JAMIA Open, 2026, **9**(2), ooag058
https://doi.org/10.1093/jamiaopen/ooag058
Research and Applications

Large language models for automated and audience-tailored labeling of latent classes

Fatemeh Gholi Zadeh Kharrat, PhD[1,2,*], Rob Werfelmann, MS[3], Glen Ep Ropella, BS[3], Wolf Mehling, MD[4], C. Anthony Hunt, PhD[5], Jeffrey Lotz, PhD[1], Thomas A. Peterson, PhD[*,1,2]; and the REACH Investigators[†]

[1]Department of Orthopaedic Surgery, University of California, San Francisco, CA 94143, United States
[2]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA 94158, United States
[3]Tempus Dictum, Inc, Olympia, WA 98502, United States
[4]Department of Family and Community Medicine, Osher Center for Integrative Health, University of California, San Francisco, CA 94115, United States
[5]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94143, United States
*Corresponding authors: Fatemeh Gholi Zadeh Kharrat, PhD, Department of Orthopaedic Surgery, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA 94143, United States (fatemeh.gholizadehkharrat@ucsf.edu); Thomas A. Peterson, PhD, Department of Orthopaedic Surgery, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA 94143, United States (Thomas.Peterson@ucsf.edu)
[†]**REACH Investigators:** Research reported in this publication was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award Number U19AR076737. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Core Center of Patient-centric, Mechanistic Phenotyping in Chronic Low Back (REACH) investigators include the following University of California, San Francisco (unless noted otherwise) personnel in alphabetical order: Zehra Akkaya, PhD, Prakruthi Amar Kumar, PhD, Jeannie Bailey, PhD, Julia Barylak, Sigurd Berven, MD, Andrew Bishara, MD, Dennis M. Black, PhD, Noah Bonnheim, PhD, Atul Butte, MD, PhD, Joel Castellanos, MD (University of California, San Diego), Jennifer Cummings, Karina Del Rosario, MD, Emilia Demarchis, MD, Sibel Demir-Deviren, MD, Susan K. Ewing, MS, Adam R. Ferguson, PhD, Aaron Fields, PhD, Scott M. Fishman, MD (University of California, Davis), Sergio Garcia Guerra, Fatemeh Gholi Zadeh Kharrat, PhD, Xiaojie (Summer) Guo, Misung Han, PhD, Trisha Hue, PhD, J. Russell Huie, PhD, C. Anthony Hunt, PhD, Anastasia Keller, PhD, Karim Khattab, Roland Krug, PhD, Gregorji Kurillo, PhD, Feng Lin, Thomas Link, MD, PhD, Jeffrey Lotz, PhD, John Lynch, PhDTong Lyu, Rob Matthew, PhD, Wolf Mehling, MD, Esmeralda Mendoza, MPH, Praveen Mummaneni, MD, MBA, Caroline Navy, Conor O'Neill, MD, Jessica Ornowski, Thomas Peterson, PhD, Ananya Rupanagunta (University of California, Berkeley), Aaron Scheffler, PhD, MS, Shalini Shah, MD (University of California, Irvine), Irina Strigo, PhD, Naoki Takegami, MD, Abel Torres-Espin, PhD (University of Waterloo), Salvatore Torrisi, PhD, Sachin Umrao, PhD, Rohit Vashisht, PhD, Joanna Veres, An (Joseph) Vu, PhD, Mark Steven Wallace, MD (University of California, San Diego), Lucy Ann Wu, MPH, Po-Hung Wu, PhD, Fadel Zeidan, PhD (University of California, San Diego), Patricia Zheng, MD, Jiamin Zhou, MS

Abstract

Objective: **This study compares multiple LLMs, including ChatGPT, DeepSeek, and Llama, to generate meaningful, audience-adapted labels for the existing latent classes among patients with chronic low back pain (cLBP).**

Methods: **Phenotypes were derived from baseline data from two cohorts within the NIH HEAL BACPAC consortium: BACKHOME, a large nationwide e-cohort (train set: $N = 3025$), and COMEBACK, a deep phenotyping cohort (test set: $N = 450$). The analysis included pain characteristics, psychosocial factors, lifestyle habits, and social determinants of health. ChatGPT-4o (OpenAI), DeepSeek-R1, and Llama 3.3 (Meta) were applied to generate class labels for each combination of audience (clinician, patient, and caregiver), tone (formal, empathetic, and informal), and technicality (high, medium, and low).**

Results: **Latent Class Model (LCM) identified four distinct behavioral phenotypes in patients with cLBP: *High Distress and Maladaptive Behaviors*, *Resilient and Adaptive Coping*, *Intermediate Maladaptive Patterns*, and *Emotionally Regulated with High Pain Burden*. Previously validated by domain experts, these profiles served as the basis for automated labeling using three LLMs (ChatGPT-4o, DeepSeek-R1, and Llama 3.3). Using different tones and complexity levels, each model produced class labels specific to clinicians, patients, and caregivers. The generated class names for all LLMs closely matched expert-defined traits like *emotional regulation*, *resilience*, and *high distress*, indicating strong conceptual alignment and the capacity of LLMs to generate precise, audience-specific labels for intricate behavioral and psychological profiles.**

Conclusions: **These results highlight the possibility of integrating LLM-driven labeling into research and clinical practice, helping to achieve more transparent knowledge translation, improved decision-making, and personalized care.**

Received: September 27, 2025. **Revised:** February 24, 2026. **Accepted:** March 26, 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of the American Medical Informatics Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Key words: Large language models (LLMs), retrieval-augmented generation (RAG), orthopedic surgery; artificial intelligence (AI), LLaMA-3, LangChain, chronic low back pain (cLBP), Latent Class Model (LCM)

