Can we use Large Language Models (LLMs) to assess the chronic pain experience?

Jacopo Amidei, Rubén Nieto, Andreas Kaltenbrunner, Jose Gregorio Ferreira De Sá, Mayte Serrat, Klara Albajes

Submitted to: Journal of Medical Internet Research
on: August 29, 2024

Table of Contents

Can we use Large Language Models (LLMs) to assess the chronic pain experience?

Jacopo Amidei[1] PhD; Rubén Nieto[2] PhD; Andreas Kaltenbrunner[1] PhD; Jose Gregorio Ferreira De Sá[1] PhD; Mayte Serrat[3] PhD; Klara Albajes[4] PhD

[1]AI and Data for Society Research Group (AID4So), Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya Barcelona ES
[2]eHealth Lab Research Group, Faculty of Psychology and Educational Sciences, Universitat Oberta de Catalunya Barcelona ES
[3]Unitat d'Expertesa en Síndromes de Sensibilització Central, Servei de Reumatologia, Vall d'Hebron Hospital Escola Universitària de Fisioteràpia, Escoles Universitàries Gimbernat, Universitat Autònoma de Barcelona Barcelona ES
[4]Psyclinic Barcelona ES

Corresponding Author:
Rubén Nieto PhD
eHealth Lab Research Group, Faculty of Psychology and Educational Sciences, Universitat Oberta de Catalunya
Rambla del Poblenou, 156.
Barcelona
ES

Abstract

Background: Chronic pain is a frequent problem in society, having an enormous impact. Tools that enhance the assessment to understand better people with pain experiences are essential to provide the care people need. In this line, a qualitative approach based in written narratives (WNs) from the people suffering chronic pain can be quite useful as supported in different studies. However, the assessment from this perspective can be time-consuming.

Objective: This study explores the feasibility of employing LLMs to assess WNs of people with chronic pain. At the end, we want to evaluate the potential of applying this LLMs to assist clinicians in assessing patients' pain.

Methods: We performed an experiment based on a list of pain narratives made by people with fibromyalgia and qualitatively evaluated in Serrat et al.[17]. Focusing on pain severity and disability, we prompt GPT-4 to assign scores and scores' explanations, to these narratives. Then we quantitatively compare GPT-4 scores with experts' scores of the same narratives, employing statistical measures such as Pearson correlations, Root Mean Squared Error (RMSE), Gwet's AC2 and Krippendorff's ?. Additionally, experts specialized in chronic pain conducted a qualitative analysis of the scores' explanation to assess their accuracy and potential applicability of GPT's analysis for future pain narrative evaluations.

Results: Our analysis reveals that GPT-4's performance in assessing pain narratives yielded promising results. GPT-4 was comparable in terms of agreement with experts, correlations with standardized measurements, and error rates. Moreover, experts generally deemed the ratings provided by GPT-4, as well as the scores' explanation, to be adequate.

Conclusions: These findings underline the potential of LLMs in facilitating the assessment of pain narratives, offering a novel approach to understanding and evaluating patient pain experiences. The integration of automated assessments through LLMs presents opportunities for streamlining and enhancing the evaluation process, paving the way for improved patient care and tailored interventions in the realm of chronic pain management.

(JMIR Preprints 29/08/2024:65903)
DOI: https://doi.org/10.2196/preprints.65903

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Jacopo Amidei[1], **Rubén Nieto**[2], **Andreas Kaltenbrunner**[1], **Jose Gregorio Ferreira De Sá**[1], **Mayte Serrat**[3,4], **Klara Albajes**[5]

[1] AI and Data for Society Research Group (AID4So), Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Barcelona, Spain
[2] eHealth Lab Research Group, Faculty of Psychology and Educational Sciences, Universitat Oberta de Catalunya, Barcelona, Spain
[3] Unitat d'Expertesa en Síndromes de Sensibilització Central, Servei de Reumatologia, Vall d'Hebron Hospital, Barcelona, Spain
4 Escola Universitària de Fisioteràpia, Escoles Universitàries Gimbernat, Universitat Autònoma de Barcelona, Sant Cugat del Vallès, Barcelona, Spain
[5] Psyclinic, Barcelona, Spain

Can we use Large Language Models (LLMs) to assess the chronic pain experience?

Abstract

Background: Chronic pain is a frequent problem in society, having an enormous impact. Tools that enhance the assessment to understand better people with pain experiences are essential to provide the care people need. In this line, a qualitative approach based in written narratives (WNs) from the people suffering chronic pain can be quite useful as supported in different studies. However, the assessment from this perspective can be time-consuming.

Objective: This study explores the feasibility of employing LLMs to assess WNs of people with chronic pain. At the end, we want to evaluate the potential of applying this LLMs to assist clinicians in assessing patients' pain.

Methods: We performed an experiment based on a list of pain narratives made by people with fibromyalgia and qualitatively evaluated in Serrat et al.[17]. Focusing on pain severity and disability, we prompt GPT-4 to assign scores and scores' explanations, to these narratives. Then we quantitatively compare GPT-4 scores with experts' scores of the same narratives, employing statistical measures such as Pearson correlations, Root Mean Squared Error (RMSE), Gwet's AC2 and Krippendorff's α. Additionally, experts specialized in chronic pain conducted a qualitative analysis of the scores' explanation to assess their accuracy and potential applicability of GPT's analysis for future pain narrative evaluations.

