RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL RESEARCH

Large language model-driven sentiment analysis for facilitating fibromyalgia diagnosis

Vincenzo Venerito , Florenzo Iannone

To cite: Venerito V, Iannone F. Large language model-driven sentiment analysis for facilitating fibromyalgia diagnosis. *RMD Open* 2024;**10**:e004367. doi:10.1136/rmdopen-2024-004367

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/rmdopen-2024-004367).

Received 27 March 2024
Accepted 8 May 2024

Rheumatology Unit - Department of Precision and Regenerative Medicine and Ionian Area, University of Bari "Aldo Moro", Bari, Italy

Correspondence to
Prof Florenzo Iannone; florenzo.iannone@uniba.it

ABSTRACT

Background Fibromyalgia (FM) is a complex disorder with widespread pain and emotional distress, posing diagnostic challenges. FM patients show altered cognitive and emotional processing, with a preferential allocation of attention to pain-related information. This attentional bias towards pain cues can impair cognitive functions such as inhibitory control, affecting patients' ability to manage and express emotions. Sentiment analysis using large language models (LLMs) can provide insights by detecting nuances in pain expression. This study investigated whether open-source LLM-driven sentiment analysis could aid FM diagnosis.

Methods 40 patients with FM, according to the 2016 American College of Rheumatology Criteria and 40 non-FM chronic pain controls referred to rheumatology clinics, were enrolled. Transcribed responses to questions on pain and sleep were machine translated to English and analysed by the LLM Mistral-7B-Instruct-v0.2 using prompt engineering targeting FM-associated language nuances for pain expression ('prompt-engineered') or an approach without this targeting ('ablated'). Accuracy, precision, recall, specificity and area under the receiver operating characteristic curve (AUROC) were calculated using rheumatologist diagnosis as ground truth.

Results The prompt-engineered approach demonstrated accuracy of 0.87, precision of 0.92, recall of 0.84, specificity of 0.82 and AUROC of 0.86 for distinguishing FM. In comparison, the ablated approach had an accuracy of 0.76, precision of 0.75, recall of 0.77, specificity of 0.75 and AUROC of 0.76. The accuracy was superior to the ablated approach (McNemar's test p<0.001).

Conclusion This proof-of-concept study suggests LLM-driven sentiment analysis, especially with prompt engineering, may facilitate FM diagnosis by detecting subtle differences in pain expression. Further validation is warranted, particularly the inclusion of secondary FM patients.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Diagnosing fibromyalgia is challenging. Sentiment analysis using natural language processing shows promise in detecting nuances in patient-reported symptoms, but its application for aiding fibromyalgia diagnosis is unexplored.

WHAT THIS STUDY ADDS

⇒ Using sentiment analysis on patient responses, a locally run, open-source, large language model can accurately distinguish fibromyalgia from other chronic pain conditions. Prompt engineering significantly improved the model's performance.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This approach could complement clinical assessment in fibromyalgia diagnosis. Further validation, integration with patient-reported outcomes and model interpretability optimisation are needed for clinical translation.

INTRODUCTION

Fibromyalgia (FM) is a chronic and complex disorder characterised by widespread musculoskeletal pain, fatigue and a plethora of other symptoms, including sleep disturbances, cognitive difficulties and emotional distress. The elusive nature of FM, which affects roughly 2%–4% of the population, poses a significant challenge for healthcare professionals, patients and researchers. Despite its prevalence, FM remains an often misunderstood condition, both within and outside the medical community. The diagnostic journey for FM patients is often fraught with challenges, delays and frustrations, and healthcare professionals struggle with this elusive disorder. The complex nature of FM involves patients' emotional sphere, encompassing emotions like anger, fear and low self-esteem that negatively affect relationships and daily functioning.