Lay Summary

People with chronic low back pain experience symptoms in different ways. Some struggle with high emotional distress and difficulty coping, while others show resilience despite ongoing pain. The Latent Class Modeling (LCM) approach is used to identify these hidden subgroups based on pain levels, mood, coping behaviors, and lifestyle factors. However, once these groups are identified, naming them clearly and appropriately for different audiences can be difficult. In this study, we considered whether large language models (LLMs) could automatically generate meaningful, audience-specific names for these subgroups. We tested three models, such as ChatGPT, DeepSeek, and Llama 3, using data from two large studies of individuals with chronic low back pain. The models were asked to create labels tailored to clinicians, patients, and caregivers, using different tones and levels of technical language. All three systems created labels that closely reflected expert-defined group characteristics. The wording was appropriately adjusted to suit the audience, making labels more technical for clinicians and more supportive and accessible for patients and caregivers. These results imply that LLMs can help simplify complex research findings into language understandable to a broad audience, enhancing communication and facilitating more personalized care.

Introduction

Latent Class Model (LCM) is a statistical method that can be applied to various fields, including education, social sciences, and healthcare.[1–3] LCM contains both Latent Class Analysis (LCA) and Latent Profile Analysis (LPA). It includes a mix of binary and continuous variables and techniques that aim to recover hidden groups from observed data.[2,4–6] LCM models use observed indicators to infer these latent groups and determine their characteristic variable patterns, assuming that the observed variable distribution results from a finite mixture of underlying, unobserved subgroups.[2] For example, Obbarius et al[7] found four subgroups based on pain burden and emotional distress in patients undergoing multimodal inpatient treatment. Rovner et al[8] determined four patient clusters based on pain acceptance levels, which were associated with specific patterns of psychological and physical function.

However, naming LCM-identified subgroups or clusters is a crucial step in making the findings interpretable and actionable. This often requires time consuming expert consensus, particularly when the LCM variables span a broad range of technical and clinical disciplines. Additionally, cluster names should be customized to the audience, as names suitable for clinicians may include language or jargon that is not interpretable or can even be offensive to patients.

Artificial intelligence (AI) has rapidly appeared as a transformative technology across numerous domains, including healthcare.[5,9–12] Large language models (LLMs), which employ transformer architectures and encode rich contextual information using billions of parameters, are one example of these developments that have made it possible to understand and apply varied topics with nuance.[13,14] Because LLMs are better at interpreting unstructured text than traditional AI models, which frequently rely on structured data and preset algorithms, they are invaluable for real-time data interpretation, supporting clinical decision-making, and improving patient engagement in clinical settings. LLMs such as ChatGPT,[15] DeepSeek,[16] and Llama[17] can process and generate human-like language, making them valuable tools in various medical applications. ChatGPT, built on OpenAI's GPT architecture, is available in multiple configurations; for example, GPT-4o[18] contains approximately 200 billion parameters, while GPT-4o-mini has around 8 billion parameters. DeepSeek is an open-source language model with up to 671 billion parameters and context lengths of up to 128K tokens. Llama 3[19] is available in different sizes, including 8 billion and 70 billion parameter versions.

LLMs are increasingly embraced for various healthcare applications, including orthopedic surgery, by enhancing diagnostic accuracy, streamlining administrative tasks, and improving patient care and education. For instance, Bains et al[20] investigated the utility of LLMs, specifically ChatGPT, in providing patient information after total knee arthroplasty (TKA) and compared its performance with that of arthroplasty-trained nurses. Lieu et al[21] examined the accuracy and comprehensibility of ChatGPT (Version 3.5) responses to frequently asked questions about scoliosis in a pediatric orthopedic context. Margetts et al[22] examined the role of AI in scientific writing related to Alzheimer's disease, osteoporosis, and fracture. Guillen et al[23] investigated the utility of LLMs (ChatGPT, Llama 3.3, Microsoft Copilot) for patient Queries on Total Knee Replacement (TKR). Caterson et al[24] demonstrated that LLMs could generate clinical letters and predict management plans for orthopedic scenarios. Sakai et al[25] used GPT-4 Turbo for multi-label classification of inpatient comments, testing prompt strategies like zero-shot, in-context, and chain-of-thought. GPT-4 Turbo outperformed traditional and pre-trained models, offering an efficient way for healthcare providers to analyze patient feedback and improve responses. Wang et al[26] showed that LLMs can be considered as implicitly inferring a hidden or "latent" variable carrying important task information, allowing for more accurate and contextually appropriate labeling, particularly in through in-context learning approaches.

LLMs and LCM aim to uncover meaningful patterns and structure within complex datasets, LLMs through contextually adaptive, language-based labeling, and LCM through statistical modeling of latent subgroups. In this current study, we combine the statistical rigor of LCM with the adaptive language capabilities of LLMs to enhance the interpretability and accessibility of

class labeling in health research. We build upon our previously established LCM[27] and leverage multiple LLMs (ChatGPT-4o, DeepSeek-R1, and Llama 3.3) to generate meaningful, audience-adapted labels (such as clinicians, patients, and caregivers) for these existing latent classes. Our results demonstrate the feasibility and advantages of integrating state-of-the-art LLMs with established statistical modeling, facilitating scientific rigor and effective communication in health research.

Methods

Phase 1: Latent class model development and characterization

We used LCM previously established in our earlier research within the NIH HEAL BACPAC consortium for this analysis. The BACKHOME,[28] a large nationwide e-cohort ($N = 3025$) utilized for model training, and COMEBACK,[28] as an external test set, a deep phenotyping cohort ($N = 450$) used for generalization. The datasets included a broad range of variables across several categories (for detailed descriptions, see Supplementary Appendix A): demographic factors (age, sex), anthropometrics (Body Mass Index [BMI]), pain-related measures (PEG score, PROMIS Pain Interference), pain characteristics (duration, frequency, intensity), and quality of life (PROMIS Physical Functioning [SF 6b]). Additional variables included mood indicators (PROMIS Depression, Pain Anxiety Symptoms [PAS]), pain persistence parameters (fear avoidance via FABQ-P, pain catastrophizing via PCS-6, pain self-efficacy via PSEQ-4), and elements of interoceptive awareness (Non-Distracting, Emotional Awareness, Self-Regulation) measured using the MAIA-2.