Results: Our analysis reveals that GPT-4's performance in assessing pain narratives yielded promising results. GPT-4 was comparable in terms of agreement with experts, correlations with

standardized measurements, and error rates. Moreover, experts generally deemed the ratings provided by GPT-4, as well as the scores' explanation, to be adequate.

Conclusions: These findings underline the potential of LLMs in facilitating the assessment of pain narratives, offering a novel approach to understanding and evaluating patient pain experiences. The integration of automated assessments through LLMs presents opportunities for streamlining and enhancing the evaluation process, paving the way for improved patient care and tailored interventions in the realm of chronic pain management.

Keywords: Large Language Models; Chronic Pain; Pain Narratives; Automated Assessment; Pain Severity

Introduction

Chronic pain poses a widespread challenge, affecting more than 20% of the global population [1,2,3,4]. Such kinds of pain are associated with restrictions in daily activities, disrupting normal functionality and diminishing the overall quality of life. Its repercussions extend to disruptions in familial, professional, and social domains [3]. Additionally, persistent pain is often linked to mental health issues, including mood and anxiety disorders. This issue significantly contributes to the demand for medical services, imposing a noteworthy economic burden on both individuals experiencing the pain and society at large. Accordingly, the economic ramifications at both the health and social levels are substantial. For example, estimates suggest that in Spain alone, the annual total cost, encompassing both direct and indirect expenses, could reach 16 billion euros[5].

Tools that enhance the assessment to understand better people with pain experiences are essential. Only with an adequate assessment, will we be able to determine the best intervention approach for each person and to evaluate the effects of interventions. Along these lines, standardized instruments (mainly self-reported measures) and established procedures exist which are available and recommended to be used routinely for assessing people with pain [6,7]. However, assessment procedures based on people's own experiences are still very scarce and could be very valuable to better understand the pain experience from a personal lens. In fact, a qualitative approach can capture subjective, intricate, and multifaceted details that standardized questionnaires can fail to capture[8].

In this context, the use of narrative holds significant potential in understanding the experiences of individuals living with pain. These provide researchers and clinicians with an opportunity to glean insights from the personal stories of individuals, allowing these individuals to highlight crucial aspects of their perspectives[9,10]. Examples of prior studies using narratives included the study by Noel et al.[11] who interviewed parents of young people with chronic pain, employing a mixed pain sample to extract and analyze the patient's narratives. Similarly, Meldrum et al.[12] conducted a narrative analysis of semi-structured interviews with children experiencing chronic pain, both at baseline and 6-12 months post-clinic intake. The use of narrative methodology has extended to written texts as well. McGowan et al.[13] solicited written narratives (WNs) from

women with chronic pelvic pain, while Dysvik et al.[14] examined WNs from individuals with chronic pain six years after completing a pain management intervention. More recently, WNs were used to explore the exerience of children with functional abdominal pain and their parents [15], adults with neck/back pain [16], and people with fibromyalgia[17]. In a related manner, Kathan et al. (2024) used a qualitative survey in which people were asked to respond to questions about pain acceptability by using their own words[18].

The growing interest in WNs is because asking for writing content to people suffering a condition offer distinct advantages over oral inquires. Primarily because writing facilitates the organization of ideas related to complex emotional experiences such as pain[19]. Additionally, WNs offer a time-efficient way to explore the subjective nature of pain, as individuals can complete the writing outside of healthcare consultations, making it an accessible initial approach to this complex phenomenon. However, on the other hand, analyzing the contents expressed by people with pain can be time-consuming for clinicians/researchers. For this reason, this study explores the feasibility of employing LLMs to assess WNs of people with chronic pain. Using LLMs provides several advantages, such as decreasing assessment time for clinicians — where LLMs serve as clinician assistants — and allowing language flexibility — meaning our methodology can be applied to various languages. This study can contribute to the open debate about the potential application of LLMs for health. While there are existing literature supporting benefits such as the capabilities for analysing massive data, there are also studies showing disadvantages such as inaccuracies with the use of LLMs[20].

Artificial Intelligence (AI) has been used in the assessment of the pain field in some studies, as presented in the recent review by Abd-Elsayedin et al.[21]. In this review, studies were found to address the following purposes: 1) Diagnostic aid, 2) Modelling Pain Progression, 3) Predicting Pain Treatment Response, and 4) Improving Treatment and Pain Maintenance. To achieve these, most of the studies used machine learning techniques. Although none of the reviewed studies used LLMs, there are few works available using them for pain research. Vaid et al.[22], used a locally-running, privacy-preserving LLMs capable of following plain language instructions to extract characteristics of musculoskeletal pain (such as location and acuity) from a heterogeneous collection of unstructured clinical notes. The study used multiple patient notes, coded by two healthcare professionals, and found great precisions of the system in classifying pain location and acuity. In another study, Shrestha et al.[23] tested the responses of GPT to clinical questions and recommendations based on an established clinical guideline. They found that the system was able to make clinical recommendations for low back pain, although it was not exempt from errors. In the same line, Gianola et al.[24] tested the ability of GPT against clinical practice guidelines to answer clinical questions about lumbosacral radicular pain. They found negative results since the internal consistency was low, as well as the precision to follow the clinical guidelines for recommendations.

