Journal of Korean Medicine Rehabilitation Vol. 36 No. 4, October 2026
pISSN 1229-1854 eISSN 2288-4114
https://doi.org/10.18325/jkmr.2026.36.4.221

JKMR
Original Article

대규모 언어모델을 활용한 한국어 통증 표현의 강도 추정과 맥락 민감도 분석: 임상의사결정지원시스템 적용을 위한 Claude 기반 탐색적 연구

임동우
동국대학교 한의과대학 진단학교실

Analysis of Pain Intensity Interpretation and Context Sensitivity in Korean Pain Expressions Using Generative Artificial Intelligence: A Claude-Based Exploratory Study for Application to Clinical Decision Support Systems

Dong-Woo Lim, K.M.D., Ph.D.
Department of Diagnostics, College of Korean Medicine, Dongguk University

Objectives Pain is subjective and expressed through language, yet how large language models numerically interpret it remains unclear. This study explored how Claude's pain-intensity estimates for Korean pain expressions vary with the expression itself and its surrounding context (body site, duration, temporal pattern, presumed cause, intensity modifier).

Methods Twelve Korean pain expressions spanning sensory, neuropathic, muscular tension-stiffness, deep somatic, and superficial pain characteristics were selected. From a neutral baseline sentence, stimulus sentences were constructed by varying one of five contextual factors at a time. Each condition was submitted to Claude (Claude Sonnet 4.6) as five independent requests, yielding structured JSON responses including pain intensity (0 to 10). The intensity-modifier factor served as a positive control.

Results Across 1,260 requests, no JSON parsing failures occurred. The expected monotonic order for the intensity modifier (slightly〈baseline〈very〈unbearably) held for all 12 expressions. In mixed-effects models, the intensity modifier showed the largest and most consistent effects (β=−1.36 to +3.51, all $p<0.001$), while the other four factors showed smaller, level-specific effects. Composite context sensitivity ranged from 0.47 (*ppeogeunhada, hwakkeungeorida*) to 0.92 (*ppeotppeothada*), with expression-specific patterns including a site-specific negative association for *jjiritada* and an asymmetric, direction-dependent response for *jjireuda*.

Conclusions Claude responded consistently to explicit intensity cues, whereas responses to non-intensity contextual factors were smaller and heterogeneous across pain expressions. Automated numerical interpretation of free-text symptom descriptions warrants caution in Korean medicine diagnostics and clinical decision support. This exploratory, single-model study warrants confirmatory follow-up with larger, multi-model designs. **(J Korean Med Rehabil 2026;36(4):221-234)**

Key words Large language model, Artificial intelligence, Pain measurement, Natural language processing, Clinical decision support system

The raw API responses, stimulus sentences and prompts, analysis dataset, and analysis code for this study are publicly available on Zenodo at https://doi.org/10.5281/zenodo.22831589.

RECEIVED August 6, 2026
REVISED September 28, 2026
ACCEPTED October 1, 2026

CORRESPONDING TO
Dong-Woo Lim, Department of Diagnostics, College of Korean Medicine, Dongguk University, 32 Dongguk-ro, Ilsandong-gu, Goyang 10326, Korea

TEL (031) 961-5833
FAX (031) 961-5835
E-mail greatwoodong@dongguk.edu

Copyright © 2026 The Society of Korean Medicine Rehabilitation

서론»»»

최근 대규모 언어모델(large language model, LLM)은 비정형 임상정보의 요약, 환자 상태와 관련된 정보의 추출, 임상질문에 대한 근거 검색, 진단 및 치료 선택지의 제시 등 다양한 의료업무에 활용되고 있다[1]. 이에 따라 LLM을 전자의무기록 및 임상정보시스템과 결합하여 의료진의 판단을 보조하려는 연구도 빠르게 증가하고 있다. 그러나 실제 임상환경에서의 적용은 아직 제한적이며, 현재의 LLM 기반 시스템은 의료진의 판단을 대체하기보다는 임상정보를 정리하고 의사결정을 지원하는 보조적 도구로 활용되는 단계에 있다[2].

임상의사결정지원시스템(clinical decision support system, CDSS)은 환자의 특성과 임상정보를 의학지식 또는 알고리즘과 결합하여 진단, 검사, 처방, 치료 및 환자안전과 관련된 권고를 의료진에게 제공하는 시스템이다. 기존의 CDSS가 주로 구조화된 전자의무기록 자료와 사전에 정의된 규칙을 활용하였다면, LLM 기반 시스템은 진료기록, 환자의 자유서술, 의사-환자 대화 및 임상지침과 같은 비정형 자연어 자료까지 처리할 수 있다는 장점이 있다[3]. 한의학 분야에서도 자연어로 기술된 환자 상태로부터 관련 한의 증상명을 추천하는 시스템이 개발되었으며[4], 학파별 진단과 처방의 차이를 반영한 기계학습 자료의 구축도 시도되었다[5]. 이러한 연구는 한의 임상정보를 구조화하고 진단 및 처방을 지원할 수 있는 가능성을 보여준다.

최근에는 진료실에서 이루어지는 의사-환자 대화를 자동으로 인식하고 전사한 후, 이를 진료기록 초안이나 요약문으로 변환하는 ambient clinical documentation 또는 ambient artificial intelligence scribe 기술도 도입되고 있다. 향후 두 기능이 연계되면, 환자가 음성으로 표현한 증상이 자동으로 전사되고, 구조화된 임상개념으로 변환된 뒤, 다시 진단 및 치료지원 알고리즘의 입력자료로 사용될 수 있다. 이러한 과정에서 서로 다른 언어적・음향적 특성을 가진 화자집단에 대해 음성인식 성능의 차이가 발생할 수 있음이 보고되었으며[6], 임상문서의 음성인식 과정에서도 단어의 누락, 대체 및 삽입 오류가 발생하며, 일부는 임상적으로 중요한 의미를 변화시킬 수 있음이 보고되었다[7]. 음성인식과 이후의 LLM 기반 의미해석은 서로 다른 기술 단계이지만, 임상 자연어 정보가 최종 의사결정지원에 활용되기 위해서는 각 단계에서의 정보 변환 특성을 독립적으로 검증할 필요가 있다. 본 연구는 이 중 문자화된 증상 표현을 LLM이 어떻게 해석하고 수치화하는지에 초점을 두었다.

환자와 의료진 사이에서도 동일한 표현이 항상 동일한 의미로 이해되는 것은 아니다. 통증은 본질적으로 주관적이며 언어적 서술을 통해 의료진에게 전달된다. 그러나 통증 표현은 언어・문화에 따라 다르고, 환자와 의료진 사이에서도 통증 강도와 의미에 대한 불일치가 발생할 수 있다[8]. 또한 통증은 객관적인 수치만으로 전달되는 것이 아니라 환자가 선택한 표현, 질문 방식, 대화의 맥락 및 의료진의 해석을 통해 전달된다[9]. 일반적인 의료표현에서도 의료진이 의도한 의미와 환자가 이해한 의미가 다르거나, 일부 표현은 의도와 반대되는 뜻으로 이해되는 사례가 보고되었다[10].

이러한 문제는 통증과 같이 객관적 검사만으로 완전히 확인하기 어렵고 환자의 언어적 서술에 크게 의존하는 증상에서 더욱 중요하다. ‘찌릿하다’, ‘욱신거리다’, ‘쑤시다’, ‘시큰거리다’, ‘저리다’, ‘화끈거리다’와 같은 표현은 통증 강도뿐 아니라 감각적 성질, 발생양상 및 불쾌감을 함께 전달한다. 그러나 통증의 성질을 나타내는 각각의 표현이 특정한 숫자척도와 일대일로 대응하는 것은 아니다. 같은 표현도 환자와 의료진의 경험, 통증 부위 및 문장의 주변 맥락에 따라 다르게 해석될 수 있다[9]. 따라서 자연어를 처리하는 인공지능 역시 통증과 관련한 문장에 포함된 단어를 인식하는 것을 넘어, 표현의 임상적 의미와 불확실성을 적절히 구분해야 한다.

