

Quantifying Patient Narratives to Strengthen Longitudinal Assessment in Chronic Pain

Ishaan Sinha¹, Arefeh Sherafati²

¹The Pingry School, Basking Ridge, NJ, 07920, USA

²Department of Psychology & Biobehavioral Sciences, St. Jude Children's Research Hospital, Memphis, TN, 38105, USA

KEYWORDS. narrative medicine, aspect-based sentiment analysis, chronic pain assessment, clinical natural language processing, wellness index

BRIEF. This study presents a computational framework that integrates patient narratives with symptom scales using aspect-based sentiment analysis to generate an interpretable longitudinal wellness index for clinical assessment.

ABSTRACT. Chronic pain assessment relies heavily on patients' descriptions of their symptoms, yet these narrative reports are difficult for clinicians to interpret consistently and are rarely incorporated into structured monitoring tools. The purpose of this research was to develop a computational system that translates patient narratives into quantitative indicators that can be tracked over time alongside standard symptom ratings. To do this, we designed a pipeline that uses natural language processing, specifically aspect-based sentiment analysis, to identify meaningful themes in patients' written descriptions and classify them using the World Health Organization's International Classification of Functioning (ICF). These sentiment-based measures were then combined with normalized 0–10 symptom scales to create a single Wellness Index ranging from 0 to 100. We evaluated the method using a synthetic dataset modeled on real fibromyalgia narratives and a six-month, 50-entry longitudinal case. The system accurately identified functional themes and emotional tone in narratives, showing strong agreement with human reviewers, and produced a stable index that reflected realistic patterns of symptom flare-ups and recovery. The study demonstrates that the proposed framework functions as intended. It reliably extracts measurable signals from patient narratives and integrates them with symptom scales into a stable, clinically plausible Wellness Index. This approach offers a transparent, interpretable tool for both patients and clinicians.

INTRODUCTION.

Symptom self-reporting remains a cornerstone of modern medical practice, particularly in conditions lacking objective biomarkers such as chronic pain, fatigue, and mental health disorders [1]. Yet subjective reports are shaped by memory limits, emotional framing, and cultural norms, creating variability that contributes to misdiagnosis, delays in care, and challenges in longitudinal monitoring [1–4]. Despite these limitations, patient narratives contain rich clinical signals that remain underutilized in structured workflows [5,6].

Recent advances in natural language processing (NLP) provide new opportunities to extract value from narrative data. Narrative features, such as adjectives, emotional terms, and contextual details correlate with pain intensity and symptom severity, showing that subjective language can be transformed into structured data [5,6]. NLP tools also extract granular symptom information from electronic health records, demonstrating the feasibility of automating interpretation of subjective reports [6]. These developments support broader calls for integrative models of pain and wellness that capture biopsychosocial dimensions beyond numeric scales [7,8].

Patient narratives embody lived experience. Qualitative narratives show recurring themes of loss, adaptation, emotional burden, and identity change, dimensions rarely reflected in quantitative measures [4]. Many patients describe their own words as the most accurate representation of their illness experience [3,8]. Narratives thus function

as both self-expression and clinically meaningful sources of diagnostic and prognostic insight. Chronic pain is particularly challenging because it persists without clear biomedical markers, leaving patients and clinicians to navigate uncertainty and invisibility. Lived experiences often involve emotional disruption, identity shifts, strained relationships, and ongoing struggle [4]. Unlike acute illness, chronic pain frequently does not follow conventional recovery trajectories, producing continuous “chronicling” of open-ended suffering [8].

Computational tools now offer pathways to integrate this narrative complexity into structured assessment. Elyazori et al. [9] used motivational interviewing combined with NLP and Aspect-Based Sentiment Analysis (ABSA) to translate chronic pain narratives into ICF-aligned quantitative indicators. Ettlin et al. [10] applied topic modeling and lexicon analysis to large-scale orofacial pain narratives, uncovering symptom clusters, emotional distress, and unmet patient needs. These studies demonstrate the scalability and clinical utility of computational approaches to patient language. Large language models (LLMs) expand these capabilities further. Venerito and Iannone [11] showed that an LLM-based sentiment classifier could distinguish fibromyalgia from other chronic pain conditions with >85% accuracy using only narrative tone and linguistic nuance. This highlights how subtle affective patterns can serve as diagnostic signals and reinforces the importance of integrating sentiment with symptom metrics.

Despite this progress, a major gap remains: current tools rarely combine narrative analysis with quantitative symptom tracking in an interpretable, longitudinal framework [4,6]. Symptom tracking apps typically emphasize numerical ratings, while journaling platforms capture rich narrative without converting it into insights [8]. Consequently, patients lack tools to observe their wellness over time, and clinicians lack ways to incorporate emotional and contextual information into decision-making [1,2].