LCM was developed using model selection criteria that balance statistical fit, clinical interpretability, and parsimony. To determine the optimal number of classes, we employed 10-fold cross-validation with 100 bootstraps and evaluated model fit using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and entropy (uncertainty) to optimize model fit.[29] AIC and BIC were used to balance fit and complexity, while entropy (uncertainty) measured the separation of latent classes. Evaluation of models ranging from 1 to 10 classes revealed that AIC and BIC values declined steeply from 1 to 4 classes, then plateaued, indicating that additional classes beyond 4 did not substantially improve model fit relative to increased complexity. Entropy values of 0.82 for BACKHOME (train) and 0.81 for COMEBACK (test) indicated well-separated, highly distinct latent classes. Furthermore, the four-class solution demonstrated strong clinical interpretability and adequate sample sizes within each class (BACKHOME: Class 1, $n = 701$; Class 2, $n = 413$; Class 3, $n = 893$; Class 4, $n = 947$), ensuring stable parameter estimates and statistical power. The resulting model was robust and generalizable, as evidenced by consistent fit metrics and class structure when applied to the independent COMEBACK test set. A detailed description of these models and their development is available elsewhere.[27]

After model estimation, the class-specific profiles of normalized indicator mean for four latent classes were obtained for both the BACKHOME (train set) and COMEBACK (test set) cohorts (see Supplementary Appendix B). The LCM was generated from extensive behavioral and clinical datasets, and all model outputs, class profiles, and variable definitions were methodically saved in DOCX format for use as input in the following stage, which involved automated class labeling with LLMs.

Phase 2: Automated class labeling using large language models

Input preparation and standardization

This standardized input structure was provided identically to all three LLMs (ChatGPT-4o, DeepSeek-R1, and Llama 3.3) to ensure consistent information was provided to each model. Specifically, all three models received, identical variable definitions and scoring guidance, and (see Supplementary Appendix A), identical class profiles (see Supplementary Appendix B), and identical instructions specifying the audience, tone, and technicality parameters for label generation. Model-specific implementation parameters were optimized for each system's architecture and access method, but these differences do not impact the comparability of outputs.

The standardized input and consistent temperature = 0.0 setting across all models (ChatGPT-4o, DeepSeek-R1, and Llama 3.3) means that any differences in the generated labels can be attributed to the models' underlying architectures, training data, and inherent stylistic preferences, rather than to discrepancies in the input information or parameter settings.

Chain-of-thought prompting with ChatGPT and DeepSeek for automated latent

Chain-of-Thought (CoT)[30] prompting is a natural language prompt engineering strategy used with LLMs to enhance their reasoning capabilities. Rather than simply instructing the model to provide an answer, CoT prompting explicitly guides the model to break down its reasoning into logical, interpretable steps before arriving at a final output. By encouraging the model to reveal its intermediary thought processes, this method can enhance its output's caliber, interpretability, and dependability.[31,32] For generating audience-tailored clinical labels from complex latent classes, where labels must simultaneously serve multiple stakeholders (clinicians, patients, caregivers) and satisfy competing requirements (clinical accuracy, accessibility, consistency), CoT offers different advantages. By requiring systematic reasoning about latent class features before label generation, CoT ensures that outputs are interpretable and justifiable to clinical users, logically consistent across audience contexts, and semantically accurate and clinically relevant. We therefore selected CoT prompting to conduct ChatGPT-4o and DeepSeek-R1 in generating audience-tailored class labels and asked the LLMs to generate label with justifications, outputting results in a standardized JSON format.

Automated latent class labeling with llama 3.3

Prompt formatting and response parsing were implemented using LangChain's PromptTemplate and StrOutputParser.[33] The Llama 3.3 model was accessed via the OllamaLLM interface and generated labels in a standardized JSON format.

Phase 3: Model execution, output, and evaluation

Each output included narrative descriptions summarizing class profile distributions and a set of candidate class labels generated by the LLMs for each latent class. The experimental design tailored the suggested class names to different audiences, tones, and technicality levels. Experts thoroughly examined each set of narrative descriptions and class labels for interpretability, contextual appropriateness, and clarity in clinical practice and health communication. This assessment considered not only the accuracy and relevance of the labels, but also their potential utility for diverse stakeholders, including clinicians, patients, and caregivers.

Overall, this approach was intended to enhance the translation of statistical findings into practical healthcare applications. The overall workflow for generating audience-specific labels from latent-class profiles is shown in Figure 1. This four-step process begins with LCM, proceeds by providing identical inputs to all three LLMs, continues with parallel processing using prompting strategies, and concludes with the generation of labeled phenotypes. Figure 2 presents a representative example of this workflow for a single latent class, illustrating the prompting approaches, model reasoning processes, and the resulting clinical labels and justifications.

Results

Sample characteristics

A descriptive summary of the demographic characteristics of the BACKHOME and COMEBACK datasets, respectively, is shown in Table 1. The demographic features of the research population for the BACKHOME and COMEBACK data sets, respectively, are summarized in Table 1. In the BACKHOME dataset, 92.3% of individuals identified as non-Hispanic or Latino; the majority were White (85.8%) and female (68.1%). Participants in the COMEBACK dataset were also White (75.7%) and female (57.1%). There was a modest difference in age between the BACKHOME and COMEBACK participants. The BACKHOME participants had a higher mean BMI than in COMEBACK.