FM patients may have difficulty disengaging from pain stimuli, which could affect their pain expression and cognitive function. Links between these factors and symptoms like anxiety and depression demonstrate the intertwining of emotion in FM.[1]

FM patients show altered cognitive and emotional processing, with a preferential allocation of attention to pain-related information. This attentional bias towards pain cues can impair cognitive functions such as inhibitory control, affecting patients' ability to manage and express their emotions.[2]

Sentiment analysis, an innovative application of natural language processing and text mining techniques, has emerged as a promising tool in the field of healthcare and biomedical informatics.[3] Large language models (LLMs) have significantly transformed the field of sentiment analysis by enhancing the accuracy and depth of understanding of text.[4 5]

Open-source LLMs such as Mistral have overcome the concerns on privacy and data protection raised by Generative Pre-trained Transformer-4 and have been trained on extensive textual data, enabling them to grasp complex grammatical structures, language patterns and contextual nuances.[5] This deep understanding allows LLMs to perform sentiment analysis with a high degree of precision, even in texts with subtle linguistic cues or in domains that require specialised knowledge. By dissecting and interpreting subjective information, such as patient-generated data from social media platforms, online forums and electronic health records, sentiment analysis can provide a unique lens through which to examine the complexity of emotions in FM.

In this proof-of-concept study, we investigated whether a local LLM-driven sentiment analysis might also catch the nuance of pain expression patterns in FM due to specific lexicon by analysing linguistic patterns and emotional cues.

METHODS

This retrospective study analysed consecutive patients with classified FM according to the 2016 American College of Rheumatology (ACR) classification criteria[6] who were referred for the first time to the outpatient clinics of a tertiary care Rheumatology Unit. Between 8 January 2024 and 20 January 2024, patients referred for the first time at the same outpatient clinic, having any other type of pain were considered as controls. Patients with FM associated with any systemic autoimmune disease or other conditions were excluded. During each visit, patients were asked in plain Italian: 'How is your pain, how do you sleep at night?'.

The response was recorded using the dictation function in Microsoft Word 365 (Microsoft, Washington, USA). An expert rheumatologist in diagnosing and treating FM, unaware of the patient's condition, made a diagnosis after the visit for each enrolled patient, eventually supported by the necessary investigations. For all patients with FM, we recorded Widespread Pain Index[6] and Symptom Severity Score,[6] together with demographic characteristics.

For the analysis, we exclusively enrolled patients who had received a clinical diagnosis of FM and fulfilled the 2016 ACR classification criteria.[6]

Coding environment

The transcript was analysed in Python V.3.9.0 by the open-source LLM Mistral-7B-Instruct-V.0.2.[7] The Mistral-7B-Instruct-V.0.2 is an improved version of the Mistral-7B-Instruct model, fine-tuned to better follow instructions. This model is part of the offerings from Mistral AI (https://mistral.ai/) and is available on the Hugging Face platform (https://huggingface.co/) with Apache 2.0 licence. It was launched locally on LM Studio V.0.2.14 (https://lmstudio.ai/) and runs on an Apple Silicon M1 Max with 64 GB RAM. LM Studio runs LLMs on the laptop entirely offline. Models can be deployed mimicking an OpenAI-compatible local server.