This paper addresses that gap by asking: Can ABSA reliably extract and quantify clinically meaningful signals from patient narratives? Can a system which utilizes ABSA of patient narratives and integrates quantitative symptom scales produce a reliable and interpretable longitudinal wellness index for chronic pain assessment? We propose a framework that merges narrative text with symptom scales into a Composite Wellness Index, paired with an interpretable breakdown of contributing factors, specifically, the Aspect Sentiment Subscore and the Numeric Scale Subscore. By quantifying subjective experience while preserving narrative nuance, this system aims to reduce recall bias, support clinical reasoning, empower patient self-insight, and provide a richer longitudinal perspective on symptom progression.

MATERIALS AND METHODS.

Study Design. We developed a computational pipeline that converts open-ended patient narratives and 0–10 symptom scales into aspect-level International Classification of Functioning (ICF) labels with sentiment and a single longitudinal Wellness Index (0–100) (Figure 1). The scheme follows an ICF + ABSA approach. Text spans, the smallest contiguous segments of patient's narrative that convey a distinct

idea, symptom, or experience, are identified. The span is then labeled with second-level ICF code. The ICF is a framework developed by the World Health Organization (WHO) to provide a standardized and globally accepted system for describing health, functioning, and disability [12] (Table S1). For example, ICF codes can represent domains such as pain (b2), mobility (d4) or support/relationships (e3), allowing narrative text to be mapped onto clinically meaningful categories. After being labeled with an ICF code, the span is then assigned sentiment (positive/neutral/negative) by an LLM.

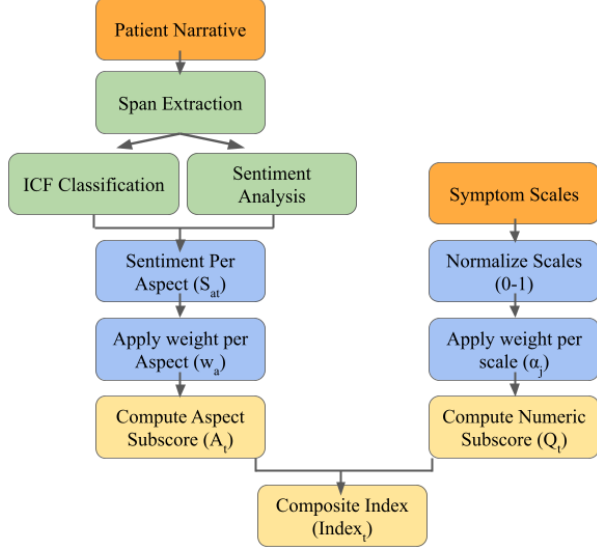


Figure 1. This diagram depicts the full computational pipeline, showing how raw patient narratives and numeric symptom ratings are transformed into ICF-coded sentiments, normalized symptom scores, and the final index. Input (orange), LLM computing (green), computing (blue), and results (yellow) are demonstrated.

Composite Wellness Index. The Primary output of this system is the Composite Wellness Index (0-100), a single interpretable score that reflects a patient’s overall well-being at a given time. To produce it, the model first maps ICF-coded narrative spans to labels, then applies sentiment. These are then converted mathematically into the Aspect Sentiment Subscore. Quantitative Scales are converted mathematically into the Numeric Scale Subscore. Both individual subscores are combined into the final Composite Wellness Index.

Aspect Sentiment Subscore. For each narrative span, the model generates two outputs: an ICF label (Table 1), which corresponded to one of the predefined categories from the ICF, and a sentiment label, which classifies the text as positive, neutral, or negative. We prompt GPT-4o with a strict, JSON-only instruction to (1) extract aspect spans, (2) assign exactly one ICF label per span, and (3) assign sentiment. The goal of this configuration is to ensure consistent, reproducible outputs across narratives. To achieve this, parameters were empirically tuned for maximizing reproducibility. The temperature was set to 0.1 to minimize randomness and to keep responses close to the model’s likely interpretation. The top-p parameter which controls randomness by letting the model choose the next token only from the smallest set of options whose combined probability was set to 0.8. GPT-4o was selected for its strength in semantic coherence, clinical text comprehension, and multi-aspect reasoning. Prior benchmarks have shown GPT-4o’s superior performance on clinical NLP and ABSA tasks [13,14].