Latent class model profiles

Our previous paper[27] constructed and detailed the latent class model profiles (see Supplementary Appendix B). Domain experts defined and categorized four classes in that work: Class 1 (“*High Distress and Maladaptive Behaviors*”), Class 2 (“*Resilient and Adaptive Coping*”), Class 3 (“*Intermediate Maladaptive Patterns*”), and Class 4 (“*Emotionally Regulated with High Pain Burden*”). Class 1 was indicated by the highest levels of anxiety/depression, fear avoidance, PCS-6, and PEG scores, indicating a greater propensity for pain and maladaptive avoidance behaviors. In contrast, Class 2 demonstrated lower scores on maladaptive measures, suggesting greater pain self-efficacy and fewer challenges with pain management. Differences across MAIA scales contributed less to class separation, although distinctions between Classes 1 and 2 were evident. Classes 3 and 4 showed intermediate means across several indicators, representing a spectrum of psychological profiles within the population. Class 4, compared to Class 3, presented with higher pain-related scores (PEG), higher fear avoidance, but showed lower emotional awareness and self-regulation. These class profiles define distinct behavioral phenotypes and form the foundation for the automated label generation process with LLMs utilized in the current study.

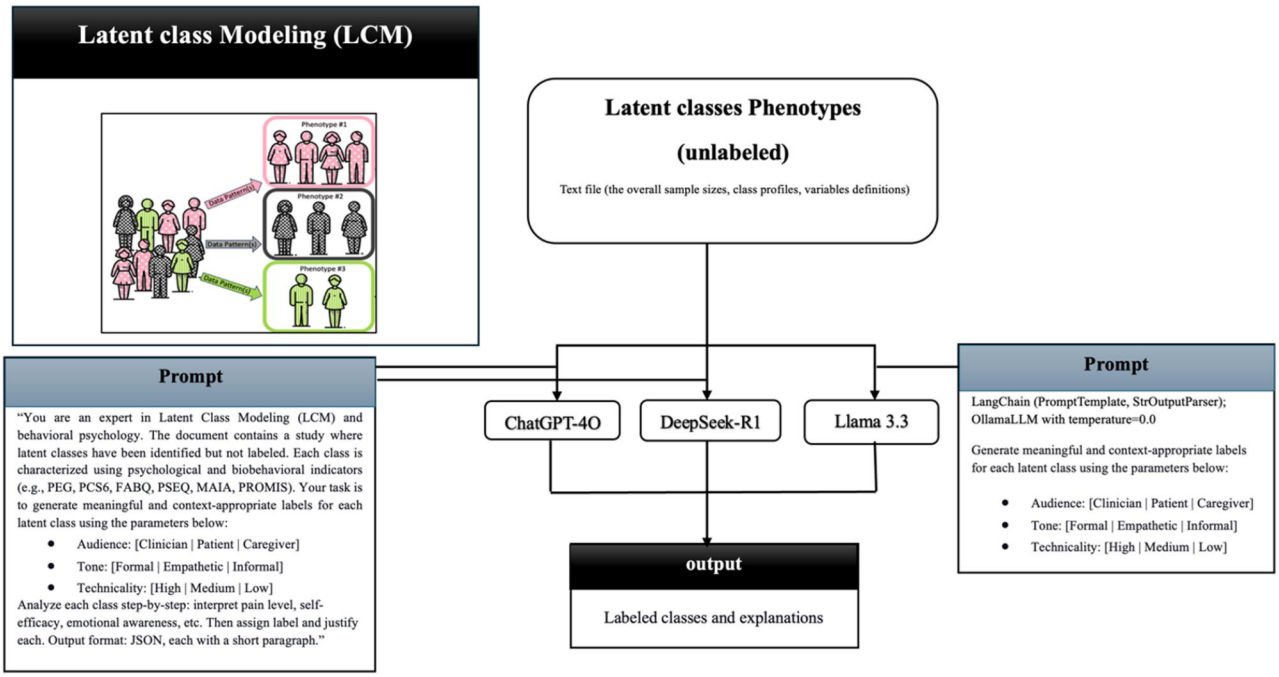


Figure 1 Schematic representation of the LLMs model.

USER-PROVIDED INPUTS & EXAMPLE PARAMETERS

==========

Class Profile (Class 1): PEG=0.70 (pain interference), PSEQ=0.39 (pain self-efficacy), PCS6=0.64 (catastrophizing), FABQ=0.73 (fear-avoidance), PROMIS Anxiety=0.54, MAIA Emotional Awareness=0.67, MAIA Not Distracting=0.30, MAIA Self-Regulation=0.49, Pain Duration=0.16

Example Parameters: Audience: Clinician | Tone: Formal | Technicality: High | Temperature: 0.0

