Smart Media Journal / Vol.15, No.2 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2026.15.2.20

대규모 언어모델을 활용한 질적 데이터의 양적 지표화 : 물리치료 융합을 위한 새로운 방법론

(Quantification of Qualitative Data Using Large Language Models: A Novel Methodology for Physical Therapy Integration)

유성훈*, 이설의***, 윤주상****

(Seong-hun Yu, Seol-eui Lee, Joo-Sang Yun)

요 약

질적 분석은 물리치료에서 환자의 주관적 경험, 치료에 대한 인식, 회복 과정에 대한 심층적 이해를 제공하여 중재 개발 및 치료 결과 향상에 필수적인 과정이다. 그러나 물리치료 환자 인터뷰에 기반한 전통적인 질적 분석은 효율 및 신뢰성 측면에서 문제에 직면하고 있다. 본 연구에서는 이러한 한계를 개선하기 위해 물리치료 연구의 질적 분석에 대형 언어 모델(Large Language Models, LLMs)을 통합하는 방안을 제안하고 있다. 뇌졸중 환자 20명을 대상으로 반구조화된 인터뷰를 실시하였으며, 체계적인 질적 코딩을 위한 정량적 지표인 대형 언어 모델 지수(LLMq, Large Language Model Quotient) 방법론을 적용하여 물리치료 맥락에서의 타당성과 효과성을 검토하였다. 뇌졸중 환자 20명의 반구조화 인터뷰를 시범 사례로 활용하여, ChatGPT, Claude, Gemini 세 LLM에 각 160회 반복 측정을 수행하고 대형 언어 모델 지수(LLMq) 방법론을 적용하여 물리치료 맥락에서의 타당성을 검토하였다. 급내상관계수(ICC) 분석 결과, ICC(3,1)은 0.802, ICC(3,k)는 0.924로 Cicchetti(1994)가 제시한 우수함의 기준에 부합하는 결과를 보였다. 결과적으로 새로운 방법론이 질적분석 효율성 향상 및 주관적 편향을 감소를 촉진한다는 것을 확인하였으며, LLM이 물리치료 연구에 유용한 가치를 제공한다는 것을 시사한다.

■ 중심어 : 질적 분석 ; 물리치료 ; 대형 언어 모델(LLM) ; 뇌졸중 재활 ; 혼합 연구

Abstract

Qualitative analysis constitutes an essential methodological approach in physical therapy research, offering comprehensive insights into patients' subjective experiences, perceptions of therapeutic interventions, and recovery trajectories, thereby facilitating evidence-based intervention development and optimizing clinical outcomes. Nevertheless, conventional qualitative analytical approaches applied to physical therapy patient interviews encounter substantial limitations regarding analytical efficiency and inter-rater reliability. The present study examines the integration of Large Language Models (LLMs) into qualitative research methodologies within the physical therapy domain to address these methodological constraints. Semi-structured interviews were conducted with one stroke patient as a pilot case, and three LLMs (ChatGPT, Claude, and Gemini) were employed with 160 iterative measurements each to implement the Large Language Model Quotient (LLMq) methodology—a quantitative framework for systematic qualitative coding—to evaluate its validity and applicability within the physical therapy research context. Intraclass correlation coefficient (ICC) analysis demonstrated ICC(3,1) of 0.802 and ICC(3,k) of 0.924, meeting Cicchetti's (1994) criteria for excellent reliability. Findings demonstrate that this analytical approach significantly enhances the efficiency of qualitative data analysis (reducing analysis time from weeks-months to hours-days) while mitigating subjective interpretive bias (ICC 0.802), thereby substantiating the considerable methodological value that LLM-based approaches confer upon physical therapy research endeavors.

■ keywords : Qualitative analysis ; Physical theerapy ; Large Language Models(LLM) ; Stroke rehabilitation ; Mixed methods research

Ⅰ. 서 론

물리치료 연구에서 질적 분석은 환자의 주관적 경험, 치료에 대한 인식, 회복 과정에 대한 심층

* 종신회원, 남부대학교 물리치료학과 교수
** 정회원 광주과학기술원 전기전자컴퓨터공학부
*** 정회원 남부대학교 물리치료학과 석사과정
본 과제(결과물)는 2025년도 교육부 및 광주광역시의 재원으로 광주RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다(2025-RISE-05-남부대-007)
접수일자 : 2025년 10월 19일
수정일자 : 2025년 12월 01일
게재확정일 : 2025년 12월 10일
교신저자 : 유성훈 e-mail : youseonghun@gmail.com

적 이해를 제공함으로써 환자 중심의 중재 개발과 치료 결과 향상에 필수적인 역할을 수행한다[1]. 특히 재활 과정에서 환자가 경험하는 복잡하고 다차원적인 감정, 동기, 그리고 적응 과정을 이해하는 데 있어 질적 연구 방법론의 중요성이 점점 더 인정받고 있다[2,3].

전통적으로 물리치료 연구에서 널리 사용되는 설문지는 구조화된 응답을 통해 양적 분석에 유용한 도구이지만, 환자의 재활 경험과 같이 개인별 맥락과 주관성이 크게 작용하는 복합적 주제를 충분히 탐색하기에는 본질적인 한계가 존재한다[4]. 이러한 한계를 극복하기 위해 반구조화 인터뷰가 주목받고 있는데, 이는 핵심 질문 틀을 유지하면서도 개별 환자의 서사적 경험을 심층적으로 탐구할 수 있어 물리치료 연구에서 보다 풍부하고 의미 있는 질적 자료를 확보하는 데 적합한 방법론으로 평가되고 있다[5,6].

하지만 물리치료 현장에서 환자 인터뷰에 대한 전통적인 질적 분석은 효율성과 신뢰성 측면에서 문제를 직면하고 있다. 첫째, 임상 업무와 연구를 병행해야 하는 물리치료사들에게 질적 분석의 노동 집약적 특성은 상당한 시간적 부담을 가하고 있다. 질적 연구는 전사, 데이터 정리, 소프트웨어 작업, 팀 협업 및 데이터 심층적 해석이 필요하며 특히 참여자와 연구자는 인터뷰 주제가 민감할 경우 별도의 신뢰 구축을 위한 시간도 필요하기 때문에 결과적으로 질적 연구자는 극심한 시간 압박을 받게 된다[7]. 둘째, 물리치료사 개인의 임상 경험, 전문적 배경, 그리고 관점의 차이에서 비롯되는 주관적 해석의 변동성으로 인해 분석 결과의 일관성과 객관성을 확보하는 것이 어려우며[8], 동일한 환자 인터뷰 자료라도 분석자에 따라 상이한 해석과 결론이 도출될 수 있어 표준화된 분석 프로세스 구축에 근본적인 한계가 있다[9].