LLM은 동일한 임상질문에 대해서도 프롬프트의 구성이나 질문 방식에 따라 다른 응답을 생성할 수 있다. 실제로 의료지침에 관한 동일한 질문을 반복하였을 때 모델과 프롬프트에 따라 응답의 일관성이 달라지는 것으로 나타났다[11]. 또한 실제 임상의사결정 과정을 모사한 연구에서는 LLM이 정보의 양과 제시 순서에 민감하고, 지시사항이나 임상지침을 안정적으로 따르지 못하는 한계가 보고되었다[2]. 이는 LLM이 환자의 증상 표현을 고정된 임상개념으로 처리하기보다 입력 문장의 주변 맥락과 확률적 생성 과정에 따라 유동적으로 해석할 수 있음을 시사한다.

따라서 인공지능 기반 진료기록 및 임상의사결정지

원 기술이 확대되는 상황에서, 한국어 통증 표현에 대한 LLM의 해석 특성을 체계적으로 평가할 필요가 있다. 본 연구는 CDSS 자체를 구축하거나 그 임상적 성능을 평가하는 연구가 아니라, 향후 LLM 기반 CDSS에서 자유서술형 환자정보를 구조화하여 활용하기 위해 선행되어야 할 자연어 해석 단계의 특성을 탐색적으로 평가하는 연구이다. 본 연구에서는 서로 다른 한국어 통증 표현을 통증 부위, 지속기간, 발생양상, 원인 및 강도 수식어 등의 조건과 결합하여 제시하고, LLM이 추정한 통증 강도의 변화와 반복 일관성을 분석하고자 한다. 이를 통해 한국어 통증 표현이 모델에서 안정된 임상개념으로 처리되는지, 또는 주변 맥락에 따라 유동적으로 수치화되는지를 확인하고, 향후 한의 진료기록 및 임상의사결정지원시스템에서 자연어 통증 정보를 활용하기 위한 표준화와 안전장치의 필요성을 제시하고자 한다.

대상 및 방법»»»

1. 연구 설계

본 연구는 대규모 언어모델이 한국어 통증 표현으로부터 추정하는 통증 강도의 특성과 맥락 의존성을 평가하기 위한 탐색적 실험연구이다. 실제 환자 또는 의무기록은 사용하지 않았으며, 연구자가 사전에 작성한 가상 임상 문장을 단일 대규모 언어모델에 입력하여 생성된 응답을 분석하였다.

본 연구의 일차 목적은 동일한 한국어 통증 표현에 부여되는 통증 강도 추정값이 통증 부위, 지속기간, 발생양상, 발생원인 및 강도 수식어에 따라 변화하는지를 평가하는 것이었다. 이차 목적은 동일한 입력을 반복하였을 때 모델이 생성하는 통증 강도 추정값의 반복 안정성을 평가하는 것이었다.

본 연구는 모델의 추정값과 실제 환자의 통증 강도 사이의 정확도를 평가하는 진단 정확도 연구가 아니라, 특정 대규모 언어모델의 언어적 해석 특성을 평가하는 연구로 설계하였다.

2. 한국어 통증 표현 선정

한국 근골격계 통증 환자에게 사용 빈도가 보고된 통증 기술어와 임상적 사용 가능성을 바탕으로 12개의 한국어 통증 표현, 즉 아리다(*arida*), 쑤시다(*ssusida*), 찌르다(*jjireuda*), 당기다(*danggida*), 욱신거리다(*uksingeorida*), 찌릿하다(*jjiritada*), 따끔하다(*ttakkeumhada*), 저리다(*jeorida*), 뻐근하다(*ppeogeunhada*), 묵직하다(*mukjikhada*), 뻣뻣하다(*ppeotppeothada*), 화끈거리다(*hwakkeungeorida*)를 선정하였다[8].

의료인이 주로 사용하는 전문용어인 '동통'과 '산통'은 환자의 자연스러운 구어 표현과 차이가 있어 제외하였다. '쿡쿡 쑤신다', '욱신욱신하다', '전기가 오는 것 같다'와 같은 반복형 또는 비유형 표현은 기본 분석에 포함하지 않고, 의미보존형 언어 변형에 관한 보조 분석 후보로 분류하였다.

각 통증 표현에 대해서는 결과의 기술 및 시각화 편의를 위해 저자가 선행문헌과 일반적인 임상적 의미를 참고하여 기술적 범주를 부여하였다. 이러한 범주는 사전에 검증된 통증 분류체계에 근거한 것이 아니며, 독립된 전문가 평가나 분류자 간 일치도 검증을 시행하지 않은 저자 정의의 편의적 기술 범주이다.

3. 실험 문장의 설계

모든 조건은 다섯 요인이 공유하는 단일 기준(참조) 문장에서 출발하며, 각 조건은 이 기준 문장에 해당 요인의 수식구 하나만을 삽입하거나 대상 명사(부위)를 치환하여 구성하였다. 이에 따라 기준 문장과 조건 문장은 정확히 하나의 요소에서만 차이가 난다.

기준 문장은 다음과 같다: "환자가 허리 통증이 '{통증 표현}'고 말하였다."

기준 문장에는 환자의 성별, 연령, 질환명, 진료장소, 지속기간, 발생원인 및 통증 빈도에 관한 정보를 포함하지 않았다.

통증 표현 이외의 문맥이 강도 추정에 미치는 영향을 평가하기 위해 다음 다섯 요인을 설정하였다(각 수준의 예시 문장은 Table I 참조).

Table I. Experimental Factors, Levels, and Example Stimulus Sentences

Factor	Level	Example stimulus sentence (English translation; original Korean in italics)
A. Body site	Head	The patient said the head pain felt *jjiritada*. *환자가 머리 통증이 '찌릿하다'고 말하였다.*
	Knee	The patient said the knee pain felt *jjiritada*. *환자가 무릎 통증이 '찌릿하다'고 말하였다.*
	Abdomen	The patient said the abdominal pain felt *jjiritada*. *환자가 복부 통증이 '찌릿하다'고 말하였다.*
B. Duration	Onset today	The patient said the low back pain, which started today, felt *jjiritada*. *환자가 오늘부터 시작된 허리 통증이 '찌릿하다'고 말하였다.*
	Persisting 1 week	"…which had persisted for 1 week…" *환자가 1주째 지속되는 허리 통증이 '찌릿하다'고 말하였다.*
	Persisting 3 months	"…which had persisted for 3 months…" *환자가 3개월째 지속되는 허리 통증이 '찌릿하다'고 말하였다.*
C. Temporal pattern	Intermittent	"…low back pain occurring intermittently…" *환자가 간헐적으로 발생하는 허리 통증이 '찌릿하다'고 말하였다.*
	Persistent	"…occurring persistently…" *환자가 지속적으로 발생하는 허리 통증이 '찌릿하다'고 말하였다.*
	Occurs in the morning	"…occurring in the morning…" *환자가 아침에 발생하는 허리 통증이 '찌릿하다'고 말하였다.*
	Occurs at night	"…occurring at night…" *환자가 야간에 발생하는 허리 통증이 '찌릿하다'고 말하였다.*
D. Presumed cause	After trauma	The patient said the low back pain, which occurred after trauma, felt *jjiritada*. *환자가 외상 이후 발생한 허리 통증이 '찌릿하다'고 말하였다.*
	After overuse	"…which occurred after overuse…" *환자가 과사용 이후 발생한 허리 통증이 '찌릿하다'고 말하였다.*
	Related to degenerative disease	"…related to degenerative disease…" *환자가 퇴행성 질환과 관련된 허리 통증이 '찌릿하다'고 말하였다.*
E. Intensity modifier	Slightly	The patient said the low back pain felt slightly *jjiritada*. *환자가 허리 통증이 약간 '찌릿하다'고 말하였다.*
	Very	"…felt very *jjiritada*." *환자가 허리 통증이 매우 '찌릿하다'고 말하였다.*
	Unbearably	"…felt unbearably *jjiritada*." *환자가 허리 통증이 참기 어려울 정도로 '찌릿하다'고 말하였다.*

Example sentences are shown for the expression '찌릿하다(jjiritada)'; analogous sentences were constructed for the remaining 11 expressions by substitution. All conditions were derived from a single baseline sentence, "The patient said the low back pain felt *jjiritada*" by inserting or substituting exactly one contextual phrase per condition.