ChatGPT-4o & DeepSeek-R1 **CHAIN-OF-THOUGHT**	**Llama 3.3** **DIRECT GENERATION**
ACTUAL PROMPT: You are an expert in Latent Class Modeling (LCM) and behavioral psychology. The document contains a study where latent classes have been identified but not labeled. Each class is characterized using psychological and biobehavioral indicators (e.g., PEG, PCS6, FABQ, PSEQ, MAIA, PROMIS). Your task is to generate meaningful and context-appropriate labels for each latent class using the parameters: Audience: [Clinician | Patient | Caregiver], Tone: [Formal | Empathetic | Informal], Technicality: [High | Medium | Low]. Analyze each class step-by-step: interpret pain level, self-efficacy, emotional awareness, etc. Then assign label and justify each. Output format: JSON, each with a short paragraph.	**ACTUAL PROMPT:** LangChain (PromptTemplate, StrOutputParser); Generate meaningful and context-appropriate labels for each latent class using the parameters: Audience: [Clinician | Patient | Caregiver], Tone: [Formal | Empathetic | Informal], Technicality: [High | Medium | Low]. Output format: JSON with label and justification.
CHAIN-OF-THOUGHT ANALYSIS: 1. PEG score is high (0.70), indicating significant daily disruption due to pain. 2. Pain self-efficacy is low (0.39), reflecting limited confidence in managing pain. 3. Catastrophizing (0.64) and fear-avoidance beliefs (0.73) are elevated, indicating maladaptive cognitive patterns. 4. Psychological distress (PROMIS = 0.54) is moderate. 5. Emotional awareness is relatively preserved (0.67), but low self-regulation (0.49) and distraction tolerance (0.30) suggest difficulties in emotional modulation. 6. Overall, the profile is consistent with a maladaptive coping phenotype with elevated psychosocial risk.	**DIRECT GENERATION:** Model generates label directly from class profile (PEG, PSEQ, PCS6, FABQ, MAIA, PROMIS scores) without explicit intermediate reasoning steps. Framework: LangChain PromptTemplate + StrOutputParser + OllamaLLM
OUTPUT (Clinician/Formal/High): `{` `  "label": "Maladaptive Pain Responders",` ` "justification": "This phenotype demonstrates significant pain-related functional impairment (PEG=0.70), markedly diminished pain self-efficacy (PSEQ=0.39), and maladaptive cognitive patterns including catastrophizing (PCS6=0.64) and pain-related fear-avoidance (FABQ=0.73). Poor emotional regulation capacity indicates heightened psychosocial vulnerability."` `}`	**OUTPUT (Clinician/Formal/High):** `{` `  "label": "Emotionally distressed patient",` ` "justification": "Characterized by high pain-related disability (PEG=0.70), limited pain management confidence (PSEQ=0.39), and maladaptive cognitive-behavioral patterns including pain catastrophizing (PCS6=0.64) and fear-avoidance beliefs (FABQ=0.73). Emotional regulation deficits suggest significant psychosocial risk."` `}`

Figure 2 Representative prompting strategies and structured input templates for generating audience-specific labels from latent class profiles.

Table 1 Socio-demographic characteristics.

Variables	BACKHOME (Train set) (*N* = 3025)	COMEBACK (Validation set) (*N* = 450)
Age		
Mean (SD)	56.5 (14.2)	55.6 (15.6)
Median [Min, Max]	58 [20,94]	58.5 [18,91]
Missing[a]	5 (0.16 %)	
Gender		
Female	2013 (68.1 %)	257 (57.1 %)
Male	941 (31.8 %)	193 (42.8 %)
BMI		
Mean (SD)	29.91 (7.5)	26.76 (5.3)
Median [Min, Max]	28.46 [13.3,71.3]	26.25 [7.8,43.9]
Missing[a]	13 (0.4 %)	12 (2.6 %)
Education		
Did not complete High school	6 (0.2 %)	3 (0.6 %)
Some Secondary school (or High School) education	47 (1.5 %)	4 (0.8 %)
High school complete	529 (17.4 %)	46 (10.2 %)
Associate's or Technical Degree complete	525 (17.3 %)	46 (10.2 %)
College or Baccalaureate Degree complete	948 (31.3 %)	190 (42.2 %)
Doctoral or postgraduate education	970 (32.0 %)	161 (35.7 %)
Race		
White	2596 (85.8 %)	341 (75.7 %)
Asian	87 (2.8 %)	45 (10.0 %)
Black or African American	142 (4.6 %)	28 (6.2 %)
Native Hawaiian or Pacific Islander	6 (0.20 %)	4 (0.8 %)
American Indian/Alaska Native	20 (0.6 %)	2 (0.4 %)
More than one race	112 (3.7 %)	12 (2.6 %)
Unknown or not reported	62 (2.05 %)	18 (4.0 %)
Ethnicity		
Not Hispanic or Latino	2793 (92.3 %)	392 (87.1 %)
Hispanic or Latino	185 (6.1 %)	51 (11.3 %)
Prefer not to answer or Unknown	47 (1.5 %)	7 (1.5 %)

[a]Observation with missing values were excluded from the analysis.

Automated audience-adaptive class labeling with LLMs

We used three sophisticated LLMs, ChatGPT-4o, DeepSeek-R1, and Llama 3.3, to automate the tagging of latent classes using a CoT prompting method, building on previously defined latent class profiles. Standardized class profiles, thorough explanations of all psychological and bio-behavioral variables, and clear instructions to create labels with a specific tone and level of sophistication for three audiences: clinicians, patients, and caregivers, were given to each model. The ChatGPT findings are shown in Table 2A. ChatGPT employed a formal tone and high technicality for clinicians, yielding labels such as "High Pain-Maladaptive Coping" and "Resilient-Adaptive Copers," which focused on behavioral patterns and diagnostic accuracy. On the other hand, patient-focused labels were more empathetic and accessible, including "Stuck in Pain Cycle" and "Managing Well." ChatGPT generated informal and encouraging labels for caregivers, such as "Stuck in Safety Mode" and "Doing Pretty Well," emphasizing consolation and useful coping.