To evaluate LLMs' capacity to assess pain narratives we used the narratives provided in Serrat et al. [17]. In our investigation, GPT-4 was employed to assign scores, as well as the scores' explanation, for *pain severity* and *disability* as expressed in the narratives. Subsequently, we conducted a quantitative analysis by comparing these scores with expert ratings reported in Serrat et al. [17], utilizing statistical measures such as Pearson correlations, Root Mean Squared Error

(RMSE), Gwet's AC2[25] and Krippendorff's α[26]. Additionally, a qualitative analysis of the scores' explanation was performed by consulting two experts specialized in chronic pain, who evaluated the accuracy of GPT's pain assessment and its potential utility for future pain WNs assessments. Altogether, the primary contribution of this paper is its pioneering exploration of applying LLMs for WNs assessment, validated through both automatic and human evaluation methods. To the best of our knowledge, this paper is the first attempt at using LLMs to assess pain WNs.

Methods

Procedure

This study uses participant data from a previous qualitative study[17] which explored the value of WNs for understanding the experience of people with fibromyalgia (FM). Specifically, we used patient WNs, their assessments by two experts, and patient answers on standardized questionnaires (see Section Dataset) to re-analyze the data using GPT-4 (see Section Experiments with GPT) and compare its outcomes with the previous results.

Dataset

A total of 46 people completed the WNs task in Serrat et al.[17]. The inclusion criteria for these participants were: (1) fulfillment of the 2010/2011 American College of Rheumatology Fibromyalgia diagnostic criteria[27] (; and (2) age 18 or older. The exclusion criteria were having terminal illnesses or programmed interventions that might interrupt the study.

In the present study, we selected data from 43 participants, i.e. the ones written in Spanish (three participants who did the task in a different language were excluded). They were requested to write about their pain experiences following instructions. The narrative's objective was to capture personal viewpoints; for this reason, the guidelines, and instructions were meticulously crafted to be clear, motivational, and inclusive, fostering a space for diverse perspectives and opinions (see more details in Serrat et al.[17]). Participants were given the option to complete the task in the language most comfortable and convenient for them. Participants were also asked to complete the following questionnaires:

- *Revised Fibromyalgia Impact Questionnaire* (FIQR): A 20-item questionnaire that measures functional impairment over the last 7 days. It has 3 dimensions: physical dysfunction (scores from 0 to 30), overall impact (scores from 0 to 20), and intensity of symptoms (scores from 0 to 50). The total sum of these scores ranges from 0 to 100, and higher scores indicate a greater impact. The Spanish version shows adequate internal consistency (Cronbach α = .93) [28].
- *Hospital Anxiety and Depression Scale* (HADS): A commonly used questionnaire that evaluates the severity of anxiety and depression symptoms by two scales (consisting each of 7 items). Scores on each scale range from 0 to 21, with higher scores indicating greater severity of symptoms. The Spanish version has shown adequate internal consistency both for anxiety (α = 0.83) and depression (α = 0.87) subscales in individuals with FM[29].

- *Tampa Scale for Kinesiophobia* (TSK): A scale composed of 11 items, to be answered on a 4-point Likert scale. The scale quantifies fear of movement or (re)injury. Its total scores can range from 11 to 44, where higher scores indicate a greater fear of pain and movement. The Spanish version shows adequate internal consistency (α = .79)[30].

In Serrat et al.[17] two independent reviewers (with expertise in pain, one psychologist, and one physiotherapist) assessed the level of severity and disability expressed in the WNs on a scale from 0 (indicating absence) to 10 (representing maximum levels). To ensure consistency in their evaluations, severity was defined as "the perceived magnitude of FM concerning pain and overall suffering conveyed in each participant text". Disability was defined as "the perceived extent to which FM disrupts the usual activities and life of the writers".

Experiments with GPT

We tasked GPT-4 to evaluate the previously described WNs one by one. Specifically, we prompted GPT-4 to provide a score for pain severity and disability, as well as the scores' explanation. Both of the scores were requested on a scale from 0 to 10 to compare with human assessments performed by Serrat et al.[17] . To test the stability of these responses, we repeat this experiment n=10 times. Both the scores of the 10 trials and the explanations (for one of the trials) were then used for the evaluation phase. The specific prompting strategy used is displayed in Table 1 and was performed with automated calls to the OpenAI API with model *gpt-4-0125-preview*. This automation saved time and ensured consistent scoring.