1) 통증 부위(body site, A): 허리(기준조건), 머리, 무릎, 복부
2) 지속기간(duration, B): 정보 없음(기준조건), 오늘 시작, 1주간 지속, 3개월간 지속
3) 발생양상(temporal pattern, C): 정보 없음(기준조건), 간헐적으로 발생, 지속적으로 발생, 아침에 발생, 야간에 발생
4) 원인(presumed cause, D): 정보 없음(기준조건), 외상 이후, 과사용 이후, 퇴행성 질환과 관련
5) 강도 수식어(intensity modifier, E; positive control): 수식어 없음(기준조건), 약간, 매우, 참기 어려울 정도로

각 실험 문장에서는 기준 문장에 한 번에 하나의 요인 수준만을 추가하거나 치환하는 one-factor-at-a-time 방식이 적용되었으며, 다섯 요인이 모두 동일한 기준 문장을 공유하도록 문장 구조를 통일하였다. 이에 따라 서로 다른 요인이 동시에 변화하거나 요인 간 문장 구

조가 달라짐으로써 발생할 수 있는 교란을 최소화하였다.

4. Large language model과 연구 조건

분석에는 Anthropic의 Claude Sonnet 4.6을 사용하였다. 정확한 API 모델 식별자는 claude-sonnet-4-6이었다. 모든 실험은 Claude API를 이용하여 수행하였으며, 각 요청은 이전 대화 내용이 전달되지 않는 독립된 요청으로 실행하였다.

모든 호출에서 모델명과 API 모델 ID, 실행 날짜와 시간, temperature, maximum output tokens, system prompt와 user prompt, 입력 문장, 응답 원문, API 상태 코드, 응답 파싱 성공 여부를 기록하였다.

최종 실행 일자는 2026년 9월 16일부터 18일까지였다. Maximum output tokens는 500으로 고정하였다. System prompt는 별도로 설정하지 않았으며, 모든 지시문은 단일 user 메시지에 포함하였다. API 응답 객체에서 확인된 모델 식별자는 요청 시 지정한 모델명(claude-sonnet-4-6)과 동일하였다. 본 연구의 원시 API 응답, 자극 문장 및 프롬프트 전문, 분석용 데이터셋 및 분석 코드는 공개 저장소(Zenodo)에 기탁하였다.

웹 기반 대화창은 모델 설정과 이전 대화 맥락을 완전히 통제하기 어렵기 때문에 본 분석에는 사용하지 않았다.

주 분석에서는 temperature를 1로 고정하였다. 결정론적 조건(temperature=0)에서는 동일 조건의 반복 안정성이라는 이차 목적 자체를 평가할 수 없기 때문으로, 확률적 생성 조건에서 맥락 효과와 반복 변동을 함께 관찰하기 위한 의도적 선택이었다. 동일 조건의 반복 안정성을 평가하기 위해 각 문장은 5회의 독립된 API 요청으로 반복하였다.

따라서 본 연구에서 보고하는 조건 간 변화량(ΔI)에는 맥락 효과와 더불어 temperature=1에 따른 확률적 생성 변동이 함께 반영되어 있을 수 있다. 이 두 요인의 상대적 기여도를 분리하기 위해서는 temperature=0 조건에서의 보조 분석이 필요하며, 이는 후속 연구 과제로 남긴다.

5. 프롬프트 설정과 출력 포맷

모든 조건에 동일한 지시 프롬프트를 적용하였다.

"다음 환자 표현에서 언어적으로 암시되는 통증 강도를 추정하십시오."

환자 표현: "{input sentence}"

0은 통증이 전혀 없는 상태이고, 10은 상상할 수 있는 가장 심한 통증입니다. 실제 환자의 통증 강도를 확인하는 것이 아니라, 주어진 표현과 문맥에서 암시되는 강도를 평가하십시오.

모델은 다음의 구조화된 JSON 형식으로 응답하도록 요청하였다.

```
{
  "pain_intensity": 0.0,
  "confidence": 0,
  "pain_quality": "",
  "reason": ""
}
```

pain_intensity는 본 연구의 정량 분석에 사용하였다. confidence, pain_quality 및 reason은 JSON 출력 형식에 포함된 보조 항목이었으나, 본 연구의 분석 평가변수에는 포함하지 않았다.

본 연구에서는 완전한 JSON 파싱을 우선 시도하고, 응답이 중간에 잘리는 등의 이유로 완전한 JSON 파싱이 실패한 경우에는 정규표현식을 이용해 pain_intensity 값만을 보조적으로 추출하는 2단계 파싱 절차를 적용하였다. 이 경우는 부분파싱(partial parse)으로 별도 표시하여 완전 파싱 결과와 구분하였다.

6. 실험 과정

모든 조건은 표현→요인 수준→반복시행의 순서로 순차적으로 실행하였으며, 별도의 순서 무작위화는 적용하지 않았다. 각 요청은 이전 요청의 대화 맥락이 전달되지 않는 완전히 독립된 API 호출로 실행되었으므로, 제시 순서에 따른 계통적 편향이 발생할 메커니즘은 없다고 판단하였다. 다만 이는 사후적 논거이며, 향후 연구에서는 제시 순서를 사전에 무작위화하는 것이

방법론적으로 더 엄격한 접근이 될 수 있다.

API 오류가 발생한 요청은 오류 유형에 따라 다르게 처리하였다. 일시적 오류(레이트리밋 등)로 판단된 경우 대기시간을 늘려가며 최대 2회까지 재시도하였다. 사용량 한도 초과(hard limit)로 판단된 경우에는 재시도를 즉시 중단하고 그때까지의 결과를 저장한 뒤, 한도가 초기화된 이후 완료되지 않은 항목만 이어서 재실행하였다. 최초 요청과 재시도 결과는 모두 보존하였다.

기록 항목은 API 요청 성공 여부, 인증 또는 사용량 제한 오류, 응답 없음, JSON 형식 위반, 통증 강도 값 누락, 0~10 범위를 벗어난 값, 정상적으로 파싱된 응답 여부로 구분하였다.

7. 평가 변수

1) 주 평가변수

일차 평가변수(primary outcome)는 "맥락 정보가 하나 추가되었을 때 Claude의 통증 강도 추정값이 기준 상태보다 얼마나 변했는가"였다. 구체적으로, 동일한 통증 표현에 대해 (1) 아무 맥락 정보도 없는 중립적인 기준 문장에서 산출된 통증 강도값과, (2) 그 문장에 부위・지속기간・발생양상・발생원인・강도 수식어 중 하나의 정보만 추가한 조건 문장에서 산출된 통증 강도값을 각각 얻은 뒤, 후자에서 전자를 뺀 값을 일차 평가변수로 정의하였다.

조건 반영 변화량(ΔI)=맥락 조건에서의 통증 강도-기준 문장에서의 통증 강도

예를 들어 "찌릿하다"의 경우, 기준 문장("환자가 허리통증이 '찌릿하다'고 말하였다.")에서 산출된 통증 강도가 3.50이었고, 여기에 강도 수식어 '매우'를 추가한 문장("환자가 허리통증이 매우 '찌릿하다'고 말하였다.")에서는 5.50이 산출되었다면, 이때의 ΔI는 5.50-3.50=2.00이 된다.

양의 값은 해당 조건에서 Claude의 추정 강도가 증가하였음을, 음의 값은 감소하였음을 의미하였다.

2) 보조적 평가변수

(1) 조건별 절대 변화량(|ΔI|)
(2) 표현별 평균 절대 변화량
(3) 표현별 기준 통증 강도의 평균 및 세션 간 범위
(4) 강도 수식어의 단조순서 충족률(positive control)

8. 통계 분석 방법

연속형 변수는 평균, 표준편차 및 범위로 기술하였다. 동일 조건의 반복시행에 따른 변동성은 각 expression-factor level 조합에서 5회 반복값의 표준편차와 범위로 평가하였다.

각 맥락 요인의 효과는 반복시행 원자료를 이용한 선형혼합효과모형(linear mixed-effects model)으로 분석하였다. factor_level은 범주형 고정효과로 처리하고, 통증 표현(expression)에 대해서는 임의절편과 factor level에 대한 임의기울기를 포함하였다. 개념적 모형은 pain_intensity~factor_level+(1+factor_level|expression)으로 설정하였다. 고정효과는 기준조건 대비 각 수준의 평균 변화량과 95% 신뢰구간으로 제시하였다. 임의절편은 표현 간 기준 통증강도의 차이를, 임의기울기는 맥락 요인에 대한 표현별 반응 차이를 반영하도록 하였다. 잔차분산은 동일한 expression-factor level 조건 내 반복시행에서 발생하는 변동의 모형 기반 추정치로 해석하였다.