이러한 방법론적 한계를 해결하기 위한 혁신적 접근법으로, 본 연구는 물리치료 연구의 질적 분석 워크플로우에 대형 언어 모델(Large Language Models, LLMs)을 통합하는 새로운 방안을 제안한다. LLM이 전통적 질적 분석의 한계를 극복할 수 있는 이론적 근거는 다음과 같다. 첫째, 분석 효율성 측면에서 LLM은 사전 학습된 방대한 언어 패턴과 의미 구조를 기반으로 텍스트를 즉각적으로 처리할 수 있으며, 명확한 코드북이 제공될 경우 인터뷰 전사 자료를 수 분 내에 분석하여 수주에서 수개월이 소요되는 전통적 과정을 획기적으로 단축할 수 있다[10]. 둘째, 신뢰성과 일관성 측면에서 LLM은 동일한 프롬프트와 코드북 기반의 반복 실행을 통해 통계적으로 안정적이고 재현 가능한 결과를 생성하며, 특히 LLMq(Large Language Model Quotient) 방법론은 다수의 반복 분석을 통해 코드 존재 여부를 확률적으로 정량화함으로써 연구자 간 주관적 해석의 변동성 문제를 완화할 수 있다[10]. 셋째, 구체적 작동 원리 측면에서 LLM은 연구자가 작성한 코드북을 구조화된 프롬프트로 입력받아, 사전 학습된 언어 이해 능력을 활용하여 단순한 키워드 매칭을 넘어서 의미론적 유사성과 맥락적 관련성을 파악하고, 다수의 반복 실행을 통해 코드 식별의 안정성을 통계적으로 검증한다[11].

최근의 방법론적 발전은 이러한 접근법의 타당성과 신뢰성을 실증적으로 입증하고 있다. Tai et al. (2024)의 연구는 LLMq 접근법을 통해 LLM 지원 질적 코딩이 다중 반복 분석을 통해 체계적이고 일관된 코드 식별을 제공할 수 있음을 보여주었으며, Gilardi 등(2023)의 연구는 ChatGPT가 텍스트 주석 작업에서 인간 크라우드 워커 대비 84% 이상의 정확도를 달성하고 질적 데이터 처리 과정에서 탁월한 일관성을 구현할 수 있음을 실증하였다. 이러한 증거들은 LLM이 단순히 분석 속도를 높이는 도구를 넘어서, 질적 연구의 신뢰성과 재현성을 강화할 수 있는 방법론적 혁신임을 시사한다[10,11].

이러한 배경에서 본 연구는 뇌졸중 환자 20명을 대상으로 반구조화된 인터뷰를 실시하고, LLM 기반 질적 분석 방법론을 적용하였다. 이를 통해 물리

치료 연구에서 LLM 기반 질적 데이터 양적 지표화 방법론의 타당성과 효과성을 검토하고자 한다.

Ⅱ. 본 론

1. 이론적 배경

가. 질적분석 LLMq 방법론

Tai et al.(2024)은 대형 언어 모델을 질적 분석에 활용하는 체계적인 방법론을 제시하였으며, 특히 연역적 코딩에서 LLM의 타당성을 입증하였다. LLMq의 수식은 다음과 같다.

$$LLMq = \frac{n}{N}$$

n은 긍정 응답 수(Postivie Response)이고, N은 총 반복횟수(Total Iterations)이다. 이들의 연구에서는 동일한 텍스트를 160회 반복 분석하여 각 반복을 새로운 코더가 분석하는 것과 동등한 과정으로 간주하였고, 이를 통해 LLMq라는 신뢰도 측정 지표를 도출하였다. 연구 결과, LLM 분석은 전통적인 인간 코더와 일관된 결과를 보였으며, 특히 40회 반복 이후 안정적인 패턴을 나타내는 것으로 확인되었다[10].

나. LLM 성능 검증

Gilardi et al.(2023)의 연구는 LLM이 텍스트 주석 작업에서 인간 크라우드워커를 능가하는 성능을 보인다는 실증적 증거를 제공하였다. 이들은 ChatGPT 3.5를 사용하여 트위터 게시물의 분류 작업을 수행한 결과, LLM이 인간보다 높은 정확성(>84%)과 일관성을 보였으며, 특히 무작위성 설정(temperature 0.2-1.0) 범위에서도 신뢰할 수 있는 반복 분류 성능을 달성하였다고 보고하였다. 이러한 결과는 LLM이 단순한 보조 도구를 넘어서 독립적인 분석 역량을 갖춘 도구로 활용될 수 있음을 시사한다[11].

다. 물리치료 연구의 LLM 활용 필요성

Naqvi et al.(2024)은 물리치료 분야에서 대형 언어 모델의 도입이 교육, 실습, 연구 전반에 걸쳐 패러다임 전환을 가져올 것이라고 전망하였다. 특히 물리치료사들이 행정 업무 간소화, 전세계적 연결, 그리고 맞춤형 치료 개발에 LLM을 활용함으로써 큰 이점을 얻을 수 있다고 주장한다. 그러나 인간의 손길과 창의성은 여전히 대체할 수 없는 가치를 지닌다는 점을 짚어, LLM의 도구적 특성을 강조하였다[20].

라. 본 연구의 이론적 근거

이러한 선행연구들의 결과는 물리치료 연구에서 LLM 지원 질적 분석의 잠재력을 뒷받침하는 강력한 이론적 근거를 제공한다. 특히 Tai et al.(2024)의 LLMq 방법론과 Gilardi et al.(2023)의 성능 우월성 입증은 본 연구가 추구하는 효율성 향상과 편향 감소 목표의 실현 가능성을 시사한다. 동시에 여러 연구들이 지적한 감정적 뉘앙스와 맥락적 해석의 한계는 본 연구에서 LLM과 인간 분석자 간의 상호 보완적 접근법의 필요성을 강조한다[10,11].

따라서 본 연구는 이러한 선행 연구들의 성과와 한계를 종합하여, 뇌졸중 환자의 재활 경험이라는 특수한 맥락에서 LLM 지원 주제분석의 타당성과 효과성을 체계적으로 검증하고자 한다.

