년 연세대학교 의료정보 석사 졸업.
2014년 비아드리나대학교 뇌인지과학 박사 졸업.

<주관심분야 : 인공지능, 디지털 헬스, 뇌인지과학 등>