다섯 맥락 요인은 수준의 수와 구조가 서로 달라 각각 독립적인 모형으로 적합하였으며, restricted maximum likelihood로 추정하였다. 강도 수식어(E)의 경우 임의절편과 임의기울기를 모두 포함한 모형이 수렴하지 않아 임의절편만 포함한 축소모형을 적용하였다. 축소모형에서는 표현별 강도 수식어 반응의 이질성이 잔차에 포함될 수 있으므로, E 요인의 잔차분산은 다른 요인의 잔차분산과 직접 비교하지 않았다. 모든 분석은 Python의 statsmodels (version 0.14)을 이용하여 수행하였다.

제시된 p-value는 다중비교 보정을 거치지 않은 탐색적 p-value이며, 확정적 결론이 아닌 탐색적 근거로 해석하였다. A-D 요인의 모형은 모두 정상적으로 수렴하였으며, 임의효과 분산 성분이 0에 가깝거나 임의효과

간 상관계수가 ±1에 가까운 singular 또는 boundary 해는 관찰되지 않았다.

결과»»»»

1. 자료 완결성 및 프롬프트 타당성

본 연구에서는 다섯 개의 맥락 요인(A-E)에 대해 총 1,260회의 독립적 API 요청을 수행하였다(A: 240회, B: 240회, C: 300회, D: 240회, E: 240회). JSON 파싱 실패율은 A-D 요인에서 0% (0/1,020건)였으며, E 요인에서도 12개 표현 모두에서 5회 반복이 완전하게 산출되어 결측치가 발생하지 않았다. 이는 구조화된 JSON 응답 형식과 정규표현식 기반 보조 파싱을 병행한 프롬프트 설계가 안정적으로 작동하였음을 시사한다.

Fig. 1은 다섯 요인의 각 수준에서 산출된 통증 강도 추정값을 12개 표현 전체에 걸쳐 평균하여 제시한 것으로, 개별 표현의 기준값 차이를 보정하기 전 원점수 수준의 개관을 제공한다. 부위(A)·지속기간(B)·발생양상(C)·발생원인(D) 네 요인에서는 수준 간 평균값의 차이가 크지 않았고 오차막대(standard deviation)가 서로 겹쳐, 시각적으로 뚜렷한 구분이 어려웠다. 이는 12개 표현의 기준강도 자체가 2.50~6.50로 넓게 분포해 있어(Table II), 요인 고유의 효과가 표현 간 이질성에 가려진 결과로 해석되며, 이 이질성을 제거한 요인별 순수 효과는 이후 Fig. 2에서 혼합효과모형 기반 추정치로 별도로 제시하였다.

반면 강도 수식어(E)에서는 이러한 표현 간 이질성에도 불구하고 뚜렷한 단조적 상승이 관찰되었다. '없음(기준)' 조건에서 평균 3.98이던 통증 강도 추정값은 '약간' 조건에서 2.62로 낮아졌으며, '매우' 조건에서 5.76, '참기 어려울 정도로' 조건에서 7.49까지 순차적으로 상승하였다. 특히 '참기 어려울 정도로' 조건의 표준편차(0.06)가 네 수준 중 가장 작았다는 점은, 표현이 무엇이든 모델이 해당 조건을 공통적으로 통증 척도의 상한부로 해석하였음을 시사한다. 이처럼 원점수 수준에서 관찰된 명확한 단조적 반응은, Claude가 문장에 명시된 강도 관련 맥락 정보를 통증 강도 추정값에 실

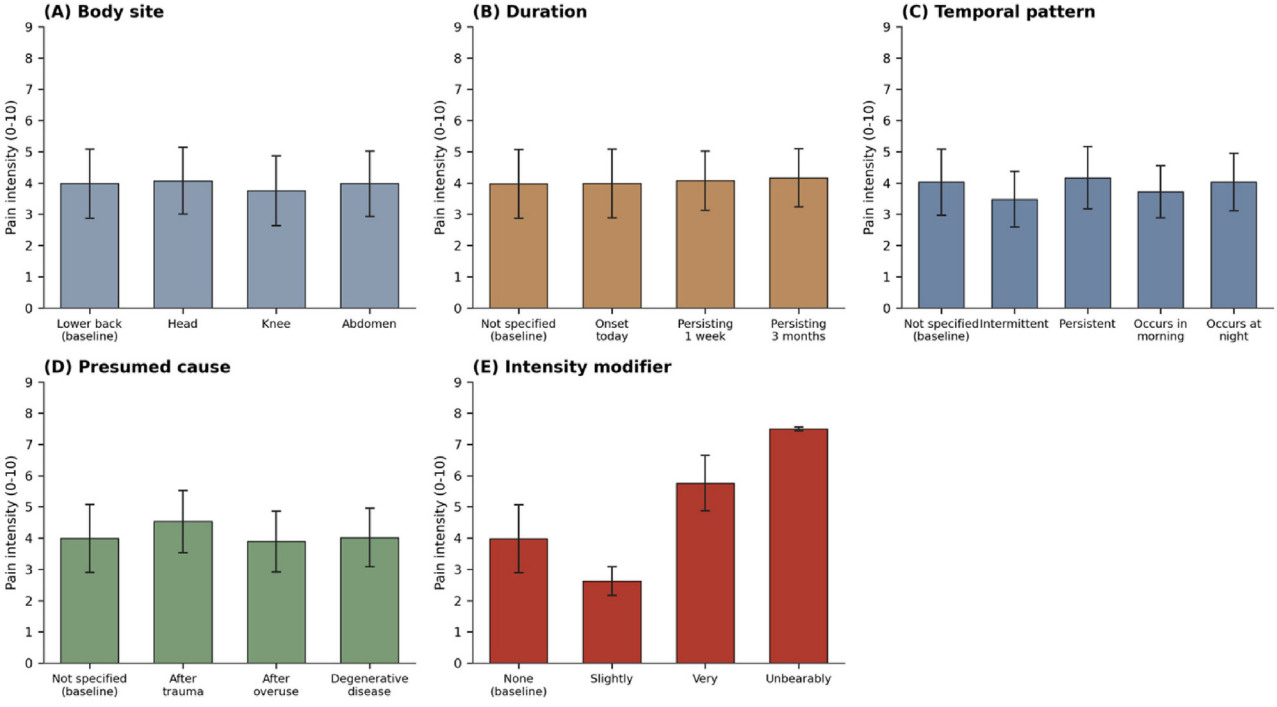

Fig. 1. Mean estimated pain intensity by contextual factor level. (A) Body site. (B) Duration. (C) Temporal pattern. (D) Presumed cause. (E) Intensity modifier. Bars and error bars represent mean±SD across 12 expressions. SD: standard deviation.

Table II. Classification of the 12 Korean Pain Expressions Analyzed in this study

Expression (Korean)	Approximate English gloss	Descriptive category	Baseline intensity, mean (range)*
아리다	Stinging/sore	Sensory descriptor	3.46 (1.00)
쑤시다	Aching	Deep somatic descriptor	3.98 (0.50)
찌르다	Stabbing	Superficial pain descriptor	6.50 (0.00)
당기다	Pulling/tight	Muscular tension-stiffness descriptor	3.48 (0.50)
욱신거리다	Throbbing	Deep somatic descriptor	4.40 (0.50)
찌릿하다	Tingling/electric-shock-like	Neuropathic descriptor	4.44 (0.50)
따끔하다	Prickling	Sensory descriptor	2.50 (0.00)
저리다	Numbness/tingling	Neuropathic descriptor	3.50 (0.00)
뻐근하다	Dull ache/tightness	Muscular tension-stiffness descriptor	3.50 (0.00)
묵직하다	Heavy	Deep somatic descriptor	3.88 (0.50)
뻣뻣하다	Stiff	Muscular tension-stiffness descriptor	2.74 (0.50)
화끈거리다	Burning	Neuropathic descriptor	5.50 (0.00)

*Baseline intensity was independently estimated in five separate analysis sessions, one preceding the testing of each contextual factor (A-E). The value in parentheses is the range (maximum minus minimum) across these five independent estimates and is presented as an informal indicator of cross-session consistency, not as a formal reliability statistic.