2. 연구 방법

가. 연구 모형

본 연구는 LLM을 활용한 질적 데이터의 양적 지표화 타당성을 검증하기 위해 단일사례 반복측정 설계를 사용하였다. 연구대상은 뇌졸중 환자 20명을 대상으로 실시한 반구조화 인터뷰 데이터이며, 보행/균형, 사회적 참여, 감정/심리, 미래 기대의 4개 영역에서 긍정 및 부정 감정을 탐색하였다.

독립변수로는 3종의 LLM을 활용한 분석 방법을 설정하였으며, 매개변수는 Tai et al.(2024)의

LLMq 방법론을 적용하여 도출하였다. 각 LLM은 동일한 인터뷰 데이터에 대해 160회 반복측정을 수행하였으며, 이를 통해 각 감정 영역별 LLMq 값을 산출하였다. 조절변수로는 선행연구에 기반한 코드북 구조를 적용하여 분석의 일관

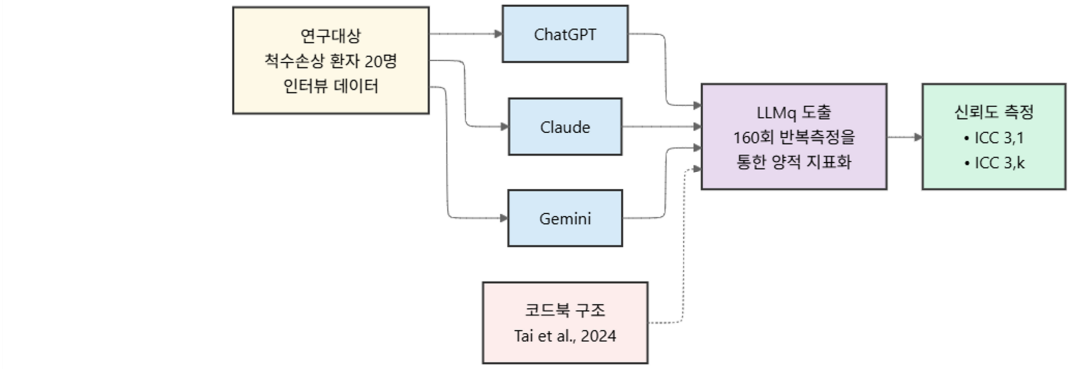


그림 1 LLM 기반 물리치료 질적 데이터 양적 지표화 검증 연구 모형. 본 연구는 뇌졸중 환자 20명의 반구조화 인터뷰 데이터를 3개의 LLM(ChatGPT, Claude, Gemini)에 입력하여 각 모델당 160회 반복 분석을 수행하였다. 각 반복 분석을 통해 도출된 LLMq(Large Language Model Quotient) 값의 신뢰도를 급내상관계수(ICC)로 검증하였으며, 이는 Tai et al.(2024)의 코드북 구조를 기반으로 설계되었다. 본 연구 모형은 질적 데이터의 양적 지표화 과정에서 LLM의 일관성과 재현성을 통계적으로 검증하기 위한 방법론적 프레임워크를 제시한다.

성을 확보하였다[10].

종속변수는 3종 LLM 간 측정값의 신뢰도로 설정하였으며, ICC(3,1)을 통해 단일 측정값의 일치도를, ICC(3,k)를 통해 평균 측정값의 신뢰도를 평가하였다. ICC 값이 0.75 이상일 경우 높은 신뢰도로 해석하였다[16]. 모든 통계분석은 Python 3.12.7 환경에서 pingouin, pandas 패키지를 사용하여 수행하였다.

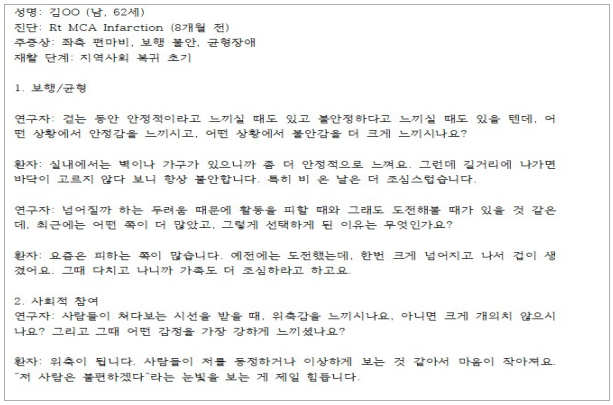
본 연구의 가설은 "3종 LLM을 통해 도출된 LLMq는 우수한(Excellent) 급내상관계수(ICC ≥ 0.75)를 나타낼 것이며, 이는 질적 데이터의 양적 지표화에 LLM 활용이 타당함을 입증할 것이다"로 설정하였다.

나. 연구 대상 및 자료

본 연구는 G시 소재 S병원에서 신경계 물리치료를 받고 있는 환자 중 연구 목적을 충분히 설명 듣고 자발적으로 참여에 동의한 20명을 대상으로 진행되었다. 질적 분석은 전통적인 질적 연구의 기본 절차를 준용하되, 대형 언어 모델 기반 질적 분석 지수(LLMq)를 활용하여 인터뷰 전사 자료를 체계적으로 구조화하였다. 연구 대상 20명의 인터뷰(1인당 약 30분)는 총 약 181,200자(A4 약 92쪽 분량)로 전사되었으며, 네 가지 핵심 영역(보행/균형, 사회적 참여, 감정·심리, 미래 기대)에 대해 다음과 같은 분량의 원자료를 작성하였다:

보행/균형: 약 38,100자, 사회적 참여: 약 33,500자
감정·심리: 약 54,200자, 미래 기대: 약 31,400자

전사 자료는 LLM 기반 자동 분석 후 연구자 검토 과정을 거쳐 의미 구조를 정제하였다.으며, 이를 통해 분석의 신뢰성과 타당성을 확보하였다.


성명: 김OO (남, 62세)
진단: Rt MCA Infarction (8개월 전)
주증상: 좌측 편마비, 보행 불안, 균형장애
재활 단계: 지역사회 복귀 초기

1. 보행/균형

연구자: 걷는 동안 안정적이라고 느끼실 때도 있고 불안정하다고 느끼실 때도 있을 텐데, 어떤 상황에서 안정감을 느끼시고, 어떤 상황에서 불안감을 더 크게 느끼시나요?