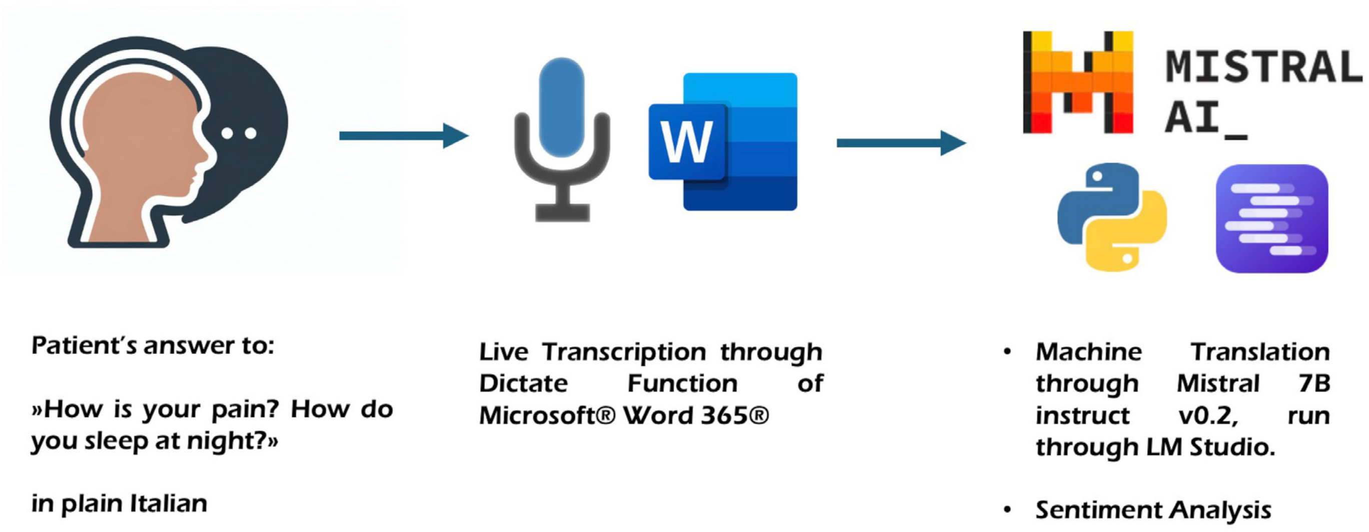
Sentiment analysis

A machine translation of the transcripts to English was first performed through the mentioned LLM (for the exact prompt, see online supplemental material). To ensure the subtleties and nuances of terminology were kept, the translation was reviewed by two English-speaking rheumatologists with substantial work experience within the UK. To distinguish patients with FM from those without, sentiment analysis was then initiated with 'prompt-engineering' (referred to as 'prompt-engineered sentiment analysis'), asking the LLM to focus on the pattern of pain expression by giving label (FM/NoFM) to each transcript (figure 1; for the exact prompts and transcripts, see online supplemental material). To test the contribution of the pattern of pain expression, we also ran the analysis with a prompt that did not focus on patterns of pain expression, simulating an ablation study (referred to as 'ablated' sentiment analysis).

Furthermore, to gain insights into the linguistic features driving the model's predictions, we conducted an attention weight analysis. Attention weights indicate the importance the model assigns to each input token when making its classification decision. By examining the attention patterns, we can identify the specific words or phrases that the model focuses on when distinguishing between FM and non-FM patients.

We extracted the attention weights from the last layer of the model for each input sentence. The attention weights were then averaged across all attention heads to obtain a single attention score for each token. To visualise the most important tokens, we selected the top 10 tokens with the highest average attention scores for each sentence.

To facilitate interpretation, we excluded certain tokens from the analysis, such as special tokens (eg, '(INST)', '(/INST)') and tokens from the instructional text used to prompt the model. However, we included the tokens 'widespread' and 'pain' in the analysis, even if they appeared in the instructional text, due to their potential relevance to FM.



Patient's answer to:

»How is your pain? How do you sleep at night?»

in plain Italian

Live Transcription through Dictate Function of Microsoft® Word 365®

- Machine Translation through Mistral 7B instruct v0.2, run through LM Studio.
- Sentiment Analysis

Figure 1 Workflow for sentiment analysis.

The attention weight analysis was performed using the HuggingFace.co Transformers application programming interface (API) V.4.39.3. The results were visualised using the library Matplotlib V.3.8.4, with each sentence represented as a separate subplot in a grid layout (online supplemental figures 1–10). The subplots displayed the top 10 tokens and their corresponding attention scores as horizontal bar charts. The tokens were rotated and labelled for readability, and the subplot titles included the sentence number and the associated diagnosis for context.

We considered the rheumatologist's diagnosis the ground truth, comparing the output of the LLM-based sentiment analysis. Hence, we plotted receiver operating characteristic curves (ROC), and the area under ROC (AUROC) was determined. Classifiers' performance was compared with the following metrics:

- Accuracy = $\frac{true\ positives+true\ negatives}{true\ positives+true\ negatives+false\ positives+false\ negatives}$,
- $\text{Recall}\left(\text{Sensitivity}\right) = \frac{true\ positives}{true\ positives+false\ negatives}$,
- Precision = $\frac{true\ positives}{true\ positives+false\ positives}$,
- Specificity = $\frac{true\ negatives}{true\ negatives\ +\ false\ positives}$

McNemar's test was also used to compare the accuracy of prompt-engineered and ablated approaches.

For each sentiment analysis approach (prompt-engineered and ablated), a univariate logistic regression model was used to examine whether a particular diagnosis was associated with misclassification as FM. The outcome variable was misclassified by the sentiment analysis model (yes/no). The predictor variable was the patient's diagnosed condition. This analysis investigated if certain conditions, based on their symptomatic and clinical overlaps with FM, posed greater challenges for accurate classification by the sentiment analysis approaches.

The analysis was conducted according to the MI-CLAIM checklist[8] (online supplemental table 2).