Table 1. Prompts used for the experiment

"As an expert psychologist specializing in evaluating pain in patients diagnosed with fibromyalgia, you are tasked with analyzing patient narratives about their pain and then scoring them on a scale from 0 (indicating no severity or disability) to 10 (indicating maximum severity and disability). These patients' explanations about their pain and how they feel it are all written in Spanish. The level of severity, defined as the perceived intensity of pain and overall suffering, and disability, defined as the extent to which fibromyalgia hinders patients' usual activities and quality of life, are to be rated based on your interpretations of the patients' texts. Scores should accurately reflect the severity and disability levels described in patient narratives without inflation. A holistic evaluation capturing the complexity of experiences is crucial. Pay attention to phrases indicating coping mechanisms, resilience, or mitigating factors that may reduce perceived severity or disability. Consider contextual understanding, including coping strategies, support systems, and adaptive behaviors, which may mitigate perceived severity and disability. Your role involves receiving a patient's narrative, enclosed within triple slashes, and analyzing it. You are expected to return your analysis in JSON format, with the following keys: "severity_score" providing the scores for severity ranging from 0 to 10, "disability_score" providing the scores for disability ranging from 0 to 10, "severity_explanation" providing an in English explanation for the severity score and "disability_explanation" providing an in English explanation for the disability score"

We performed the experiments with two different values for GPT's temperature parameter. This parameter allows controlling the randomness in the answers of the LLMs. We use 0 (less randomness) and 1 (average randomness), the latter is the default value of GPT-4.

Experts evaluation

We asked two experts in pain management and assessment to analyse the scores given by GPT-4 and the corresponding textual explanations[1]. Specifically, we asked them to read the original narratives, the scores, and their explanations given by GPT-4 for pain severity and disability, and to assess on a seven-point scale (from strongly disagree to strongly agree) to what extent the explanation: 1) could have been written by a psychologist expert in fibromyalgia, 2) adequately represents the scores for severity or disability, and 3) they would use the score and explanation provided by GPT-4 for patient assessment.

Evaluation

We performed a four-stage analysis. First, we used standard deviation analysis to test the stability of assessments given in 10 trials by GPT-4 for pain severity and disability.

Second, we compared GPT-4 scores with experts' scores (from Serrat et al.[17]) and a naive baseline predictor (which always predicts the average experts' score) using three strategies: 1) Inter-Annotator Agreement (IAA) to quantify the agreement between GPT-4 and the experts, 2) RMSE to measure the average squared difference between GPT-4, expert scores and the naive baseline, and 3) Mean Absolute Error (MAE) to determine if GPT-4 systematically overestimates or underestimates the expert scores (i.e. if GPT-4 exhibits bias). IAA was assessed using four coefficients to ensure data reliability[2]: *percent agreement*, *weighted percent agreement*[3], *Krippendorff's α[26]*, and *Gwet's AC2* coefficient[25]. We reported Gwet's AC2 coefficient due to its ability to address some paradoxical behaviour compared to other popular chance-corrected coefficients like Krippendorff's α, Cohen's kappa[31] or Fleiss's kappa[32]. Our experiments show indeed an imbalance in ratings towards higher categories (6-10), making chance-corrected coefficients less suitable due to the prevalence paradox[33].

Third, we compute descriptive statistics for the experts' evaluation of the GPT-4 assessments (for both the scores and their explanations), as well as IAA (in this case, Krippendorff's α was not reported because of the significant imbalance in scores, particularly towards the higher end).

Fourth, we tested correlations between GPT-4 assessments, expert assessments, and standardized pain measurements (from Serrat et al.[17]).

Results

GPT-4 assessment stability

The mean and standard deviation (SD) for pain severity and disability were assessed for 10 trials at two different GPT-4 temperatures (0 and 1). At temperature 0, the mean severity score was 8.13 (SD

1 Although they form part of the list of authors of this paper, they only were aware of the general objective of the study and the instructions provided for their task. After doing their task, they were given aces to all the details of the study and the manuscript.

2 The coefficients of agreement were measured by the irrCAC library "https://github.com/afergadis/irrCAC/tree/master". The weighted percent agreement, Krippendorff's α and Gwet's AC2 coefficient were measured with ordinal weight.

3 This is a weighted version of percent agreement that takes into account the ordinal nature of the data.

1.09) and the mean disability score was 7.25 (SD 1.28). At temperature 1, the mean severity score was 8.08 (SD 1.02) and the mean disability score was 7.33 (SD 1.32). These results suggest stability in both pain severity and disability across the different 10 trials and temperatures.

GPT-4 assessment stability

Results related to the IAA are presented in Table 2. IAA between experts is acceptable (both for pain severity and disability) considering Krippendorff's α and Gwet AC2 coefficients. Delving deeper into the IAA analysis, the low percentage agreement (0.29 for pain severity and 0.31 for disability), combined with a high weighted percentage agreement (0.96 for pain severity and 0.95 for disability), suggests that while experts rarely chose the exact same score, their disagreements were typically within adjacent scores. Given the subjectivity inherent in assessing WNs, which is influenced by numerous factors, the IAA demonstrates a strong agreement among experts.

Table 2. Mean (± Standard Deviation) of the agreement between experts and GPT-4. Values for GPT-4 are averages over 10 experiments.