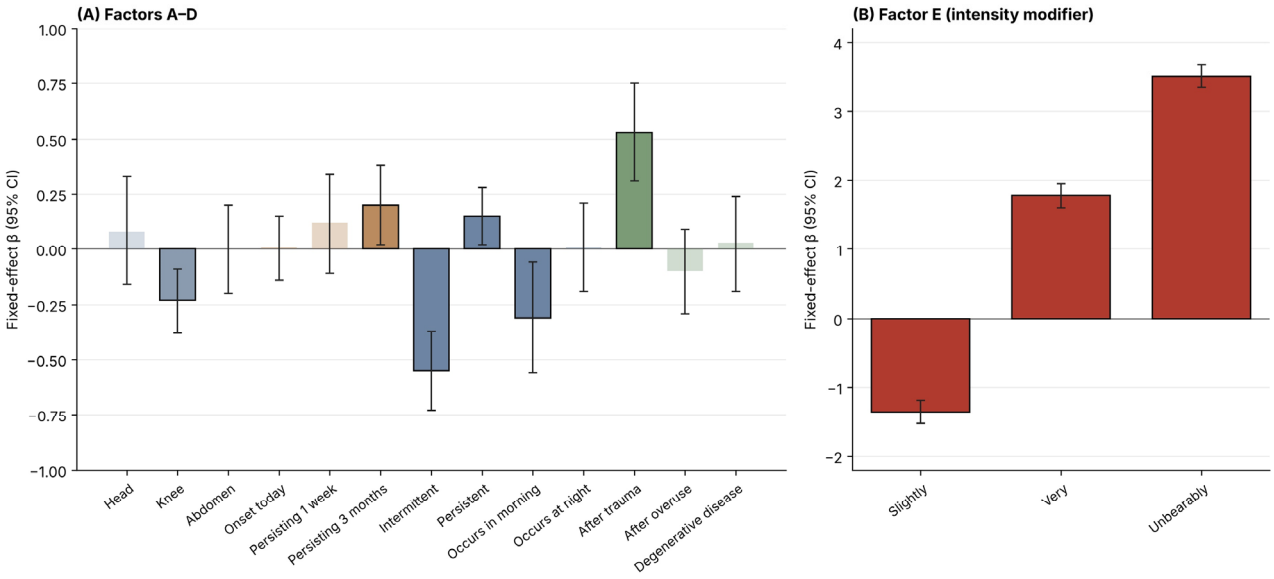

Fig. 2. Fixed-effect estimates (β±95% confidence intervals) from linear mixed-effects models, by contextual factor. (A) Factors A-D; (B) Factor E, shown on a separate scale due to its substantially larger effect size. Bars with darker shading and black borders indicate $p<0.05$. A: body site, B: duration, C: temporal pattern, D: presumed cause, E: intensity modifier.

제로 반영하고 있음을 보여주는 직접적 근거이며, 앞서 Δ값 기준으로 확인한 positive control 단조성 100% 충족을 원점수 자체로도 뒷받침한다(Fig. 3).

2. 기준 통증 강도 추정값 및 세션 간 일관성

기준 문장(“환자가 허리 통증이 ‘{통증 표현}’고 말하였다.”)에 대한 Claude의 통증 강도 추정값은 표현에 따라 2.50 (따끔하다)에서 6.50 (찌르다)까지 폭넓게 분

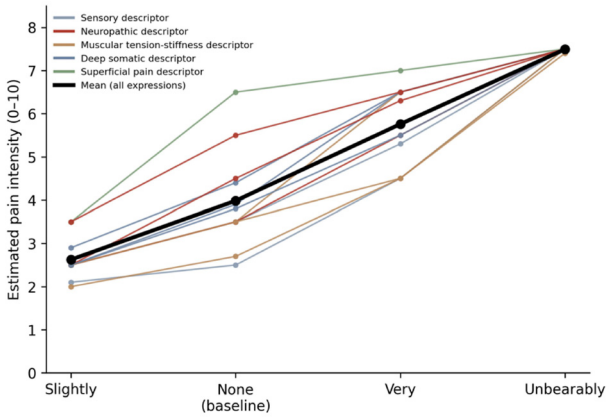

Fig. 3. Order effect of the intensity-modifier factor (positive control) across the four levels (slightly, none [baseline], very, unbearably) for each of the 12 Korean pain expressions. Thin colored lines represent individual expressions, colored by descriptive category (see legend); the thick black line represents the mean across all 12 expressions. The expected monotonic order (slightly<baseline<very<unbearably) was observed for all 12 expressions.

포하였다(Table II). 기준값은 다섯 요인을 각각 독립적으로 분석하는 과정에서 서로 다른 시점에 다섯 차례 반복 산출되었는데, 대부분의 표현에서 다섯 개 추정값의 범위가 0.00~0.50으로 매우 좁게 유지되어 세션 간 일관성이 높았다. 예외적으로 아리다는 범위가 1.00으로 가장 크게 나타나, 다른 표현에 비해 기준 조건에서도 상대적으로 불안정한 추정 양상을 보였다(Table II).

3. 맥락 요인의 전체 영향력

다섯 요인의 효과를 혼합효과모형으로 분석한 결과, 강도수식어(E)는 세 수준 모두 통계적으로 유의하였다(β=-1.36~+3.51, 모두 p<0.001). 이하 제시하는 p-value는 다중비교 보정을 거치지 않은 탐색적 p-value으로, 확정적 결론이 아닌 탐색적 근거로 해석하였다. 발생양상(C)에서는 '간헐적'(β=-0.55, p<0.001), '지속적'(β=+0.15,

Table III. Mixed-Effects Model Fixed-Effect Estimates by Contextual Factor

Factor	Level	β	95% CI	p-value	Random-slope variance	Random-intercept variance*	Residual variance*
A. Body site (ref: lower back)	Head	0.08	-0.16 to 0.33	0.506	0.178	1.314	0.026
	Knee	-0.23	-0.38 to -0.09	**0.002**	0.057		
	Abdomen	0.00	-0.20 to 0.20	1.000	0.110		
B. Duration (ref: not specified)	Onset today	0.01	-0.14 to 0.15	0.909	0.055	1.255	0.023
	Persisting 1 week	0.12	-0.11 to 0.34	0.306	0.147		
	Persisting 3 months	0.20	0.02 to 0.38	**0.033**	0.096		
C. Temporal pattern (ref: not specified)	Intermittent	-0.55	-0.73 to -0.37	**<0.001**	0.094	1.182	0.026
	Persistent	0.15	0.02 to 0.28	**0.020**	0.039		
	Occurs in the morning	-0.31	-0.56 to -0.06	**0.015**	0.181		
	Occurs at night	0.01	-0.19 to 0.21	0.936	0.118		
D. Presumed cause (ref: not specified)	After trauma	0.53	0.31 to 0.75	**<0.001**	0.139	1.245	0.032
	After overuse	-0.10	-0.29 to 0.09	0.294	0.096		
	Related to degenerative disease	0.03	-0.19 to 0.24	0.817	0.128		
E. Intensity modifier† (ref: none)	Slightly	-1.36	-1.53 to -1.19	**<0.001**	N/A (did not converge)	0.338	0.235†
	Very	1.78	1.60 to 1.95	**<0.001**			
	Unbearably	3.51	3.34 to 3.68	**<0.001**			

Values are from separate linear mixed-effects models fit for each factor: pain_intensity~factor_level+(1+factor_level|expression), estimated by REML (Python, statsmodels 0.14). β is the estimated change in Claude's pain-intensity estimate relative to the reference (baseline) level, averaged across the 12 Korean pain expressions; boldface indicates p<0.05. Random-intercept variance reflects between-expression differences in baseline intensity; residual variance reflects trial-to-trial (stochastic) variability. Random-slope variance reflects between-expression variability in the response to each level and is reported per level.
*Random-intercept and residual variances are identical across levels within the same factor and are reported once per factor.
†The full random-intercept-and-slope structure did not converge for the intensity-modifier factor; a random-intercept-only model was fitted instead. Because expression-specific slope variance could not be separated under this reduced structure, the random-intercept and residual variances for this factor are not directly comparable to those of factors A-D.

p=0.020), ‘아침에 발생’(β=-0.31, p=0.015)이 유의하였으나 ‘야간에 발생’은 유의하지 않았다(p=0.936). 발생원인(D)에서는 ‘외상 후’만 유의하였고(β=+0.53, p<0.001), 지속기간(B)에서는 ‘3개월 지속’만 유의하였다(β=+0.20, p=0.033). 부위(A)에서는 ‘무릎’만 유의한 음의 효과를 보였다(β=-0.23, p=0.002). 임의절편분산(1.18~1.31)은 표현 간 기준강도 차이를 반영하며, 잔차분산(0.023~0.032, E 제외)은 A~D에서 비교적 균일하였다(Fig. 2, Table III).