RESULTS

A total of 80 patients were analysed, 40 of whom were classified with primary FM according to 2016 ACR criteria,[6] with a median WPI of 18 (IQR 12–19) and a median SSS of 10 (IQR 6–12). The mean age was 48.22±9.46 years, and females were the most prevalent (38/40, 95.00%) in the cohort. FM patients' details have been reported in online supplemental table 1. Other diagnoses included psoriatic arthritis (PsA, 10/40, 27.50%), subacromial bursitis in calcifying tendinopathy (5/40, 12.5%), axial spondyloarthritis (AxSpA, 4/40, 10%), hand and/or knee osteoarthritis (4/40, 10%), rheumatoid arthritis (RA, 4/40, 10%), spinal stenosis (3/40, 7.50%) and other miscellaneous conditions (21/40, 50.00%) reported in table 2online supplemental table 1. None of such patients met the ACR2016 criteria[6] for FM nor had a previous or concomitant clinical diagnosis of FM. The machine translation run flawlessly and with manual adjustments needed.

The prompt-engineered sentiment analysis had an accuracy of 0.87, a precision of 0.92, a recall of 0.84 and a specificity of 0.82. The AUROC for the prompt-engineered sentiment analysis was 0.86 (Confusion matrix in figure 2, left panel).

The patients misclassified with FM (false positives) for the latter approach were 7/40 (17.50%): n.3 patients with subacromial bursitis due to calcifying tendinopathy (7.50%), n.1 patient with AxSpA (2.50%), n.1 patient with DeQuervain's tenosynovitis (2.50%), n.1 patient with RA (2.50%), n.1 patient with spinal stenosis (2.50%).

On the other hand, the ablated sentiment analysis had an accuracy of 0.76, a precision of 0.75, a recall of 0.77 and a specificity of 0.75. The AUROC for the ablated model was 0.76 (Confusion matrix in figure 2, right panel).

McNemar's test showed that the prompt-engineered sentiment analysis had superior accuracy compared with the ablated approach (χ^2=16.57, p<0.001, figure 3).

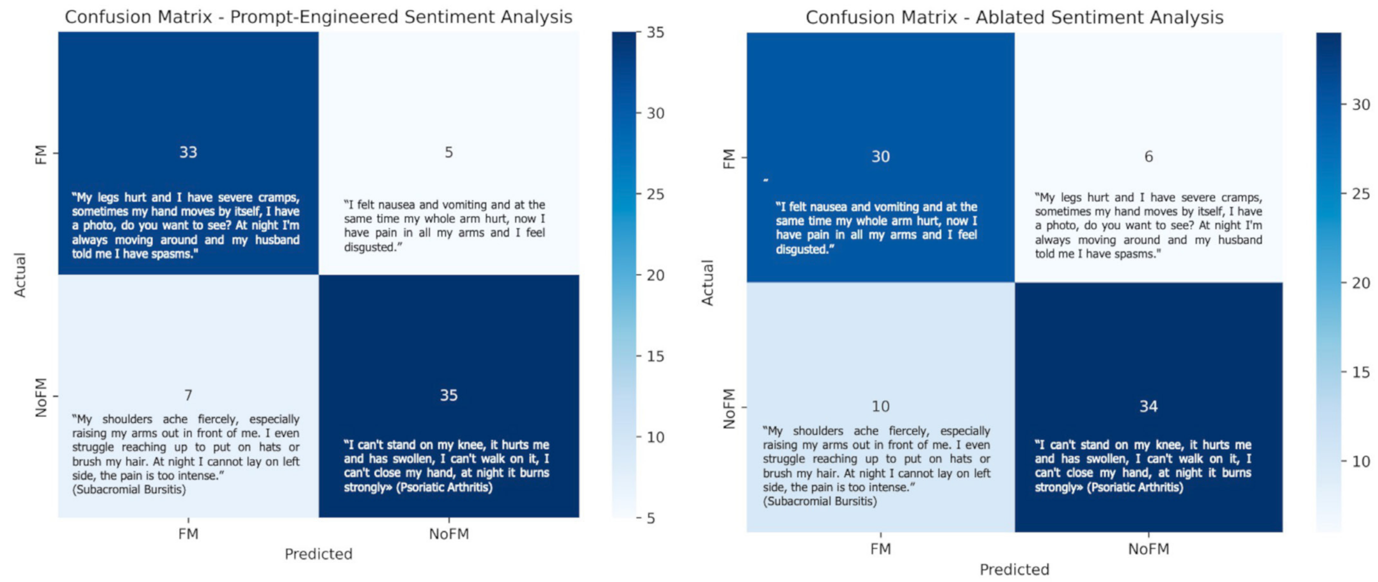


Figure 2 Left panel: Confusion matrix for prompt-engineered sentiment analysis. Right panel: Confusion matrix for ablated sentiment analysis. The number of true positives decreased while false positives increased with ablated sentiment analysis. FM, fibromyalgia.