환자: 실내에서는 벽이나 가구가 있으니까 좀 더 안정적으로 느껴요. 그런데 길거리에 나가면 바닥이 고르지 않다 보니 항상 불안합니다. 특히 비 온 날은 더 조심스럽습니다.

연구자: 넘어질까 하는 두려움 때문에 활동을 피할 때와 그래도 도전해볼 때가 있을 것 같은데, 최근에는 어떤 쪽이 더 많았고, 그렇게 선택하게 된 이유는 무엇인가요?

환자: 요즘은 피하는 쪽이 많습니다. 예전에는 도전했는데, 한번 크게 넘어지고 나서 겁이 생겼어요. 그때 다치고 나니까 가족도 더 조심하라고 하고요.

2. 사회적 참여
연구자: 사람들이 쳐다보는 시선을 받을 때, 위축감을 느끼시나요, 아니면 크게 개의치 않으시나요? 그리고 그때 어떤 감정을 가장 강하게 느끼셨나요?

환자: 위축이 됩니다. 사람들이 저를 동정하거나 이상하게 보는 것 같아서 마음이 작아져요. "저 사람은 불편하겠다"라는 눈빛을 보는 게 제일 힘듭니다.

그림 2 인터뷰 설문지 질문 및 답변 예시

다. 분석 모델 및 평가 지표

(1) 분석 모델 선택

본 연구는 LLM을 활용한 질적 데이터의 정량화 가능성을 탐색하고자, 상용화된 LLM 중 성능과 접근성이 뛰어난 3종의 모델을 분석 도구로 선정하였다. 선정된 모델은 OpenAI사의 ChatGPT 3.5-turbo, Anthropic사의 Claude 3.5-haiku, Google DeepMind의 Gemini 2.5-flash-lite 총 3종의 LLM 모델이다. 세 개의 모델은 Transformer 알고리즘을 기반으로 하지만 각기 다른 아키텍쳐를 갖고 있어, 특정 알고리즘이나 모델에 편향되지 않는 결과를 얻기에 적합하다.

(2) 질적 데이터 정량화 지표: LLMq

본 연구는 Tai et al.(2024)이 제안한 Large Language Model Quotient(LLMq) 접근법과 Gilardi et al.(2023)의 실증적 연구 결과를 기반으로 기획되었다. Tai et al.(2024)는 환자의 주관적 경험을 정량적으로 분석하기 위해 LLMq라는 질적데이터 정량화 지표를 제시하였다[10,11]

본 연구에서는 환자들이 네 가지 주요 영역(보행/균형, 사회적 참여, 감정/심리, 미래/기대)에서 경험하는 정서를 긍정(Positive)과 부정(Negative)으로 구분하여 세 종류의 LLM에 대한 LLMq를 측정하였다. Tai et al.(2024)가 제안한 160회 반복 측정 프로토콜을 적용하여 LLMq 값의 안정성을 확보하였다.

(3) 신뢰도 검증 지표: ICC

본 연구에서는 세 종류의 LLM이 일관성 있는 평가를 수행하는지 검증하기 위해 평가자 간 신뢰도(Inter-Rater Reliability)를 측정하였다. 이를 위해 급내상관계수(Intraclass Correlation Coefficient, ICC)를 활용하였다.

체계적 분석을 위해, ICC 지표의 판단 기준은 Cicchetti(1994)가 제시한 심리학 분야의 표준화된 평가 도구 측정 기준을 활용하였다[16]. 그 기준은 다음과 같다.

- 우수함(Excellent): 0.75 ~ 1.00
- 좋음(Good): 0.60 ~ 0.74
- 보통(Fair): 0.40 ~ 0.59
- 낮음(Poor): 0.00 ~ 0.39

세부적으로는 ICC 모형 중 ICC(3,1)을 선택하였다. ICC(3,1)은 "Two-way Mixed Effects, Single Measures" 모형으로, 평가자들이 고정된(fixed) 집단이며 단일 평가자의 측정값에 대한 신뢰도를 평가한다. Two-way Mixed Model은 특정 평가자의 개별적인 차이를 인정하는 것과 동시에 선택한 기준에 따른 평가 일관성에 초점을 맞출 수 있으며, 이러한 접근 방식은 인간 평가의 신뢰성과 AI 모델 성능과 비교할 때 중요한 요소다[17][18].

따라서 ICC(3,1)을 선택한 근거는 다음과 같다. 첫째, 본 연구는 ChatGPT, Claude, Gemini라는 특정 LLM 세 종류를 의도적으로 선택하여 사용하였다. 둘째, 연구 결과를 이 세 가지 특정 LLM에 한정하여 해석하는 것이 적절하며, 무작위로 선택된 표본으로 간주하기보다는 고정된 평가자 집합으로 보는 것이 타당하다. 셋째, ICC(3,1)은 절대적 일치도(absolute agreement)를 측정하여 평가자 간 체계적 편향까지 고려한다.

ICC(3,1)과 함께 ICC(3,k) 또한 측정하였다. ICC(3,k)는 "Two-way Mixed Effects, Average Measures" 모형으로, k명의 고정된 평가자(본 연구에서는 3개 LLM)의 평균 측정값에 대한 신뢰도를 나타낸다. 이는 Spearman-Brown 예언 공식에 기반하여 계산되며, 여러 평가자의 평균을 사용할 경우 개별 측정 오차가 상쇄되어 더 높은 신뢰도를 보인다. ICC(3,k)는 향후 연구에서 이 세 LLM 합의 점수나 평균값을 활용할 경우의 신뢰도를 제시하는 보조 지표로 활용된다.

라. 코드북 및 프롬프트 설계

코드북은 연구의 4가지 핵심 영역(보행/균형,

사회적 참여, 감정/심리, 미래기대)을 기준으로, 각 영역에서 나타날 수 있는 긍정적 및 부정적 핵심 키워드를 포함하여 구성하였다. 특히, 참여자의 정서 상태를 객관적으로 측정하고 지표의 명확성을 높이기 위해 정서가, 각성가 및 구체성 평정을 통한 한국어 정서단어 목록 개발 연구[12]에서 제시된 표준화된 감정 단어들을 코드북에 통합하였다. 이를 통해 LLM이 인터뷰 텍스트에 내포된 감정을 보다 객관적이고 일관된 기준으로 판단하도록 유도하였다.

코드북의 4가지 핵심영역은 다음과 같다.