The DeepSeek-generated class labels are displayed in Table 2B. For clinicians' labels, such as "Maladaptive Pain Responders" and "Resilient Copers," also prioritized clinical clarity and diagnostic language. Patient-focused labels, such as "Feeling Overwhelmed by Pain" and "Staying Strong with Pain," captured the emotional and experiential aspects of the latent profiles. In contrast, caregiver labels like "Pain and Worry Mix" and "Doing Well" were crafted to acknowledge the challenges and resilience observed by those in a supportive role. Table 2C shows Llama 3.3 audience-specific labeling performed using LangChain's PromptTemplate and StrOutputParser. For clinicians, Llama 3.3 provided formal, high-technicality descriptors such as "Emotionally Distressed Patient" and "Resilient Pain Manager." Patient-focused outputs, like "Challenged by Pain" and "Resilient Copers," emphasized personal experience and emotional nuance. Labels for caregivers, such as "High Distress Patient" and "Steady Progress Patient," combined a recognition of emotional burden with an optimistic view on adaptation and progress.

Even though the inputs were the same, the three LLMs showed different stylistic trends. Llama highlighted emotional regulation and resilience across all classes, even when describing high-distress presentations ("Emotionally Distressed Patient" vs other models' "Maladaptive Pain Responders").

Table 2 Audience-adapted latent class labels generated by (A) ChatGPT-4o, (B), DeepSeek-R1 and (C) Llama 3.3.

Audience	Tone	Technicality	Class1	Class2	Class3	Class4
(A)						
Clinicians	Formal	High	Maladaptive Pain Responders	Resilient Copers	Transitional Responders	Pain-Enduring Copers
Patients	Empathetic	Low	Feeling Overwhelmed by Pain	Staying Strong with Pain	Managing but Sometimes Struggling	Coping Even with High Pain
Caregivers	Informal	Low	Pain and Worry Mix	Doing Well	Finding Their Way	Gets Through with Grit
(B)						
Clinicians	Formal	High	High Pain-Maladaptive Coping	Resilient Adaptive Copers	Moderate Pain Intermediate Coping	High Pain Emotionally Aware
Patients	Empathetic	Low	Stuck in Pain Cycle	Managing Well	Mixed Feelings	Pain with Understanding
Caregivers	Informal	Low	Stuck in Safety Mode	Doing Pretty Well	Middle Ground	Thinks It Through
(C)						
Clinicians	Formal	High	Emotionally distressed patient	Resilient Pain Manager	Moderate Challenged Individual	Adapting Patient with Resilience
Patients	Empathetic	Low	Challenged by Pain	Resilient Copers	Balanced Mangers	Emotionally Aware Copers
Caregivers	Informal	Low	High Distress Patient	Resilient Pain Manager	Balanced Coping Patient	Steady Progress Patient

DeepSeek prioritized emotional experience and coping persistence, incorporating emotionally resonant language ("Feeling Overwhelmed by Pain," "Pain and Worry Mix"). ChatGPT made the most granular severity distinctions, using transitional language to capture the range of presentations ("Managing but Sometimes Struggling"). These stylistic variations may be significant when choosing models for specific clinical settings or organizational communication styles, as they reflect different model architectures and training data.

Across all LLMs, the generated class names reliably reflected the core psychological and behavioral characteristics identified by expert manual labeling, such as high distress and maladaptive behaviors, resilience and adaptive coping, or emotional regulation with high pain burden. This conceptual alignment shows how well LLMs interpret latent class profiles and produce audience-specific yet semantically accurate labels.

Discussion

In this study, we compared three advanced large language models (LLMs), ChatGPT-4o, DeepSeek-R1, and Llama 3.3, for automating the labeling of latent classes using a CoT prompting strategy. Our approach was grounded in previously established latent class profiles,[27] in which domain experts had identified and categorized four distinct classes: Class 1 ("*High Distress and Maladaptive Behaviors*"), Class 2 ("*Resilient and Adaptive Coping*"), Class 3 ("*Intermediate Maladaptive Patterns*"), and Class 4 ("*Emotionally Regulated with High Pain Burden*"). The models were then prompted to generate class labels while systematically considering the intended audience (clinician, patient, or caregiver), the suitable tone (formal, empathetic, or informal), and the selected level of technicality (high, medium, or low). Analysis of the LLM-generated labels revealed consistent audience adaptation and contextual nuance (Table 2). For clinicians, all models used a formal tone and high technicality, yielding labels that emphasized diagnostic precision and behavioral patterns. For instance, ChatGPT labeled Class 1 as "High Pain-Maladaptive Coping," while DeepSeek and Llama generated "Maladaptive Pain Responders" and "Emotionally Distressed Patient," respectively. For patients, each model adopted an empathetic tone with low technicality, resulting in more accessible, supportive labels such as "Stuck in Pain Cycle" (ChatGPT), "Feeling Overwhelmed by Pain" (DeepSeek), and "Challenged by Pain" (Llama) for Class 1. For caregivers, labels were informal and simple, often reflecting practical concerns or encouragement, such as "Stuck in Safety Mode" (ChatGPT), "Pain and Worry Mix" (DeepSeek), and "High Distress Patient" (Llama). It was also clear that the models differed both stylistically and substantively: Llama frequently focused on emotional regulation and resilience, DeepSeek emphasized coping persistence, and ChatGPT outputs tended to provide more thorough gradations between moderate. There was significant conceptual overlap between the labels produced by LLM and the manual expert labels that had previously been developed.[27] Recent advances in the field have demonstrated the unique capabilities of LLMs for annotation and interpretability tasks. For instance, Törnberg et al[34] provided compelling evidence that LLMs, particularly GPT-4, substantially enhance text annotation in the social

sciences, outperforming both traditional methods and human coders in accuracy, and enabling more nuanced, context-sensitive analysis. Nahum et al[35] highlighted the promise of LLMs for improving annotation reliability, including using ensemble approaches ("LLM-as-a-judge") to flag potential label errors in existing datasets systematically. Building on these findings, our work shows that LLMs can also be effectively leveraged for automated class labeling in the health and behavioral sciences. According to the thematic and content analysis results, the LLM pipeline systematically assigned class labels to key bio—behavioral features for every kind of audience. The CoT approach increased transparency since most labels included clear, understandable explanations of the reasoning behind each assignment. This promotes the practical conversion of statistical findings into valuable insights for end users maintaining scientific rigor.