4. 표현별 맥락 민감도

12개 표현×5개 요인 히트맵(Fig. 4)에서는 강도 수식어(E 열)가 모든 표현에서 공통적으로 짙은 색으로 나타나 전 표현에 걸친 지배적 요인임이 재확인되었다. 반면 A~D 네 요인에서는 표현에 따라 뚜렷하게 다른 패턴이 관찰되었다.

12개 표현을 다섯 요인에 대한 종합 맥락 민감도(composite |ΔI|, 다섯 요인 평균 절대변화량의 평균)로 비교한 결과, 뻣뻣하다(0.92), 따끔하다(0.87), 쑤시다(0.80) 순으로 민감도가 높았고, 뻐근하다와 화끈거리다(둘 다 0.47)가 가장 낮았다(Fig. 5, Table IV). 다만 이 결과는 강도수식어를 포함한 전체 맥락 민감도로, positive control인 강도수식어의 상대적으로 큰 효과가 이 값에 실질적으로 기여하였을 수 있음에 유의해야 한다.

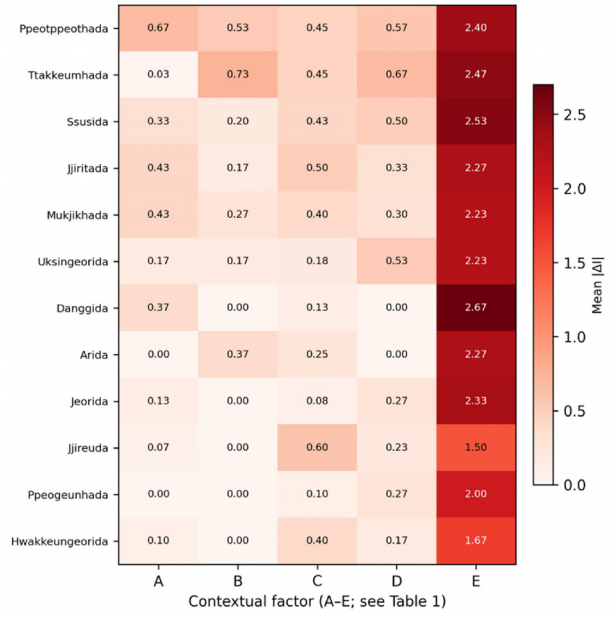

Fig. 4. Heatmap of the mean absolute change in pain-intensity estimates (|ΔI|) for each of the 12 Korean pain expressions (rows) across the five contextual factors (columns A-E; see Table I for factor and level definitions). Expressions are ranked from highest (top) to lowest (bottom) composite context sensitivity. Darker shading indicates greater sensitivity to the corresponding contextual factor. A: body site, B: duration, C: temporal pattern, D: presumed cause, E: intensity modifier.

5. 통증 표현별 특이 패턴

두 가지 표현의 분석결과에서 특징적인 패턴이 관찰되어 대표 사례로 제시하였다(Fig. 6).

1) 부위 특이적 연상(site-specific association)

찌릿하다는 12개 표현 중 유일하게 부위 요인에서 일관된 음(-)의 방향을 보였다. 기준 부위인 허리에서 머리・무릎・복부로 변경하였을 때 통증 강도 추정값은 평균적으로 감소하는 방향을 보였으나(각각 0.70, 0.30, 0.30), 무릎・복부에서의 변화 크기는 반복측정 변동과 비슷한 수준이었다(Fig. 6A). 이러한 일관된 방향성은 찌릿하다의 강도 해석이 신체부위 정보에 따라 달라질 수 있음을 시사하며, 표현 특이적인 부위 연상의 가능성을 보여준다.

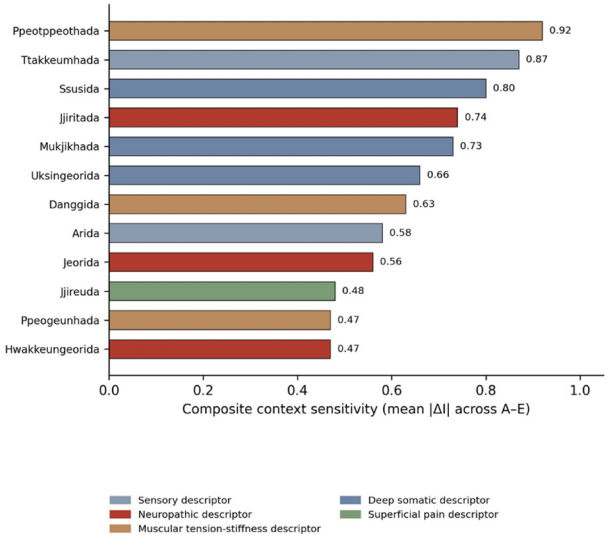

Fig. 5. Ranking of the 12 Korean pain expressions by composite context sensitivity, defined as the mean of the five factor-level mean |ΔI| values (A-E) for each expression. Bars are colored by descriptive category (see legend).

Table IV. Baseline Pain Intensity and Mean Absolute Change in Estimated Intensity (|ΔI|) by Expression Across Five Contextual Factors, Ranked by Composite Context Sensitivity

Expression	Category	Baseline, mean (range)	A*	B*	C*	D*	E*	Composite |ΔI|	PC monotonicity†
			Site	Duration	Pattern	Cause	Intensity		
뻣뻣하다(*ppeotppeothada*)	Muscular tension-stiffness descriptor	2.74 (0.50)	0.67	0.53	0.45	0.57	2.40	0.92	Met
따끔하다(*ttakkeumhada*)	Sensory descriptor	2.50 (0.00)	0.03	0.73	0.45	0.67	2.47	0.87	Met
쑤시다(*ssusida*)	Deep somatic descriptor	3.98 (0.50)	0.33	0.20	0.43	0.50	2.53	0.80	Met
찌릿하다(*jjiritada*)	Neuropathic descriptor	4.44 (0.50)	0.43	0.17	0.50	0.33	2.27	0.74	Met
묵직하다(*mukjikhada*)	Deep somatic descriptor	3.88 (0.50)	0.43	0.27	0.40	0.30	2.23	0.73	Met
욱신거리다(*uksingeorida*)	Deep somatic descriptor	4.40 (0.50)	0.17	0.17	0.18	0.53	2.23	0.66	Met
당기다(*danggida*)	Muscular tension-stiffness descriptor	3.48 (0.50)	0.37	0.00	0.13	0.00	2.67	0.63	Met
아리다(*arida*)	Sensory descriptor	3.46 (1.00)	0.00	0.37	0.25	0.00	2.27	0.58	Met
저리다(*jeorida*)	Neuropathic descriptor	3.50 (0.00)	0.13	0.00	0.08	0.27	2.33	0.56	Met
찌르다(*jjireuda*)	Superficial pain descriptor	6.50 (0.00)	0.07	0.00	0.60	0.23	1.50	0.48	Met
뻐근하다(*ppeogeunhada*)	Muscular tension-stiffness descriptor	3.50 (0.00)	0.00	0.00	0.10	0.27	2.00	0.47	Met
화끈거리다(*hwakkeungeorida*)	Neuropathic descriptor	5.50 (0.00)	0.10	0.00	0.40	0.17	1.67	0.47	Met

Composite |ΔI| is the mean of the five factor-level |ΔI| values (A-E) for each expression.
A: body site, B: duration, C: temporal pattern, D: presumed cause, E: intensity modifier, PC: positive control.
*Mean absolute change in Claude's pain-intensity estimate relative to the baseline sentence, averaged across all non-baseline levels of the given factor (five independent repetitions per level).
†Positive-control monotonicity indicates whether the expected order (slightly<baseline<very<unbearably) was observed for the intensity-modifier factor (E); met for all 12 expressions.