The patients misclassified with FM (false positives) for the ablated approach were 10/40 (25,00%): n.2 patients with AxSpA (5.00%), n.2 patients with subacromial bursitis due to calcifying tendinopathy (5.00%), n.2 patients with spinal stenosis (5.00%), n.1 patients with PsA (2.50%), n.1 patients with RA (2.50%), n.1 patient with DeQuervain's tenosynovitis (2.50%), n.1 patient with idiopathic transient osteoporosis of the hip (2.50%).

Patients with subacromial bursitis due to calcifying tendinopathy had a higher likelihood of being misclassified as having FM by the prompt-engineered sentiment analysis, with an OR of 29.57 (95% CI 2.70 to 323.69). However,

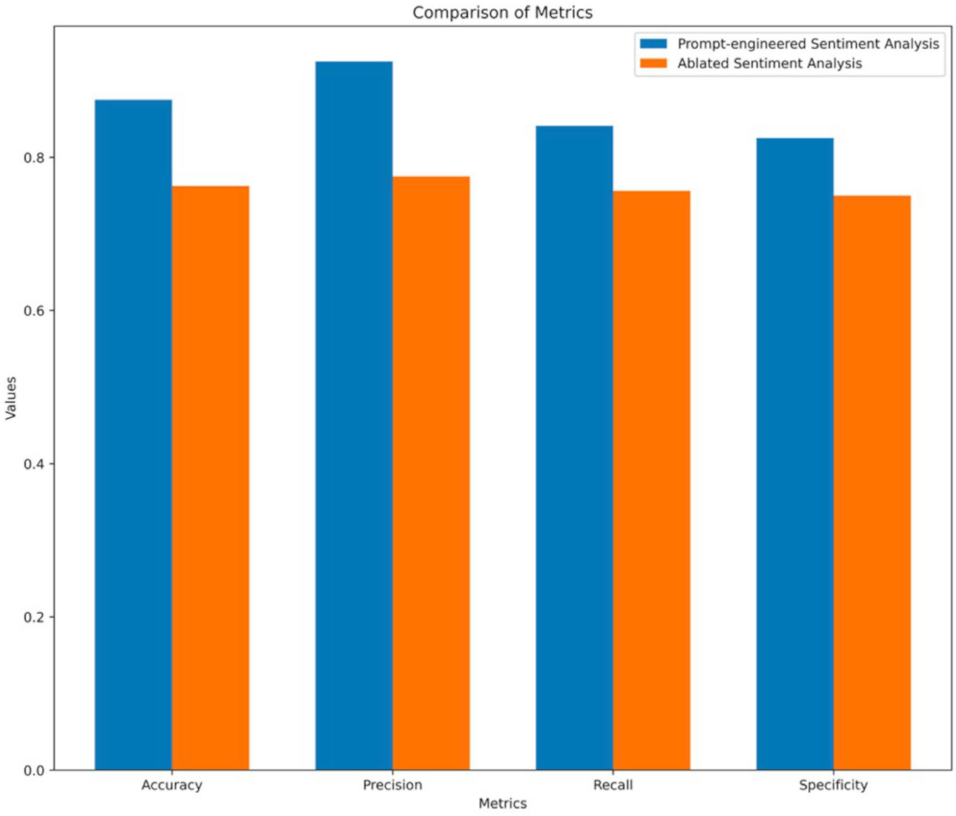


Figure 3 Performance of the prompt-engineered sentiment analysis versus the ablated sentiment analysis.

patients with other diagnoses, including AxSpA and PsA, did not demonstrate a significantly greater probability of misclassification as FM by either the prompt-engineered or ablated sentiment analysis approaches (data are not shown),

The attention weight analysis demonstrated that the prompt-engineered sentiment analysis model highlighted gave a notable emphasis on words associated with widespread pain, fatigue, depressed mood and dysesthesia, such as 'everywhere', 'spot' (used to communicate a 'leopard-spot' pain), 'exhaust', 'depressed', 'electric', 'burning' (online supplemental figures 1–10).