	Expert 1 vs.			Expert 2 vs.	
	Expert 2	**GPT-4 (Temp 0)**	**GPT-4 (Temp 1)**	**GPT-4 (Temp 0)**	**GPT-4 (Temp 1)**
Pain Severity					
Percent agreement	0.29	0.36 (± 0.01)	0.36 (± 0.04)	0.27 (± 0.01)	0.32 (± 0.03)
Weighted percent agreement	0.96	0.94 (± <0.01)	0.94 (± 0.01)	0.95 (± <0.01)	0.95 (± <0.01)
Gwet´s AC2	0.87	0.83 (± <0.01)	0.84 (± 0.02)	0.84 (± <0.01)	0.84 (± 0.01)
Krippendorff´s α	0.66	0.46 (± 0.01)	0.45 (± 0.05)	0.51 (± 0.01)	0.49 (± 0.04)
Disability					
Percent agreement	0.31	0.21 (± 0.02)	0.23 (± 0.04)	0.33 (± 0.02)	0.35 (± 0.08)
Weighted percent agreement	0.95	0.94 (± <0.01)	0.94 (± 0.0)	0.94 (± <0.01)	0.94 (± <0.01)
Gwet´s AC2	0.83	0.79 (± <0.01)	0.79 (± 0.01)	0.8 (± <0.01)	0.8 (± 0.02)

Krippendorff´s α	0.69	0.47 (± 0.01)	0.49 (± 0.03)	0.57 (± 0.01)	0.57 (± 0.04)

When comparing GPT-4 scores with expert scores, Krippendorff's α indicated slightly lower agreement compared to the agreement between the experts. However, percent agreement, weighted percent agreement and Gwet's AC2 are comparable to the agreement between experts, with the notable exception that, for pain severity (temperature 1), percent agreement was higher than the one reached by the experts.

We compare the scores of pain severity and disability obtained with GPT-4 to those of the experts and the naive baseline using RMSE and MAE (by definition, MAE is 0 for the naive baseline).

Table 3. Predicting severity and disability scores with GPT with different temperature values. Root Mean Squared Error (RMSE) and Mean Average Error (MAE) of GPT-4 compared to two expert evaluations and a naive baseline. * indicate a significant difference between the RSMEs of GPT-4 and the baseline (two-sided p-value equivalent of the z-score of 10 experiments with GPT-4 and the baseline, * p<0.1, ** p<0.01, *** p<0.001).

	Expert 1 vs.				Expert 2 vs.			Average of experts vs.		
	Expert 2	GPT-Temp 0	GPT-Temp 1	Baseline	GPT-Temp 0	GPT-Temp 1	Baseline	GPT-Temp 0	GPT-Temp 1	Baseline
Pain severity										
RSME	1.15	1.27 ***	1.25 *	1.38	1.39 ***	1.40 **	1.72	1.20 ***	1.19 **	1.45
MAE	0.12	-0.71	-0.66	0.00	-0.83	-0.78	0.00	-0.77	-0.72	0.00
DISABILITY										
RSME	1.56	1.75 ***	1.74 **	2.20	1.50 ***	1.53 **	2.05	1.44 ***	1.44 **	1.98
MAE	0.05	-0.25	-0.33	0.00	-0.30	-0.37	0.00	-0.27	-0.35	0.00

GPT-4 can approximate the expert ratings with RMSEs between 1.25 and 1.40 for severity and 1.5 and 1.75 for disability (see Table 3). In both cases, these values are significantly lower than the ones obtained with the naive baselines with p<0.001 for GPT-4 with temperature 0 and p<0.01 for temperature 1 on five cases out of six, having p<0.1 in the remaining case. Furthermore, we can observe that the RSMEs between the two experts are very close to the ones obtained by GPT-4 when compared to the average of the two experts for disability and even smaller for pain severity. Overall these errors are acceptable, especially when compared to the differences between the two

experts.

We also analyze the MAEs (i.e. expert score minus GPT-4 score) which indicate a potential tendency to underrate or overate the expert scores. For both temperatures, we find that GPT-4 on average slightly overestimates the experts´ scores. More precisely, between 0.66 (temperature 1) and 0.83 (temperature 0) for severity and 0.25 (temperature 0) and 0.37 (temperature 1) for disability.

Finally, all the results commented are quite similar when comparing Temperature 0 and 1. The comparison between GPT-4 scores and those provided by experts highlights a significant alignment in their assessments of WNs. This finding holds promise, especially given the inherent subjectivity involved in evaluating WNs. To delve deeper into this alignment, we enlisted two pain assessment experts to evaluate the GPT-4 scores and their accompanying descriptions.

Experts' evaluation

The IAA among experts is acceptable, as shown in Table 4. The low percent agreement is compensated by a notably high weighted percent agreement (with the exception of pain severity in question 1). This suggests that, although experts rarely assign the same score, when a disagreement arises, they tend to choose adjacent ratings. This phenomenon indicating acceptable IAA is also reflected in Gwen's AC2 values (with the exception of disability in question 2).

Table 4. Agreement between experts on the 3 questions. The coefficients were calculated by assigning numerical values to the categories: Strongly disagree (1), Disagree (2), Somewhat disagree (3), Neither agree nor disagree (4), Somewhat agree (5), Agree (6), and Strongly agree (7). Ordinal weights were then applied to these categories based on their positions in the scale (Gwet, 2014). This method acknowledges that experts may differ more significantly in their disagreement if one selects "Somewhat disagree" while the other selects "Agree" compared to if one chooses "Agree" while the other chooses "Strongly agree".