- 보행/균형 영역: 보행 능력과 균형감에 대한 긍정적/부정적 감정
- 사회적 참여 영역: 사회활동과 대인관계에 대한 긍정적/부정적 감정
- 감정/심리 영역: 전반적인 정서 상태와 심리적 적응에 대한 긍정적/부정적 감정
- 미래 기대 영역: 회복과 미래에 대한 긍정적/부정적 전망

각 영역은 홍영지(2016)의 표준화된 한국어 정서단어 목록을 참조하여[12], 긍정/부정 지표를 명확히 정의하였다.

```
"보행_균형_긍정감정": {"definition": "보행이나 균형 능력에 대해 자신감, 만족감, 희망적 기대, 개선에 대한 기쁨 등 긍정적 감정을 표현하는 내용", "indicators": ["자신감", "좋아졌다", "편해졌다", "안정적", "향상", "덜무서워", "할 수 있다", "가능하다"]}
```

표 4 표준화된 한국어 정서단어 목록[12] 기반 보행 균형 긍정 단어 예시

실제로 사용한 프롬프트는 [표 2]에 작성하였다. 보행/균형, 사회적참여, 감정/심리, 미래기대 4가지 영역에 대하여 긍정 및 부정 분석을 진행할 수 있도록 코드북을 설계하였다.

```
아래 인터뷰 텍스트를 분석하여 4개 영역에서 긍정적 감정이 우세한지 부정적 감정이 우세한지 판단해주세요.

## 분석 영역 및 기준
### 1. 보행/균형 영역
긍정적 감정 (1):** 보행이나 균형 능력에 대한 자신감, 만족감, 희망적 기대, 개선에 대한 기쁨
- 핵심 지표: 자신감, 좋아졌다, 편해졌다, 안정적, 향상, 덜무서워, 할 수 있다, 가능하다
부정적 감정 (0):** 보행이나 균형 능력에 대한 두려움, 불안, 좌절감, 절망감
- 핵심 지표: 무섭다, 겁이난다, 두렵다, 불안, 어렵다, 힘들다, 넘어질까봐, 휘청
...
... (이하 다른 영역들도 동일한 형식)

## 분석 대상 인터뷰
{interview}

## 분석 지침
1. 각 영역에서 긍정적 감정과 부정적 감정을 모두 고려하되, 어느 쪽이 더 우세한지 판단
2. 직접적인 언급뿐만 아니라 맥락상 드러나는 감정도 고려
3. 해당 영역에 대한 언급이 없거나 중립적인 경우, 전체 맥락을 고려하여 판단
4. 각 영역별로 1(긍정 우세) 또는 0(부정 우세)으로 결정

## 결과 형식
반드시 다음 형식으로만 답변하세요:
[보행/균형, 사회적참여, 감정/심리, 일상자립, 미래/기대]

예: [1,0,1,1]
```

표 3 인터뷰 분석용 코드북 프롬프트

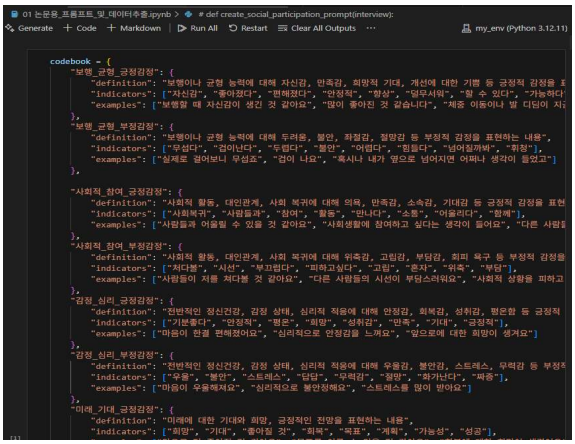


그림 2 네 가지 주요 영역에 대한 코드북

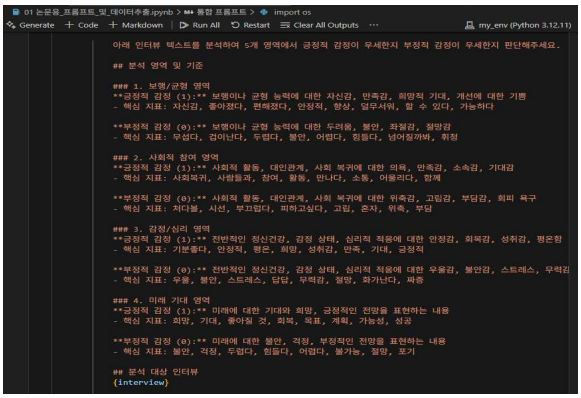


그림 3 코드북 기반 프롬프트

3. 연구 결과

가. LLMq 측정 결과 및 분석

ChatGPT 3.5-turbo, Claude 3.5-haiku, Gemini 2.5-flash-lite 세 가지 LLM을 활용하여 뇌졸중 환자 인터뷰 데이터의 4개 감정 영역(보행/균형, 사회적 참여, 감정/심리, 미래기대)에 대하여 160회 반복 측정을 한 후 LLMq를 산출하였다.

우측 후두엽 뇌경색 진단 후, 좌측 동측성 반맹을 겪은 68세 여성 인터뷰를 대상으로 LLMq를 산출한 결과는 다음과 같다.