DISCUSSION

This proof-of-concept study demonstrates the potential of using open-source LLM-driven sentiment analysis to aid in the diagnosis of FM. The high accuracy, precision and recall of the prompt-engineered sentiment analysis suggest it could effectively distinguish between patients with and without FM based on subtle differences in their expression of pain symptoms.

These findings have meaningful implications for clinical practice. By detecting linguistic and emotional cues that may reflect the central sensitisation and negative effects associated with FM, sentiment analysis could help clinicians overcome some of the challenges in diagnosing FM.

All key accuracy metrics of LLM-driven sentiment analysis showed sizeable gains with the incorporation of prompts designed to pick up on nuanced elements of how FM patients verbalise their pain experiences. The attention weight analysis of the prompt-engineered sentiment analysis model provided valuable insights into the language nuances associated with FM. The model consistently assigned high attention weights to words related to widespread pain, fatigue, depressed mood and dysesthesia, accurately capturing the key linguistic markers of FM. These findings highlight the model's ability to detect the subtle language patterns that distinguish FM from other chronic pain conditions. The targeted prompt engineering played a crucial role in guiding the model to focus on these specific linguistic features, enhancing its sensitivity and accuracy in FM classification. The analysis demonstrates the potential of sentiment analysis in developing novel tools for FM diagnosis and understanding the patient experience. These results underscore the importance of considering language nuances in the development of accurate and interpretable models for complex conditions like FM.

This approach offers several distinct perspectives with potential benefits. On the one hand, LLM-driven sentiment analysis may aid in accurately identifying patients who should be referred from primary care to rheumatology for evaluation of possible FM. This study demonstrates the potential for LLM-driven sentiment analysis to aid in accurately identifying patients who should be referred from primary care to rheumatology for evaluation of possible FM. The high precision and recall metrics suggest the tool could become an asset for general practitioners (GPs) encountering patients with diffuse chronic pain complaints. Providing GPs with more objective, data-driven insights could overcome inherent challenges in distinguishing early FM from other benign pain conditions at the primary care level. This could allow timely specialist assessment and initiation of appropriate treatment. In this regard, the LLM-driven sentiment analysis approach presented in this study could serve as the foundation for developing tools that integrate seamlessly with electronic medical records (EMRs). By incorporating such tools directly into the EMR system, GPs could access the sentiment analysis results alongside other patient data, facilitating a more comprehensive assessment of the patient's condition. This integration would enable GPs to make more informed decisions about referrals and management plans, potentially improving patient outcomes and reducing healthcare costs associated with delayed or misdiagnosed FM cases.

Furthermore, the sentiment analysis approach could be extended to incorporate speech recognition systems like Whisper (https://openai.com/research/whisper), allowing for direct spoken sentiment analysis. This could be particularly valuable in multilingual settings, where open-source LLMs trained in various languages could be employed. Such an implementation would enable real-time analysis of patient narratives, providing GPs with immediate insights and support during consultations.

However, if such tools are directly provided to patients, there are risks of inappropriate self-diagnosis and self-management. Patient-entered data would lack the context of a formal clinical evaluation. There may be overinterpretation of chronic regional pain as widespread centralised pain, unclear case definitions between chronic fatigue and FM fatigue, and assumptions that negative sentiments equate to an FM diagnosis even when clinical criteria are not met. This could propagate myths about FM, undermine provider–patient relationships and promote untargeted therapy.

On the other hand, by using an open-source LLM like Mistral locally, healthcare providers may analyse patient data more ethically and responsibly than relying on third-party APIs. Running the model on a local machine rather than sending data to an external service allows healthcare providers to have full control and visibility into how patients' private health information is being used.

For an innovative application like LLM-powered sentiment analysis, assuring data protection and oversight is key to ethical adoption. By keeping both the data and model on-site, healthcare providers can analyse sensitive information fairly and safely.