	Question 1	Question 2	Question 3
Pain Severity			
Percent agreement	0.08	0.21	0.12
Weighted percent agreement	0.50	0.95	0.91
Gwet´s AC2	0.93	0.83	0.66
Disability			

Percent agreement	0.17	0.21	0.33
Weighted percent agreement	0.95	0.96	0.94
Gwet´s AC2	0.81	0.43	0.72

As shown in Table 5, a t-test analysis indicates that Expert 2's assessments were significantly higher ($p<0.001$) for all three questions compared to Expert 1's assessments. This phenomenon implies a divergence in experts' scoring interpretations. Despite this difference, both experts appear to adhere consistently to their respective scoring criteria during the annotation process. This finding suggests that while there may be slightly individual variations in scoring approaches between experts, they maintain internal consistency in their assessments.

Table 5. Mean (± Standard Deviation) of the expert evaluation scores (max score is 7) and two-sided t-test (comparing the two experts) for each question. The coefficients were calculated by assigning numerical values to the categories: Strongly disagree (1), Disagree (2), Somewhat disagree (3), Neither agree nor disagree (4), Somewhat agree (5), Agree (6), and Strongly agree (7). **Question 1**: The explanation could have been written by a psychologist expert in fibromyalgia; **Question 2:** The explanation adequately represents the pain severity expressed in the narrative; and **Question 3:** I would use the pain severity score and explanation above to help myself assess the patient's pain.

		Expert 1	Expert 2	t-statistic
Question 1	**Pain Severity**	5.72 (± 0.45)	6.88 (± 0.32)	-13.67*
	Disability	5.93 (± 0.26)	6.83 (± 0.37)	-13,10*
Question 2	**Pain Severity**	5.93 (± 0.26)	6.79 (± 0.41)	-11.62*
	Disability	6 (± 0)	6.79 (± 0.41)	-12.60*
Question 3	**Pain Severity**	5.44 (± 0.63)	6.72 (± 0.45)	-10.82*
	Disability	5.77(± 0.43)	6.65 (± 0.48)	-8.99*

*p<0.001

The mean assessment scores for both for pain severity and disability are presented in Table 5. Assessments for the three questions were high (scores ranging from Somewhat agree (5) and Agree (6) for Expert 1 and Agree (6), and Strongly agree (7) for Expert 2), and with low variability (standard deviation, ranging from 0 to ± 0.63). In other words, the experts agreed to consider GPT-4 assessments accurate and usable as clinician assistants.

In conclusion, the disagreement observed between the experts appears to stem from differing interpretations of scoring criteria, as evidenced by the consistent trend of one expert assigning higher scores than the other expert during disagreements. This indicates a structural disparity in their evaluation approaches. Despite this variation, both experts concur on the utility of GPT-4 as a valuable tool for pain assessment tasks. This alignment in their assessment of GPT-4's effectiveness underscores its potential to complement and enhance traditional expert evaluations in pain narrative analysis.

Correlations of human and GPT-4 assessments with standardized measurements

The mean of the ratings assigned by experts for severity and disability (from Serrat et al. [17]) significantly correlated with scores from the FIQR questionnaire, and the anxiety and

depression scores from the HAD questionnaire. When analyzing the corresponding correlations with the scores given by GPT-4, for temperature 0, results were very similar to the ones found for experts with the exception that in this case there were no significant correlations between pain severity and disability and anxiety scores. However, the correlations between scores assigned by GPT-4 with temperature 1 and depression and anxiety, were in the same line as the ones found for the experts (see Table 6 for further details). Also in this case, the results indicate that GPT-4's performance in assessing pain WNs closely approximates humans' assessment. In addition, a further encouraging outcome is found: GPT-4's pain WNs assessment shows a favourable alignment with standard pain assessment tests (for example, see GPT-4 with temperature 1 in the assessments of FIQR and HADs Depression in Table 6).

Table 6. Pearson correlations of expert and GPT assessments with standardized measurements. *p<0.05; **p<0.01.
FIQ-R: Fibromyalgia Impact Questionnaire; HADs: Hospital Anxiety and Depression Inventory;
TSK: Tampa Scale of Kinesiophobia.

				HADs	
		FIQ-R	**TSK**	**Anxiety**	**Depression**
Experts	**Pain Severity**	0.41**	0.18	0.34*	0.44**
	Disability	0.44**	0.19	0.38*	0.46**
GPT-4 (Temp 0) Mean of 10 Trials	**Pain Severity**	0.36*	0.17	0.25	0.35*
	Disability	0.42**	0.21	0.28	0.36*
GPT-4 (Temp 1) Mean of 10 Trials	**Pain Severity**	0.43**	0.18	0.32*	0.41**
	Disability	0.49**	0.16	0.34*	0.45**

Discussion

Principal findings

Our preliminary study highlights the potential of utilizing LLMs, such as GPT-4, for automatizing the assessment of pain severity and disability levels in patient narratives. Pain narratives can be very useful for people's pain assessment, but are time-consuming for clinicians. The methodology based on LLMs presented in this paper conducts an automated assessment of the levels of pain severity and disability in the patient's written content. Our results indicate that experts in pain assessment can make use of LLMs for faster patient assessment. Indeed, the conducted analysis, bolstered by various statistical measures, reveals a significant resemblance between expert scores and those generated by GPT-4. This observation is further supported by the comparable correlation values observed between standardized measurements and assessments by both experts and GPT-4. Moreover, the positive reception from experts regarding the scores and explanations generated by GPT-4 underscores the potential applicability of automated systems in pain assessment. It is worth nothing that both experts agree on the perceptions about GPT-4's scores and explanations: although one of them seems more positive in her assessments, leading to some variations in agreement indices.