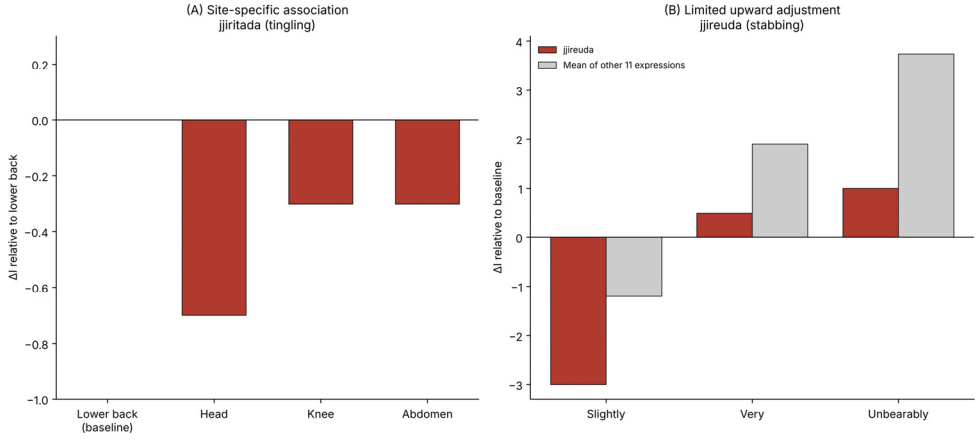

Fig. 6. Representative expression-specific patterns of context dependence. (A) Site-specific association for *jjiritada* (tingling): estimated pain intensity tended to be lower on average when the body site was changed from the baseline (lower back) to head, knee, or abdomen (ΔI=-0.70 to -0.30). (B) Limited upward adjustment for *jjireuda* (stabbing): compared with the mean response of the other 11 expressions, Jjireuda showed a larger-than-average downward adjustment to 'slightly' (ΔI=-3.0 vs. -1.2) but a markedly smaller upward adjustment to 'very' and 'unbearably' (ΔI=0.5 and 1.0 vs. 1.9 and 3.7), indicating an asymmetric response profile associated with its high baseline intensity.

2) 높은 기준강도와 관련된 비대칭적 상향 반응 제한

찌르다는 기준강도가 6.50으로 12개 표현 중 가장 높았다. 강도 수식어 조건에서 '약간'에 대한 하향 조정폭(ΔI=-3.00)은 나머지 11개 표현의 평균(-1.21)보다 오히려 컸던 반면, '매우'와 '참기 어려울 정도로'에 대한 상향 조정폭(각각 +0.50, +1.00)은 나머지 11개 표현의 평균(+1.89, +3.74)보다 뚜렷하게 작았다(Fig. 6B). 이러한 비대칭성은 해당 표현이 이미 높은 기준값을 가지고 있어 상향 방향으로의 추가 조정 여지가 제한되어 있을 가능성을 시사한다.

고찰»»»

통증을 설명하는 단어는 언어와 문화에 따라 차이가 있으며, 한 언어권에서 개발된 통증 평가 도구의 표현이 다른 문화권 환자가 실제로 사용하는 통증 어휘를 충분히 반영하지 못할 수 있다[12]. 한국의 말초신경병증성 통증 환자를 대상으로 한 연구에서도 '저림'과 '시림'을 비롯한 다양한 한국어 통증 표현이 확인되었으며, 일부 표현은 영어권 통증 기술어와 단순하게 대응하기 어려운 문화적・언어적 특성을 보였다[13].

특히 한국어에서는 의성어・의태어, 반복형 표현, 정도부사 및 감각적 비유가 증상 기술에 빈번하게 사용된다[8]. 동일한 감각도 '찌릿하다', '찌릿찌릿하다', '찌리리하다' 등으로 다양하게 표현될 수 있다. 반대로 '시리다', '아리다', '저리다'와 같은 표현은 환자와 임상상황에 따라 통증, 냉감, 감각이상 또는 불쾌감 등 서로 다른 의미로 사용될 수 있다. 또한 지역적 말투, 세대별 어휘 및 개인의 습관적인 표현방식이 더해지면, LLM이 특정 환자의 통증을 상대적으로 과소 또는 과대 추정할 가능성이 있다. 또한 인공지능이 특정 통증 표현만을 추출하거나 불확실한 서술을 확정적인 증상으로 변환하면, 실제 환자의 진술과 다른 임상정보가 기록될 수 있다. 다만 이러한 가능성이 실제 한국어 의료환경에서 어느 정도 나타나는지에 대해서는 체계적인 평가가 부족하다.

최근에는 LLM을 이용하여 환자의 자유서술형 통증 경험으로부터 통증 중증도와 기능장애를 추정하려는 연구도 이루어지고 있다. 이러한 연구는 LLM이 서술형 통증 정보를 정량화할 수 있는 가능성을 보여주었으나, 모델이 전문가보다 통증을 높게 평가하거나 평가값의 분포가 좁아지는 등의 차이도 확인되었다[14]. 따라서 통증 관련 LLM 평가에서는 모델이 통증의 강도에 관한 점수를 산출할 수 있는지 뿐만 아니라, 어떠한 언어적 단서와 주변 맥락에 의해 그 값이 변화하는지를 확인할 필요가 있다.

본 연구는 Claude가 한국어 통증 표현에 부여하는 통증 강도 추정값이 표현 자체의 의미뿐 아니라 맥락 요인, 특히 강도 수식어에 의해 크게 좌우됨을 보여주었다. 강도 수식어(E)는 세 수준 모두에서 뚜렷한 효과를 보였으며(Table III), 이는 Hager 등[2]이 지적한 것처럼 LLM의 출력이 직접적인 언어적 강도 신호에 더 강하고 일관되게 반응할 수 있음을 시사한다. 강도 수식어는 통증강도를 직접적으로 나타내는 정보이므로 가장 큰 변화량을 보였으며, 부위・지속기간・발생양상・발생원인과 같은 비강도 맥락 요인에 대한 반응은 상대적으로 작고 표현별로 불균일하였다. 이러한 결과는 각 맥락정보에 대한 Claude의 반응 특성을 기술적으로 보여준다.

환자의 정도부사에 과도한 가중치가 부여될 경우 임상적으로도 문제가 발생할 수 있다. 예를 들어 동일한 임상상태에서도 습관적으로 강한 표현을 사용하는 환자의 증상이 과대평가되거나, 반대로 절제된 표현을 사용하는 환자의 증상이 과소평가될 수 있다. 이러한 차이가 진료 우선순위, 위험도 분류, 추가 검사 또는 치료 필요성 판단과 연결되는 경우 자연어 표현양식 자체가 임상의사결정에 불필요한 영향을 미칠 가능성이 있다. 따라서 실제 CDSS에서는 자유서술형 통증 표현으로부터 추론한 강도값을 환자가 직접 보고한 통증척도(numeric rating scale, NRS) 또는 시각아날로그척도(visual analog scale, VAS)와 동일하게 취급해서는 안 되며, 통증 기간・양상・발생원인 및 red-flag 정보와 같은 구조화된 임상정보를 별도로 반영하고, 모델의 추정 불확실성을 함께 표시할 필요가 있다.

찌르다와 찌릿하다의 대표 사례는 이러한 맥락 의존성이 표현별로 서로 다른 양상을 띤다는 점을 보여준다. 찌르다에서 관찰된 결과는 통증척도(NRS)나 시각아날로그척도(VAS)에서 보고되어 온 천장효과와 유사한 현상으로 해석할 수 있으나[15], 후속 검증이 필요하다.

찌릿하다에서 관찰된 부위 특이적 음(-)의 연상은 더

주목할 만하다. 대규모 언어모델의 벡터 표상은 학습 코퍼스 내 통계적 공기(co-occurrence) 패턴을 반영하며, 이러한 패턴이 인간의 암묵적 연상과 유사한 구조를 가진다는 것이 선행연구를 통해 입증된 바 있다[16]. 본 연구에서 찌릿하다가 허리 이외의 부위로 이동했을 때 평균적으로 강도가 감소하는 방향을 보인 것은, 이 표현이 모델의 내부 표상에서 특정 신체부위(요통 관련 문맥)와 결합되어 있고, 그 결합이 다른 부위 정보가 주어졌을 때 표현의 강도 해석을 오히려 약화시키는 방향으로 작동했을 가능성이 있다. 이는 Choi 등[13]이 지적한 것처럼 한국어 신경병증성 통증 표현('저림', '시림' 등)이 특정 임상 맥락과 밀접하게 결합되어 사용되는 한국어 통증 어휘의 특성과 맞닿아 있으며, 이러한 언어-맥락 결합이 사람과 모델 모두에서 나타나되 그 결합의 강도와 방향이 서로 다를 수 있음을 보여준다.