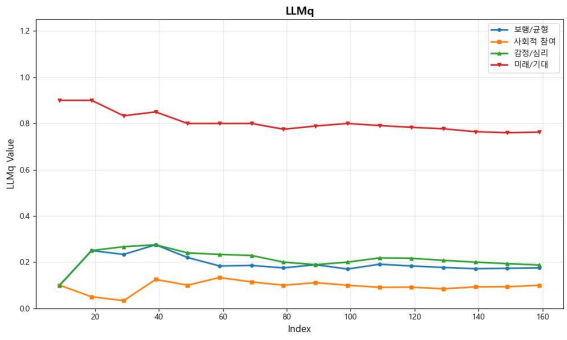


그림 5 ChatGPT 3.5 LLMq 추이

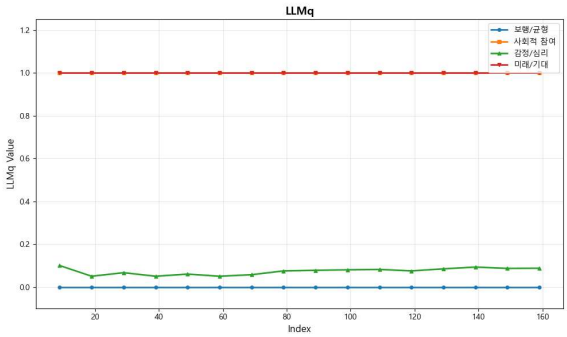


그림 6 Claude 2.5 LLMq 추이

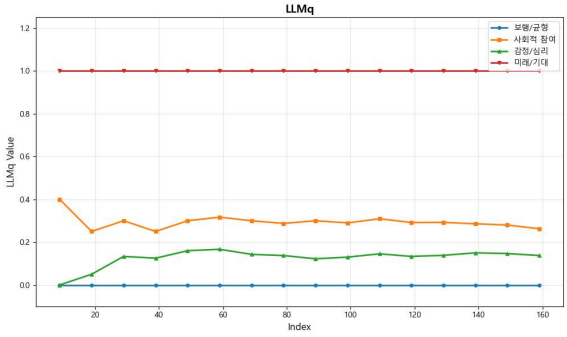


그림 7 Gemini 2.5 LLMq 추이

각 변화를 표로 정리하면 다음과 같다.

표 1 ChatGPT 3.5 LLMq 변화

횟수	보행 /균형	사회적 참여	감정 /심리	미래 기대
20	0.250	0.050	0.250	0.900
40	0.275	0.125	0.275	0.850
60	0.183	0.133	0.233	0.800
80	0.175	0.100	0.200	0.775
100	0.170	0.100	0.200	0.800
120	0.183	0.092	0.216	0.783
140	0.171	0.093	0.200	0.764
160	0.175	0.100	0.188	0.763

표 2 Claude 2.5 LLMq 변화

횟수	보행 /균형	사회적 참여	감정 /심리	미래 기대
20	0	1	0.050	1
40	0	1	0.050	1
60	0	1	0.050	1
80	0	1	0.075	1
100	0	1	0.080	1
120	0	1	0.075	1
140	0	1	0.093	1
160	0	1	0.088	1

표 3 Gemini 2.5 LLMq 변화

횟수	보행 /균형	사회적 참여	감정 /심리	미래 기대
20	0	0.250	0.050	1
40	0	0.250	0.125	1
60	0	0.316	0.167	1
80	0	0.288	0.138	1
100	0	0.290	0.130	1
120	0	0.292	0.133	1
140	0	0.286	0.150	1
160	0	0.263	0.138	1

LLMq를 분석하였을 때, 이 여성은 세 개의 LLMq 검사에서 공통적으로 보행/균형, 감정/심리 측면에서는 부정적인 감정을 내비쳤고, 미래 기대 측면에서는 긍정적인 판단을 하고 있다고 볼 수 있다. 사회적 참여 측면에서는 ChatGPT 3.5, Gemini 2.5는 각각 최종 0.100, 0.263으로 부정적인 상태라고 판단하였으나, Claude 2.5는 긍정적인 상태라고 평가하고 있다.

(2) ICC 측정 결과 및 분석

본 연구는 네 개의 분석 영역(보행/균형, 사회적 참여, 감정/심리, 미래/기대)별로 ICC를 측정하고, 전체 통합 ICC를 산출하였다. 그 결과는 다음과 같다.

표 4 ICC(3,1), ICC(3,k) 측정 결과

분야	F	p	ICC (3,1)	95% CI ICC(3,1)	ICC (3,k)	95% CI ICC(3,k)
보행 /균형	11.1	<.001	0.77	[0.59, 0.89]	0.91	[0.81, 0.96]
사회적 참여	9.97	<.001	0.75	[0.55, 0.88]	0.9	[0.79, 0.96]
감정 /심리	48.5	<.001	0.94	[0.88, 0.97]	0.98	[0.96, 0.99]
미래 기대	14.6	<.001	0.82	[0.66, 0.92]	0.93	[0.86, 0.97]
전체	13.1	<.001	0.80	[0.73, 0.86]	0.92	[0.89, 0.95]

모든 분석 영역에서 ICC(3,1) 값이 0.74 이상으로 나타나 Cicchetti(1994)의 기준에 따른 "좋음" 이상의 신뢰도를 보였다. 특히 감정/심리 영역에서 ICC(3,1) = 0.941 (95% CI [0.88, 0.97])로 "우수함" 수준의 매우 높은 일치도를 나타냈으며, F-통계량 48.47로 가장 강력한 통계적 유의성을 보였다($p<.001$). 이는 세 LLM이 감정 및 심리 관련 주제에 대해 특히 일관된 평가를 수행함을 시사한다.

다른 영역에서도 미래/기대(ICC=0.819), 보행/균형(ICC=0.771), 사회적 참여(ICC=0.749) 순으로 높은 신뢰도를 나타냈다. 사회적 참여 영역이 상대적으로 낮은 ICC를 보였으나, 여전히 "좋음" 수준을 유지하며 95% 신뢰구간 하한이 0.55로 적절한 신뢰도를 확보하였다. 모든 분야에서 F-검정 결과가 통계적으로 유의미하여($p<.001$), 평가자 간 일치도가 우연에 의한 것이 아님을 확인하였다.

(3) 전체 통합 분석

네 개 분야를 통합한 전체 ICC(3,1)은 0.802 (95% CI [0.73, 0.86])로 나타났다. 이는 ChatGPT, Claude, Gemini가 다양한 질적 분석 영역에서 전반적으로 높은 일관성을 유지함을 입증한다. 통합 ICC 값이 개별 분야들의 평균적 수준을 반영하며, 분야별 변동성에도 불구하고 안정적인 신뢰도를 보인다.

ICC(3,k) 분석 결과, 모든 영역에서 0.90 이상의 "우수함" 수준 신뢰도를 보였다. 특히 감정/심리 영역에서 ICC(3,k) = 0.979로 거의 완벽에 가까운 일치도를 나타냈으며, 전체 통합 ICC(3,k)는 0.924로 매우 높게 나타났다. 이는 세 LLM의 평균 점수를 활용할 경우 모든 분석 영역에서 매우 안정적이고 신뢰할 수 있는 측정이 가능함을 시사한다.

(4) LLM 영역별 불일치 원인 분석

Claude의 사회적 참여 영역 극단적 판단 경향

본 연구에서 Claude-3.5-Haiku는 사회적 참여 영역에서 다른 두 LLM(GPT-4o, Gemini-2.5-Flash-Lite)과 뚜렷하게 구별되는 판단 패턴을 보였다. 영역별 일치도 분석 결과, Claude는 특정 인터뷰에서 완전한 긍정(1.0) 또는 완전한 부정(0.0)의 극단적 판단을 일관되게 내리는 경향을 나타냈다.