This study has several strengths. First, the LLM-based labeling approach is flexible enough to accommodate changing data landscapes since labels can be produced quickly for any new class solution or dataset without recurring expert consensus. Due to the high degree of adaptation of the framework, labels can be consistently and reproducibly adapted to different audiences and communication contexts using standardized output formats. The method also makes it easier to update and improve label sets in response to new clinical or scientific requirements by enabling quick prototyping and iterative development. Other significant advantages of CoT prompting are transparency and interpretability, which enable more candid conversations with stakeholders and let users comprehend the rationale behind each label. Further expanding the usefulness of latent class modeling, the LLMs' structured, machine-readable outputs facilitate smooth integration with clinical decision support systems, patient education resources, and digital health platforms. However, this study also has limitations. Due to the biases in their training data, LLMs may not produce results that accurately reflect all patient populations' contextual or cultural nuances. Importantly, this study did not include direct feedback from end-users (clinicians, patients, or caregivers) on the interpretability and utility of the generated labels, a key area for future research.

Further research is needed to test these findings in different clinical settings, assess user acceptance and comprehension, and enhance prompt engineering methods to maximize contextual fit and impact. Direct engagement with end-users is essential before broader clinical deployment. Future work should include cognitive interviews with patients to evaluate label clarity and personal relevance, focus groups with clinicians (orthopedic surgeons, pain specialists, physicians) to evaluate diagnostic accuracy and clinical utility, and brief implementation studies in clinical settings to assess real-world adoption and impact on clinical decision-making. For more generalizability, the framework must be expanded to encompass non-English speaking populations and a greater range of healthcare scenarios. LLM-generated labels' contextual sensitivity and clarity will also need further improvement through ongoing model updates and prompt design enhancements. Finally, advancing clinical decision assistance and tailored communication at scale may be possible by integrating LLMs' labeling pipelines into digital health infrastructures.

Conclusion

This study shows that LLMs can efficiently and effectively generate interpretable, audience-specific labels for clinicians, patients, and caregivers for latent class analysis in health and behavioral research. Our approach bridges the gap between statistical modeling and practical, stakeholder-centered communication. Embracing LLM-based labeling frameworks improves the clarity, consistency, and scalability of latent class analysis while decreasing time to obtain expert consensus. These results highlight the possibility of integrating LLM-driven labeling into research and clinical practice, helping to achieve more transparent knowledge translation, improved decision-making, and personalized care.

Acknowledgments

The REACH Investigators would like to express their gratitude to the members of the comeBACK clinical site research team, especially the Clinical Research Coordinators (CRCs) for their commitment and contributions toward the successful conduct of the study (in alphabetical order): Jamie Ahn[3], Kristina Benirschke[1], Alexandra Bryson[1], Katherine Bunda[4], Briana Davis[1], Carolina Dorofeyev[2], Rosalee Espiritu[4], Pirooz Fereydouni[1], Aamna Haq[1], Nicholas Harris[1], Sara Honardoost[3], Gabriel Johnson[1], Jennifer Johnson[1], Edward Lingayo, Jr[2], Robert Miller[3], Phirum Nguyen[4], Christopher Orozco[1], Lindsay Ruiz-Graham[2], Kie Shidara[1], Kaitlyn Smith[1], John (Boyuan) Xiao[1], Michelle Yang[1]. **CRC Affiliations:** [1]University of California, San Francisco; [2]University of California, Davis; [3]University of California, Irvine; [4]University of California, San Diego.

Author contributions

Fatemeh Gholi Zadeh Kharrat (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing), Rob Werfelmann (Formal analysis, Methodology, Validation, Writing—review & editing), Glen ep Ropella (Conceptualization, Investigation, Methodology, Validation, Writing—review & editing), Wolf Mehling (Conceptualization, Data curation, Methodology, Validation, Writing—review & editing), C. Anthony Hunt (Conceptualization, Methodology, Writing—review & editing), Jeffrey C. Lotz (Funding acquisition, Project administration, Resources, Writing—review & editing), and Thomas A. Peterson (Conceptualization, Data curation, Funding acquisition, Methodology, Resources, Supervision, Writing—review & editing)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Conflicts of interest

All authors declare to have no conflict of interests.

Funding

The research reported in this publication was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health (Award number U19AR076737). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data availability

The data that support the findings of this study are available in the Dryad Digital Repository (https://doi.org/10.5061/dryad.1jwstqk9d). The original datasets are available through the Vivli repository (https://doi.org/10.25934/PR00010820 and https://doi.org/10.25934/PR00010819) and require approval for access.