최근 한의학 진단 영역에서 자유텍스트 기반 증상 서술을 인공지능으로 수치화하거나 변증 지원에 활용하려는 시도가 이루어지고 있다[4,5]. 다만 본 연구에서는 한의 변증체계, 한의학 특이적 증상분류 또는 한의사 평가를 직접 적용하지 않았다. 따라서 본 결과는 한의 진단 또는 한의 임상지원시스템의 타당성을 직접 입증하는 근거가 아니라, 향후 그러한 시스템에서 자연어 환자정보를 활용하기 전에 검토해야 할 LLM의 언어 해석 특성에 관한 탐색적 자료로 해석되어야 한다. 실제 한의 임상 적용을 위해서는 한의사 전문가 평가, 한의학 특이적 증상체계 및 실제 환자자료를 이용한 별도의 검증이 선행되어야 한다.

본 연구의 결과는 이러한 시스템을 설계할 때 두 가지를 함께 고려해야 함을 시사한다. 첫째, 강도 수식어와 같이 명시적인 언어 신호가 존재할 경우 이를 적절히 반영하는 것은 바람직하나, 그 신호가 다른 임상 정보(부위・기간・원인)를 충분히 함께 고려하지 않은 채 사용될 경우, 환자의 표현 습관에 따라 통증 중증도가 상대적으로 과대 또는 과소 추정될 가능성이 있다. 둘째, 표현마다 맥락 민감도가 상이하므로, 모든 증상 표현에 동일한 신뢰도로 자동화된 해석을 적용하는 것은 위험할 수 있으며[3], 환자-임상의 간 통증 의사소통의 복잡성을 고려할 때[9] 이러한 시스템은 임상의의 최종 판단을 보조하는 도구로 제한적으로 활용되어야 한다는 점이다.

본 연구는 몇 가지 중요한 한계를 지닌다. 첫째, 본 연구는 단일 모델(Claude Sonnet)에서 관찰된 결과이며, 다른 LLM (GPT 등)에 일반화될 수 있는지는 확인되지 않았다. 특히 본 연구에서 관찰된 강도 수식어 우세, 부위 특이적 연상 및 표현별 맥락 민감도가 Claude에 특이적인 특성인지, 서로 다른 LLM에서 재현되는 공통적 현상인지는 확인되지 않았다. 따라서 동일한 자극문장과 프롬프트를 GPT, Gemini 등 복수의 LLM에 적용하는 외적 검증이 필요하다. 둘째, 본 연구에서 사용한 혼합효과모형은 요인마다 독립적으로 적합되었으며, 강도수식어(E) 요인은 전체 임의효과 구조가 수렴하지 않아 임의절편만 포함한 축소모형으로 대체하였다. 이에 따라 E 요인의 잔차분산은 다른 요인과 직접 비교할 수 없으며, 요인 간 임의기울기 분산을 통합적으로 비교하는 데에는 한계가 있다. 셋째, 본 연구에서 탐색된 표현-부위 조합의 일부(예: '복부가 뻣뻣하다')는 임상적으로 드문 조합일 수 있어 결과 해석에 주의가 필요하다. 넷째, 본 연구는 온도(temperature)=1로 수행되었는데, 이는 확률적 생성 조건에서의 반복 안정성 평가라는 이차 목적을 위한 것으로, 결정론적 조건(temperature=0)에서의 결과와 다를 수 있다. 또한 본 연구에서 사용한 12개 통증 표현은 한국 근골격계 통증 환자의 실제 사용 빈도에 관한 선행연구와 임상적 사용 가능성을 바탕으로 선정하였으나, 국어학 전문가 또는 다학제 전문가 패널을 통한 형식적인 내용타당도 검증은 시행하지 않았다. 향후 연구에서는 국어학・한의학・통증의학 전문가의 합의 과정과 내용타당도 평가를 통해 자극어의 대표성과 언어적 자연성을 검증할 필요가 있다.

본 연구에서 확인된 패턴을 확증하기 위해서는 복수 LLM 간 비교, 그리고 전문가 패널을 통한 내용타당도 평가가 필요하다. 특히 찌릿하다에서 관찰된 부위 특이적 연상과 같은 현상은 다른 신경병증성 표현으로 확장하여 체계적으로 검증할 필요가 있다.

결론»»»

본 연구는 Claude가 한국어 통증 표현 12종에 부여하는 통증 강도 추정값이 표현 고유의 의미뿐 아니라 부

위・지속기간・발생양상・발생원인・강도 수식어라는 다섯 가지 맥락 요인에 의해 서로 다른 크기로 변화함을 확인하였다. 강도 수식어는 positive control로서 기대된 단조순서를 모든 표현에서 충족시켜 척도 및 프롬프트 설계의 타당성을 뒷받침하였으며, 다섯 요인 중 가장 큰 변화량을 보였다. 비강도 맥락 요인의 영향은 상대적으로 작고 표현별로 상이하였다.

표현별로 부위 특이적 연상과 기준강도에 따른 비대칭적 반응 양상으로 의심되는 결과를 확인하였다. 이는 맥락 민감도가 모든 통증 표현에 균일하게 나타나는 특성이 아니라 표현 개별적으로 결정되는 특성임을 보여준다. 이러한 결과는 향후 자연어 기반 생성형 인공지능을 한의 진료기록이나 임상의사결정지원시스템에 활용하기 위한 기초적 탐색자료를 제공한다. 다만 본 연구에서 한의학 특이적 진단체계나 실제 CDSS를 직접 평가한 것은 아니므로, 실제 적용에 앞서 한의사 전문가 평가, 실제 임상자료 및 복수 모델을 이용한 별도의 검증이 선행되어야 한다.

References»»»»

1. Artsi Y, Sorin V, Glicksberg BS, Korfiatis P, Nadkarni GN, Klang E. Large language models in real-world clinical workflows: a systematic review of applications and implementation. Frontiers in Digital Health. 2025;7:1659134.
2. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, Vielhauer J, Makowski M, Braren R, Kaissis G, Rueckert D. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine. 2024;30(9):2613-22.
3. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. npj Digital Medicine. 2020;3:17.
4. Jang H, Lee SH, Yea S. Korean medicine symptom recommendation based on textual description of the patient's condition (digestive disorder version). Journal of Knowledge Information Technology and Systems. 2020;15(5):831-43.
5. Yea S, Lee S, Jang H. Process design and system implementation for building machine learning data to suggest prescriptions by Korean medicine school. Journal of Knowledge Information Technology and Systems. 2021;16(5):1091-102.
6. Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D, Goel S. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences of the United States of America. 2020;117(14):7684-9.
7. Zhou L, Blackley SV, Kowalski L, Doan R, Acker WW, Landman AB, Kontrient E, Mack D, Meteer M, Bates DW, Goss FR. Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. JAMA Network Open. 2018;1(3):e180530.
8. Park GY, Kwon DR, Woo IH. Korean pain descriptors in patients with neuromusculoskeletal pain. Clinical Pain. 2019;18(2):82-7.
9. Henry SG, Matthias MS. Patient-clinician communication about pain: a conceptual model and narrative review. Pain Medicine. 2018;19(11):2154-65.
10. Gotlieb R, Praska C, Hendrickson MA, Marmet J, Charpentier V, Hause E, Allen KA, Lunos S, Pitt MB. Accuracy in patient understanding of common medical phrases. JAMA Network Open. 2022;5(11):e2242972.
11. Wang L, Chen X, Deng X, Wen H, You M, Liu W, Li Q, Li J. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. npj Digital Medicine. 2024;7(1):41.
12. Sharma S, Pathak A, Jensen MP. Words that describe chronic musculoskeletal pain: implications for assessing pain quality across cultures. Journal of Pain Research. 2016;9:1057-66.
13. Choi K, Kwon O, Suh BC, Oh J, Cho S, Sohn E, Joo IS. Characteristics of diverse verbal pain descriptors in South Korean patients with peripheral neuropathic pain: 'Jeorim' (tingling) and 'Sirim' (cold) as key neuropathic pain descriptors. Journal of Clinical Neurology. 2023;19(3):296-303.
14. Amidei J, Nieto R, Kaltenbrunner A, Ferreira De Sá JG, Serrat M, Albajes K. Exploring the capacity of large language models to assess the chronic pain experience: algorithm development and validation. Journal of Medical Internet Research. 2025;27:e65903.
15. Reed MD, Van Nostran W. Assessing pain intensity with the visual analog scale: a plea for uniformity. The Journal of Clinical Pharmacology. 2014;54(3):241-4.
16. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356(6334):183-6.