Future research

While these findings are promising, further research is essential to advance this area. In this study, we used pain severity and disability as main indicators of the texts, since we wanted to explore variables relevant in clinical context. International guidelines support measuring these in the pain field[6,7]. Moving forward, it would be beneficial to explore in future studies additional variables beyond pain severity and disability that could enhance the clinical relevance of automated assessments. For example, future research could instruct LLMs to assess levels of catastrophizing thoughts in the texts. These are common patterns of thinking in people with pain, related to a worse adaptation to pain in multiple studies (e.g. Quartana et. al.[34]). Assessing this variable adequately would be very useful in the clinical context to identify people at risk of suffering more complex problems that need more attention.

Apart from studying other variables, an additional path for future research could be to study ways of improving the performance of the LLMs model, for example through few-shot learning methods. Our findings suggest the need for ongoing efforts to enhance the precision of automated LLMs in pain narrative assessments. While the agreement between GPT-generated scores and expert ratings was generally favourable, there is room for improvement, particularly in aligning with certain evaluation metrics like Krippendorff's alpha. By refining the training data and algorithms used by LLMs, we can strive to achieve even greater accuracy and reliability in automated pain assessments.

Another line of research can involve the comparison between the explanations of scores provided by an LLM and the corresponding explanations provided by experts. For example, a blind annotation can aim to evaluate if experts would be able to distinguish the origin of these explanations. Regarding the LLM scores´ explanation, as reported in Section *3. Experts´ evaluation*, upon a detailed examination of the experts' evaluation, it becomes apparent that the disagreement stems from one expert consistently assigning slightly higher scores across all parameters, displaying a more optimistic outlook. The identification of a clear pattern in the evaluation process is particularly intriguing as it suggests a consistent trend in how assessments were conducted and interpreted by the experts. On the one hand, further analysis of this pattern could provide valuable insights into the underlying factors influencing the evaluation outcomes and shed light on the reliability and consistency of the assessment process. Understanding and leveraging such patterns can enhance the effectiveness and accuracy of automated systems like LLMs in pain narrative assessments. On the other hand, understanding this pattern in expert evaluations can provide valuable insights into the subjective nature of pain assessments and the potential impact on IAA in pain narrative evaluations.

Finally, this study, due to its preliminary nature, focused on people with fibromyalgia. Future research would benefit of including people with different chronic pain problems, and bigger samples to compare the performance of GPT-4 in different pain conditions. Big differences are not expected, since we assume that GPT-4 could be able to assess equally texts from people with other pain problems. However, empirical tests are needed to support this assumption.

Conflicts of Interest

None of the authors have conflicts of interest do declare.

Abbreviations

AI: Artificial Intelligence
FIQR: Revised Fibromyalgia Impact Questionnaire
HADS: Hospital Anxiety and Depression Scale
IAA: Inter-Annotator Agreement
LLMs: Large Language Models
MAE: Mean Absolute Error
RMSE: Root Mean Squared ErrorRCT: randomized controlled trial
SD: Standard Deviation
TSK: Tampa Scale for Kinesiophobia
WNs: Written Narratives

References

1. Johannes CB, Le TK, Zhou X, et al. The prevalence of chronic pain in United States adults: results of an Internet-based survey. *J Pain* 2010; 11:1230-9. PMID: 20797916.

2. Leadley RM, Armstrong N, Lee YC, et al. Chronic diseases in the European Union: the prevalence and health cost implications of chronic pain. *J Pain Palliat Care Pharmacother*. 2012; 26:310-325. PMID: 23216170.

3. Breivik H, Collett B, Ventafridda V, et al. Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. *Eur J Pain* 2006; 10: 287-333.PMID: 16095934.

4. Rikard SM, Strahan AE, Schmit KM, Guy GP Jr. Chronic Pain Among Adults - United States, 2019-2021. *MMWR Morb Mortal Wkly Rep*. 2023;72(15):379-385. Published 2023 Apr 14. doi:10.15585/mmwr.mm7215a1. PMID: 37053114.

5. Torralba A, Miquel A, Darba J. Situación actual del dolor crónico en España: iniciativa" Pain Proposal". *Rev. Soc. Esp. Dolor* 2004; 21: 16-22. https://dx.doi.org/10.4321/S1134-80462014000100003.

6. Turk DC, Dworkin RH, Allen RR, et al. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. *Pain* 2003; 106: 337-345. PMID: 14659516.

7. Dworkin RH, Turk DC, Farrar JT, et al. Core outcome measures for chronic pain clinical trials:

IMMPACT recommendations. *Pain* 2005; 113: 9-19. PMID: 15621359.