Claude의 극단적 판단 경향은 Anthropic의 Constitutional AI(CAI) 학습 방식에서 비롯된 구조적 특성으로 해석된다. Constitutional AI는 명확한 원칙(constitution) 기반으로 판단을 내리도록 설계되어, 모호한 중간 지대보다는 확실한 결론을 선호하는 경향이 있다. 본 연구의 프롬프트가 "긍정적 감정이 우세한지 부정적 감정이 우세한지 판단"을 요구했을 때, Claude는 이를 "명확한 이진 분류 과제"로 인식하여 중립적 표현을 최소화하고 0 또는 1의 극단값으로 수렴했다. 또한 Claude는 긍정적 표현이 일부라도 존재하면 이를 강하게 가중하는 맥락 가중치 해석 편향(Context Weighting Bias)을 보였다. 예를 들어, 인터뷰에서 "사람들과 만나고 싶긴 하지만, 몸이 불편해서 나가기 힘듭니다"라는 혼재된 표현이 나타났을 때, Claude는 "만나고 싶다"는 긍정적 의지에 높은 가중치를 부여하여 사회적 참여를 긍정(1)으로 판단한 반면, ChatGPT와 Gemini는 "나가기 힘들다"는 실제 제약에 초점을 맞춰 부정(0)으로 평가한 것으로 분석된다.

사회적 참여 영역에서 나타난 LLM 간 불일치는 이 개념 자체가 가진 다층적 특성에서 기인한다. 임상적으로 뇌졸중 환자의 사회적 참여는 크게 두 가지 분야로 구분할 수 있는데, (1) 물리적 참여(실제로 외출하고 모임에 나가는 행동), (2) 심리적 참여(참여하고자 하는 의지와 소속감)으로 볼 수 있다. Claude는 환자가 "사람들과 어울리고 싶다", "친구들이 그립다"와 같은 심리적 의지를 표현하면, 실제로 외출을 회피하거나 시선 부담으로 모임을 피하는 행동이 공존하더라도 전체 영역을 긍정으로 판단하는 경향을 보였다. 반면 ChatGPT와 Gemini는 "사람들과 멀어졌다", "모임에 나가기 어렵다"와 같은 실제 행동 제약에 더 높은 가중치를 부여했다.

Ⅲ. 결 론

본 연구는 물리치료 분야의 질적 연구에서 Large Language Model(LLM)을 활용한 분석 방법론의 타당성과 실용성을 검증하였다. 뇌졸중 환자 20명의 재활 경험에 대한 반구조화 인터뷰 분석을 통해 개발된 LLMq(Large Language Model quotient) 지표를 CChatGPT 3.5-turbo, Claude 3.5-haiku, Gemini 2.5-flash-lite에 측정한 결과, LLMq 값이 점차 수렴되는 추이를 확인할 수 있다.

평가자 간 일치도 분석 결과, ICC(3,1)은 0.802, ICC(3,k)는 0.924로 Cicchetti(1994)가 제시한 우수함의 기준에 부합한다. 이러한 결과는 ChatGPT, Claude, Gemini를 이용한 질적 분석이 분야에 관계 없이 높은 평가자 간 일치도를 보이며, 단일 LLM 사용 시에도 충분한 신뢰도를 확보할 수 있음을 입증한다. 특히 감정/심리와 같은 주관적 해석이 요구되는 영역에서도 높은 일관성을 보인다는 점은 이들 LLM의 질적 연구 도구로서의 타당성을 강력히 뒷받침한다.

하지만 LLM의 한계도 명확히 인식되고 있다. De Paoli(2024)의 연구는 GPT 3.5-Turbo를 사용한 귀납적 주제분석 실험을 통해, LLM이 대부분의 주요 주제를 추론할 수 있지만 감정적 뉘앙스와 맥락적 깊이를 파악하는 데는 한계가 있음을 보여주었다고 설명한다[19]. 우측 후두엽 뇌경색 진단 후, 좌측 동측성 반맹을 겪은 68세 여성 인터뷰 LLMq 측정 결과, ChatGPT, Gemini LLM는 각각 LLMq를 0.100, 0.263으로 사회적 참여 분야에서 환자가 부정적이라고 판단하였으나, Claude는 1.000으로 환자가 긍정적인 상태라고 판단하였다. 즉, LLM이 절차적이고 구조적인 요소에는 강점을 보이지만, 정서적이고 심리적인 차원의 복합적인 맥락 해석에는 인간 연구자의 전문성이 여전히 필수적이라고 볼 수 있다.

사회적 참여 영역에서 나타난 LLM 간 판단 불일치는 이 구성개념이 가진 본질적 이중성, 심리적 욕구(wanting to engage)와 실제 행동(actual engagement)에서 기인한다. 본 연구의 코드북은 "의욕, 만족감, 소속감, 기대감"과 같은 내적 동기와 "위축감, 고립감, 회피 욕구"와 같은 정서적 반응을 단일 이진 판단(긍정/부정)으로 압축하도록 설계되었으나, 임상 현장에서 뇌졸중 환자는 "친구들과 만나고 싶지만, 몸이 불편해서 집에만 있습니다"와 같이 심리적 의지와 실제 행동이 불일치하는 진술을 빈번히 한다. 이는 코드북 설계 단계에서 개념적 경계를 명확히 하지 못한 데서 비롯된 것으로, 향후 연구에서는 ICF 프레임워크에 따라 사회적 참여를 '활동(Activity)'과 '참여(Participation)'로 세분화하거나, 프롬프트에 "실제 행동이 심리적 욕구보다 우선한다"는 명확한 판단 기준을 제시함으로써 LLM 판단의 일관성을 높일 필요가 있다.

본 연구에서 제안한 LLMq 지표와 분석 방법론은 물리치료 분야의 환자 중심 연구와 근거 기반 실무 발전에 기여할 수 있는 혁신적 도구로서, 임상 현장과 연구 영역 모두에서 질적 분석의 효율성과 객관성을 동시에 향상시킬 수 있는 방법론적 기반을 제공한다.