One limitation of the LLM-driven sentiment analysis is that it may miss some cases of inflammatory pain conditions that have overlapping features with FM. For example, AxSpA, PsA and subacromial bursitis due to calcifying tendinopathy may present with diffuse chronic pain and negative emotions. A comprehensive clinical assessment that includes history, physical examination, laboratory tests and eventual imaging is still necessary to rule out alternative diagnoses and confirm FM. On the other hand, the lower false positive rate for AxSpA and PsA, for which FM is a known confounder, suggests prompt engineering specifically helped the sentiment analysis model better recognise key differences in the pain quality, fatigue, stiffness, emotional distress and other symptoms between these diseases. This may allow more precise classification that could facilitate appropriate specialist referral and management.

An additional limitation of this study is that we exclusively enrolled patients with primary FM. Patients with autoimmune disorders who may also have secondary FM manifestations were omitted. However, in real-world clinical settings, there is frequently an overlap between FM and other chronic pain-predominant conditions.[9] Hence, there is a need for future studies to assess whether LLM-driven sentiment analysis could aid in detecting the central sensitivity syndrome of FM amidst the complex pain presentations of comorbid diseases. Expanding the test cohort to include secondary FM patients with conditions like, for instance, PsA who have concomitant widespread pain is an essential next step on the research agenda.

The small sample size, with only 40 FM patients and the retrospective design warrants caution in generalising the accuracy metrics. Additionally, while prompt engineering was used to target keywords related to FM pain experiences, the semantics may not fully capture the complex psychosocial aspects of living with chronic widespread pain because of interindividual differences in pain expression due to factors like culture, demographics, comorbidities and traits, especially alexithymia.

To fully realise the potential of LLM-driven sentiment analysis in FM diagnosis, validation studies in international cohorts are necessary. These studies should include patients from diverse cultural and linguistic backgrounds to assess the generalisability and robustness of the approach. Additionally, future research should explore the integration of sentiment analysis with other clinical data and patient-reported outcomes to develop comprehensive diagnostic support tools for FM.

In conclusion, this study provides early evidence that LLM-driven sentiment analysis could be a useful tool to complement clinical assessment in diagnosing complex conditions like FM. Further validation in larger prospective cohorts is warranted. Additionally, optimising model interpretability and integrating findings with patient-reported outcomes data could help translate these analytics into clinical impact for patients.

X Vincenzo Venerito @vincevenerito

Contributors VV and FI conceived of the presented idea. VV developed the theory and performed the computations. VV and FI verified the analytical methods. FI supervised the findings of this work. FI is responsible for the overall content as guarantor. All authors discussed the results and contributed to the final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Consent obtained directly from patient(s).

Ethics approval This study involves human participants and was approved by (Biopure registry, IRB Approval no. 5940, Azienda Ospedaliera Universitaria di Bari). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Vincenzo Venerito http://orcid.org/0000-0002-2573-5930
Florenzo Iannone http://orcid.org/0000-0003-0474-5344

REFERENCES

1 Galvez-Sánchez CM, Duschek S, Reyes Del Paso GA. Psychological impact of fibromyalgia: current perspectives. *Psychol Res Behav Manag* 2019;12:117–27.
2 Pidal-Miranda M, González-Villar AJ, Carrillo-de-la-Peña MT. Pain expressions and inhibitory control in patients with fibromyalgia: behavioral and neural correlates. *Front Behav Neurosci* 2018;12:323.
3 Denecke K, Reichenpfader D. Sentiment analysis of clinical narratives: a scoping review. *J Biomed Inform* 2023;140:104336.
4 Venerito V, Gupta L. Large language models: rheumatologists' newest colleagues *Nat Rev Rheumatol* 2024;20:75–6.
5 Venerito V, Bilgin E, Iannone F, *et al*. AI am a rheumatologist: a practical primer to large language models for rheumatologists. *Rheumatology (Oxford)* 2023;62:3256–60.
6 Wolfe F, Clauw DJ, Fitzcharles M-A, *et al*. Revisions to the 2010/2011 fibromyalgia diagnostic criteria. *Semin Arthritis Rheum* 2016;46:319–29.
7 Jiang AQ, Sablayrolles A, Mensch A, *et al*. Mistral 7B, 2023. Available: https://doi.org/10.48550/arXiv.2310.06825
8 Norgeot B, Quer G, Beaulieu-Jones BK, *et al*. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 2020;26:1320–4.
9 Iannone F, Nivuori M, Fornaro M, *et al*. Comorbid Fibromyalgia impairs the effectiveness of biologic drugs in patients with psoriatic arthritis. *Rheumatology (Oxford)* 2020;59:1599–606.