8. Morse JM. Using qualitative methods to access the pain experience. *Br J Pain*. 2015;9:26-31. PMID: 26516553.

9. Hall JM, Powell J. Understanding the person through narrative. *Nurs Res Pract* 2011; 2011:293837.PMID: 21994820.

10. Vindrola-Padros C, Johnson GA. The narrated, nonnarrated, and the disnarrated: conceptual tools for analyzing narratives in health services research. *Qual Health Res* 2014; 24:1603-11. PMID: 25192757.

11. Noel M, Beals-Erickson SE, Law EF, et al. Characterizing the pain narratives of parents of youth with chronic pain. *Clin J Pain* 2016;32:849–58.PMID: 26736026.

12. Meldrum ML, Tsao JCI, Zeltzer LK. "I can't be what I want to be": Children's narratives of chronic pain experiences and treatment outcomes. *Pain Med* 2009;10:1018–34. PMID: 19594848.

13. McGowan L, Luker K, Creed F, et al. How do you explain a pain that can't be seen? The narratives of women with chronic pelvic pain and their disengagement with the diagnostic cycle. *Br J Health Psychol* 2007;12:261–74. PMID: 17456285.

14. Dysvik E, Natvig GK, Furnes B. A narrative approach to explore grief experiences and treatment adherence in people with chronic pain after participation in a pain-management program: A 6-year follow-up study. *Patient Prefer Adherence* 2013;7:751–9.PMID: 23990710.

15. Nieto R, Sora B, Boixadós M, et al. Understanding the Experience of Functional Abdominal Pain Through Written Narratives by Families. *Pain Med* 2020; 21:1093-1105. PMID: 31361016.

16. Sora B, Nieto R, Vall-Roqué H, et al. Chronic neck and low back pain from personal experiences: a written narrative approach. *Pain Manag 2024*; 14:183-194. PMID: 38717373.

17. Serrat M, Sora B, Ureña P, Vall-Roqué H, Edo-Gual M, Nieto R. Written narratives to understand the experience of individuals living with fibromyalgia. *Musculoskeletal Care*. 2024;22(2):e1905. PMID: 39031673.

18. Kahtan H, Jordan A, Forget P. Is pain ever acceptable? A qualitative exploration concerning adult perceptions of chronic pain. *Eur J Pain*. 2024;28(7):1213-1225. PMID: 38400800.

19. Pennebaker JW, Seagal JD. Forming a story: the health benefits of narrative. *J Clin Psychol*. 1999;55(10):1243-1254. PMID: 11045774.

20. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* 2023;11:887. PMID: 36981544.

21. Abd-Elsayed A, Robinson CL, Marshall Z, et al. Applications of Artificial Intelligence in Pain Medicine. *Curr Pain Headache Rep* 2024;28:229-238. PMID: 38345695.

22. Vaid A, Landi I, Nadkarni G, et al. Using fine-tuned large language models to parse clinical notes in musculoskeletal pain disorders. *Lancet Digital Health* 2023; 12: e855–e858. DOI: 10.1016/S2589-7500(23)00202-9.

23. Shrestha N, Shen Z, Zaidat B, et al. Performance of ChatGPT on NASS Clinical Guidelines for the Diagnosis and Treatment of Low Back Pain: A Comparison Study. *Spine* 2024; 49:640-651. PMID: 38213186.

24. Gianola S, Bargeri S, Castellini G, et al. Performance of ChatGPT Compared to Clinical Practice Guidelines in Making Informed Decisions for Lumbosacral Radicular Pain: A Cross-sectional Study. *J Orthop Sports Phys Ther* 2024; 54:1-7. PMID: 38284363.

25. Gwet, Kilem L. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

26. Krippendorff K. Measuring the reliability of qualitative text analysis data. *Quality and Quantity* 2004; 38: 787-800. https://doi.org/10.1007/s11135-004-8107-7

27. Bennett RM, Friend R, Jones KD, et al. The Revised Fibromyalgia Impact Questionnaire (FIQR): validation and psychometric properties. *Arthritis Res Ther* 2009;11:R120. PMID: 19664287.

28. Luciano JV, Aguado J, Serrano-Blanco A, et al. Dimensionality, reliability, and validity of the revised fibromyalgia impact questionnaire in two Spanish samples. *Arthritis Care Res* 2013;65(10):1682-9. PMID: 23609980.

29. Luciano JV, D'Amico F, Cerdà-Lafont M, et al. Cost-utility of cognitive behavioral therapy versus U.S. Food and Drug Administration recommended drugs and usual care in the treatment of patients with fibromyalgia: an economic evaluation alongside a 6-month randomized controlled trial. *Arthritis Res Ther* 2014;16:451. PMID: 25270426.

30. Gómez-Pérez L, López-Martínez AE, Ruiz-Párraga GT. Psychometric Properties of the Spanish Version of the Tampa Scale for Kinesiophobia (TSK). *J Pain* 2011; 12:425-35. PMID: 20926355.

31. Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 1960; 20:37–46. https://doi.org/10.1177/001316446002000104

32. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971;76:378, 1971. https://doi.org/10.1037/h0031619.

33. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34:555–596, 2008. https://doi.org/10.1162/coli.07-034-R2

34. Quartana PJ, Campbell CM, Edwards RR. Pain catastrophizing: a critical review. *Expert Rev Neurother* 2009;9:745-58. PMID: 19402782.