References

1. Weller BE, Bowen NK, Faubert SJ. Latent class analysis: a guide to best practice. *J Black Psychol*. 2020;46:287–311.
2. lSinha P, Calfee CS, Delucchi KL. Practitioner's guide to latent class analysis: methodological considerations and common pitfalls. *Crit Care Med*. 2021;49(1):e63–79.
3. Bacher J, Vermunt JK. Analyse latenter klassen. Handbuch der sozialwissenschaftlichen Datenanalyse. 2010;553–74.
4. Porcu M, Giambona F. Introduction to latent class analysis with applications. J Early Adolesc. 2017;37(1):129–58.
5. Gholi Zadeh Kharrat F, Shydeo Brandão Miyoshi N, Cobre J, Mazzoncini De Azevedo-Marques J, Mazzoncini de Azevedo-Marques P, Cláudio Botazzo Delbem A. Feature sensitivity criterion-based sampling strategy from the Optimization based on Phylogram Analysis (Fs-OPA) and Cox regression applied to mental disorder datasets. PLoS One. 2020;15(7):e0235147.
6. Kuo JC, Chan W, Leon-Novelo L, Lairson DR, Brown A, Fujimoto K. Latent classification model for censored longitudinal binary outcome. Stat Med. 2024;43(20):3943–57.
7. Obbarius A, Fischer F, Liegl G, Obbarius N, van Bebber J, Hofmann T, et al. A step towards a better understanding of pain phenotypes: latent class analysis in chronic pain patients receiving multimodal inpatient treatment. J Pain Res. 2020;1023–38.
8. Rovner G, Vowles KE, Gerdle B, Gillanders D. Latent class analysis of the short and long forms of the chronic pain acceptance questionnaire: further examination of patient subgroups. J Pain. 2015;16(11):1095–105.
9. Gholi Zadeh Kharrat F, Gagne C, Lesage A, Gariépy G, Pelletier JF, Brousseau-Paradis C, et al. Explainable artificial intelligence models for predicting risk of suicide using health administrative data in Quebec. PLoS One. 2024;19(4):e0301117.
10. Srinivasu PN, Sandhya N, Jhaveri RH, Raut R. From blackbox to explainable AI in healthcare: existing tools and case studies. Mobile Information Systems. 2022;2022.
11. Wang J, Kharrat FGZ, Pelletier JF, Rochette L, Pelletier E, Lévesque P, et al. A case–control study on predicting population risk of suicide using health administrative data: a research protocol. BMJ Open. 2023;13(2):e066423.
12. Holmes J, Sacchi L, Bellazzi R. Artificial intelligence in medicine. Ann R Coll Surg Engl. 2004;86:334–8.
13. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017); 2017; Long Beach, CA, USA. pp. 5998–6008.
14. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak. 2025;25(1):117.
15. Roumeliotis KI, Tselikas ND. Chatgpt and open-ai models: A preliminary review. Future Internet. 2023;15(6):192.
16. Liu A, Feng B, Xue B, Wang B, Wu B, Lu C, et al. Deepseek-v3 technical report. arXiv preprint arXiv:241219437. 2024.
17. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The llama 3 herd of models. arXiv preprint arXiv:240721783. 2024.
18. Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, et al. Gpt-4o system card. arXiv preprint arXiv:241021276. 2024.
19. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. The llama 3 herd of models. arXiv e-prints. 2024; arXiv-2407.
20. Bains SS, Dubin JA, Hameed D, Sax OC, Douglas S, Mont MA, et al. Use and application of large language models for patient questions following total knee arthroplasty. J Arthroplasty. 2024;39(9):2289–94.
21. Lieu B, Crawford E, Laubach L, Yeramosu T, Sharps C, Horstmann J, et al. Patient education strategies in pediatric orthopaedics: using ChatGPT to answer frequently asked questions on scoliosis. Spine Deform. 2025;1–13.
22. Margetts TJ, Karnik SJ, Wang HS, Plotkin LI, Oblak AL, Fehrenbacher JC, et al. Use of AI language engine ChatGPT 4.0 to write a scientific review article examining the intersection of Alzheimer's disease and bone. Curr Osteoporos Rep. 2024;22(1):177–81.
23. Guillen B, Murambadoro AK, Elizondo V, Hnatow M, Sander M. Evaluating AI Language Models for Patient Queries on Total Knee Replacement (TKR). 2024.
24. Caterson J, Ambler O, Cereceda-Monteoliva N, Horner M, Jones A, Poacher AT. Application of generative language models to orthopaedic practice. BMJ Open. 2024;14(3):e076484.
25. Sakai H, Lam SS, Mikaeili M, Bosire J, Jovin F. Large language models for patient comments multi-label classification. arXiv preprint arXiv:241023528. 2024.
26. Wang X, Zhu W, Saxon M, Steyvers M, Wang WY. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. Adv Neural Inf Process Syst. 2023;36:15614–38.
27. Gholi Zadeh Kharrat F, Amar Kumar P, Mehling W, Strigo I, Lotz J, Peterson TA, REACH Investigators Biobehavioral phenotypes of chronic low back pain: psychosocial subgroup identification using latent profile analysis. *Pain Med*. 2026;27:15-25. https://doi.org/10.1093/pm/pnaf095
28. Hue TF, Lotz JC, Zheng P, et al. Design of the COMEBACK and BACKHOME studies, longitudinal cohorts for comprehensive deep phenotyping of adults with chronic Low-Back

pain (cLBP): a part of the BACPAC research program. medRxiv. 2024, preprint: not peer reviewed.
29. Chakrabarti A, Ghosh JKA. BIC and recent advances in model selection. *Philos Stat*. 2011;583-605.
30. Yu Z, He L, Wu Z, Dai X, Chen J. Towards better chain-of-thought prompting strategies: A survey. arXiv preprint arXiv : 231004959. 2023;, preprint: not peer reviewed.
31. Zhang Z, Zhang A, Li M, Smola A. Automatic chain of thought prompting in large language models. arXiv preprint arXiv : 221003493. 2022;, preprint: not peer reviewed.
32. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst*. 2022;35:24824-24837.
33. Topsakal O, Akinci TC. Creating large language model applications utilizing langchain: a primer on developing llm apps fast. *ICAENS*. 2023;1:1050-1056.
34. Törnberg P. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Soc Sci Comput Rev.* 2025;43:1181-1195.
35. Nahum O, Calderon N, Keller O, Szpektor I, Reichart R. Are llms better than reported? detecting label errors and mitigating their effect on model performance. arXiv preprint, arXiv:241018889. 2024, preprint: not peer reviewed.

© The Author(s) 2026. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
JAMIA Open, 2026, 9, 1–10
https://doi.org/10.1093/jamiaopen/ooag058
Research and Applications