하지만 더 큰 표본 크기에서의 검증과 다양한 환자군에 대한 적용 연구가 필요하다. 또한 LLM의 한계를 보완하기 위한 인간-AI 협력 모델 개발과 도메인 특화 모델의 활용 방안에 대한 후속 연구가 요구된다.

본 연구에서 제안한 LLMq 지표와 분석 방법론은 물리치료 분야의 환자 중심 연구와 근거 기반 실무 발전에 기여할 수 있는 혁신적 도구로서, 임상 현장과 연구 영역 모두에서 질적 분석의 효율성과 객관성을 동시에 향상시킬 수 있는 방법론적 기반을 제공한다.

후속 연구로써 다음과 같은 구체적인 연구 질문을 제시하고자 한다. 첫째, 뇌졸중 환자에서 확장하여, 신경계 및 근골격계 환자의 기능적 움직임 회복 속도와 심리적 회피 행동 간의 시간적 인과관계를 규명하여, 재활 초기 단계에서 공포-회피 신념이 장기 재활 성과에 미치는 영향을 정량화하는 방법을 도입하는 것을 제안한다. 예를 들어 로봇 보조 보행 훈련(RAGT)이 전통적 수동 훈련에 비해 환자의 "넘어질 것 같은 두려움"을 얼마나 감소시키는지, 그리고 이러한 심리적 안전감이 실제 보행 속도 개선으로 이어지는지를 LLMq 기반 종단 연구로 검증할 필요가 있다. 둘째, 재활 기기의 심리적 수용도를 연령대, 기술 친숙도, 통제감 상실에 대한 민감도에 따라 분석하여, 고령 환자나 기술 거부감이 높은 환자에게 적합한 로봇 적응 프로토콜을 개발하는 것이 가능하다.

또한 LLM의 한계를 보완하기 위한 인간-AI 협력 모델 개발과 도메인 특화 모델의 활용 방안에 대한 연구가 요구된다. 본 연구는 범용 LLM(GPT-4o, Claude-3.5-Haiku, Gemini-2.5-Flash-Lite)을 사용했으나, 향후 물리치료 전문 용어와 한국어 감정 표현에 최적화된 파인튜닝(fine-tuning) 모델을 개발하면 일상/자립 영역 같은 낮은 일치도 영역의 신뢰도를 추가로 향상시킬 수 있을 것이다.

REFERENCES

[1] Bastemeijer, C. M., Voogt, L. P., Hazelzet, J. A., & Ewout, W. S., "Patient values in physiotherapy practice, a qualitative study," *Physiotherapy Research International,* vol. 25, no. 3, e1862, 2020.

[2] O'Keeffe, M., et al., "What influences patient-therapist interactions in musculoskeletal physical therapy? Qualitative systematic review and meta-synthesis," *Physical Therapy,* vol. 96, no. 5, pp. 609-622, 2016.

[3] Morera-Balaguer, J., et al., "Physical therapists' perceptions and experiences about barriers and facilitators of therapeutic patient-centred relationships during outpatient rehabilitation: a qualitative study," *Brazilian Journal of Physical Therapy,* vol. 22, no. 6, pp. 484-492, 2018.

[4] Kallio, H., Pietilä, A. M., Johnson, M., & Kangasniemi, M., "Systematic methodological review: developing a framework for a qualitative semi structured interview guide," *Journal of Advanced Nursing,* vol. 72, no. 12, pp. 2954-2965, 2016.

[6] Wijma, A. J., et al., "Patient-centeredness in physiotherapy: What does it entail? A systematic review of qualitative studies," *Physiotherapy Theory and Practice,* vol. 33, no. 11, pp. 825-840, 2017.
[7] McGrath, C., Palmgren, P. J., & Liljedahl, M., "Twelve tips for conducting qualitative research interviews," *Medical Teacher,* vol. 41, no. 9, pp. 1002-1006, 2019.
[8] Malterud, K., "Qualitative research: standards, challenges, and guidelines," *The Lancet,* vol. 358, no. 9280, pp. 483-488, 2001.
[9] Noble, H., & Smith, J., "Issues of validity and reliability in qualitative research," *Evidence-Based Nursing,* vol. 18, no. 2, pp. 34-35, 2015.
[10] Tai, R. H., et al., "An examination of the use of large language models to aid analysis of textual data," *International Journal of Qualitative Methods,* vol. 23, pp. 1-14, 2024.
[11] Gilardi, F., Alizadeh, M., & Kubli, M., "ChatGPT outperforms crowd workers for text-annotation tasks," *Proceedings of the National Academy of Sciences,* vol. 120, no. 30, 2023.
[12] 홍영지, 남예은, 이윤형, "정서가, 각성가 및 구체성 평정을 통한 한국어 정서단어 목록 개발," *인지과학*, 제27권, 제3호, 377-406쪽, 2016년 9월.
[13] Reed, G. F., Lynn, F., & Meade, B. D., "Use of coefficient of variation in assessing variability of quantitative assays," *Clinical and Diagnostic Laboratory Immunology,* vol. 9, no. 6, pp. 1235-1239, 2002.
[14] McHugh, M. L., "Interrater reliability: The kappa statistic," *Biochemia Medica,* vol. 22, no. 3, pp. 276-282, 2012.
[15] Landis, J. R., & Koch, G. G., "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159-174, 1977.
[16] Cicchetti, D. V., "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychological Assessment,* vol. 6, no. 4, pp. 284-290, 1994.
[17] Koo, T. K., & Li, M. Y., "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine,* vol. 15, no. 2, pp. 155-163, 2016.
[18] Yavuz, F., "Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments," *British Journal of Educational Technology,* vol. 56, no. 1, pp. 125-143, 2024.
[19] De Paoli, S., "Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach," *Social Science Computer Review,* vol. 42, no. 4, pp. 997-1019, 2024.
[20] Naqvi, W., Shaikh, R., & Mishra, A., "Large language models in physical therapy: Time to adapt and adept," *Frontiers in Public Health,* vol. 12, 1364660, 2024.

저 자 소 개

유성훈(종신회원)

남부대학교 물리치료학과 부교수.

<주관심분야 : 스마트미디어,헬스케어>

이설의(정회원)

2020년 광주과학기술원 전기전자컴퓨터공학부 학사 졸업.

<주관심분야 : 인공지능, LLM, 소프트웨어 공학>

윤주상(정회원)

2025년 남부대학교 물리치료학 석사과정

<주관심분야 : 물리치료학, AI융합, 헬스케어>