Each span’s sentiment is mapped to a numeric value: positive = +1, neutral = 0, negative = -1. We average the span scores to get the aspect sentiment $S_{a,t} \in [-1, 1]$ (Eq. 1), per entry t and per aspect a (ICF label). N_a is the number of narrative text spans mapped to aspect a . s_k is the k th narrative span (+1, 0, -1).

$$S_{a,t} = \frac{1}{N_a} \sum_{k=1}^{N_a} s_k \quad (1)$$

Calculating Aspect Sentiment Subscore (A_t). This subscore quantifies the emotional tone expressed solely in the patient’s own words, capturing how patients narratively describe their health across ICF functional domains. It reflects the qualitative, language derived dimension of the Wellness Index. $\frac{S_{a,t}+1}{2}$ rescales $[-1, 1]$ to $[0, 1]$, aligning it with standardized scales. Aspect weights (w_a) prioritize pain-proximal categories (e.g., b2, b7, d4). Clinically, A_t captures how patients feel about their health, providing context on emotion and function beyond numeric pain scores (Eq. 2). $\mathcal{A}$ is the set of all ICF aspects.

$$A_t = \sum_{a \in \mathcal{A}} w_a \cdot \frac{S_{a,t} + 1}{2} \quad (2)$$

Numeric Scale Subscore. The Numeric Scale Subscore represents the objective symptom burden derived entirely from standardized 0–10 scales (pain, fatigue, sleep, function). Each scale is min–max normalized to $[0, 1]$. For symptoms where higher values indicate worse severity (e.g., pain, fatigue), the normalized score is reversed so that higher values consistently reflect better health. Symptoms where higher values already indicate improvement (e.g., sleep quality) are left as-is. This ensures all symptom scales align directionally before being integrated into the composite index.

Calculating Numeric Scale Subscore (Q_t). Here, $x_{t,j}$ represents normalized 0–1 values from symptom scales (pain, fatigue, sleep, mood, function), and $\sum_j \alpha_j = 1$. The weights (α_j) reflect the relative importance of each clinical measure. j is the index over numeric symptom scales. Q_t retains objectivity and comparability (Eq. 3). It translates conventional symptom metrics into a standardized contribution toward the overall index.

$$Q_t = \sum_j \alpha_j \cdot x_{t,j} \quad (3)$$

Calculating Composite Index ($Index_t$). The final index combines the narrative and numeric components, scaling the result from 0–100 for interpretability, where scores closer to 0 indicate greater symptom burden and scores closer to 100 show lower symptom burden. A default $\beta = 0.5$ assigns equal weight to both sources of data but this can be tuned. In psychological or chronic pain settings, a higher β gives more influence to patient narratives. Clinically, β allows for personalization. Patients who express more through narrative text can have their qualitative data weighted more heavily. Quantitative-heavy users maintain robust comparability. The result is a transparent, adaptive index that tracks both symptom severity and emotional recovery over time (Eq. 4).

$$Index_t = 100 \cdot [\beta \cdot A_t + (1 - \beta) \cdot Q_t] \quad (4)$$

Synthetic Datasets. We evaluated this framework using two synthetic cohorts derived from qualitative fibromyalgia (FM) literature. FM was selected because it lacks definitive biomarkers and relies heavily on subjective self-report, making it an ideal test case for narrative-based assessment. The first cohort consisted of 5 synthetic FM patients, each generating paired free-text narratives and 0–10 symptom ratings grounded in published FM narrative studies. The narratives encapsulated unpredictable widespread pain and fatigue, healthcare and diagnostic struggles, treatment frustration, coping and self-management, and work/role loss [15–18]. This first cohort was used to evaluate ICF label classification performance. The second cohort consisted of 50 narrative–numeric entries from a single synthetic FM patient over a simulated six-month period. This second cohort was used to evaluate longitudinal system–level performance and ensure the Wellness Index produced clinically plausible trends.

Sentiment Analysis. Aspect-based sentiment analysis was performed using GPT-4o, and a subset of outputs was independently hand-annotated by two human reviewers.

Evaluation. Performance was evaluated using precision (proportion of model-assigned ICF labels that matched human annotation), recall (proportion of human-annotated labels the model successfully identified), and the F1-score (harmonic mean of precision and recall). Cohen’s κ was used to assess agreement. For the purposes of initial evaluation, given the limited number of examples per ICF label in the synthetic dataset, both misclassified labels and missing labels were treated as false negatives. This approach ensured that performance metrics reflected the model’s ability to correctly identify subtle or low-frequency aspects rather than overestimating accuracy due to sparse class representation. For system-level evaluation, one synthetic FM patient generated 50 entries over a simulated six-month period. The computational pipeline was executed with $\beta = 0.7$ and equal aspect weights.

RESULTS.

LLM ICF Label Classification Performance. Evaluation focused on the accuracy of aspect-level ICF code assignment and sentiment detection across the five-patient synthetic FM cohort. Narrative spans were manually annotated and compared against GPT-4o outputs. This evaluation establishes a performance benchmark for the system’s ability to reliably classify functional aspects and sentiment, ensuring that all downstream analyses of the Wellness Index are grounded in validated ICF mappings. From these ground truth annotations, we derived precision, recall, F1-score, and Cohen’s κ (Table 1), supplemented by qualitative inspection of the confusion matrix (Figure 2).

Table 1. This table reports the precision, recall, F1-score, and Cohen’s κ based on comparison with human-annotated narrative spans.

Precision	0.887
Recall	0.825
F1 Score	0.855
Cohens κ	0.673

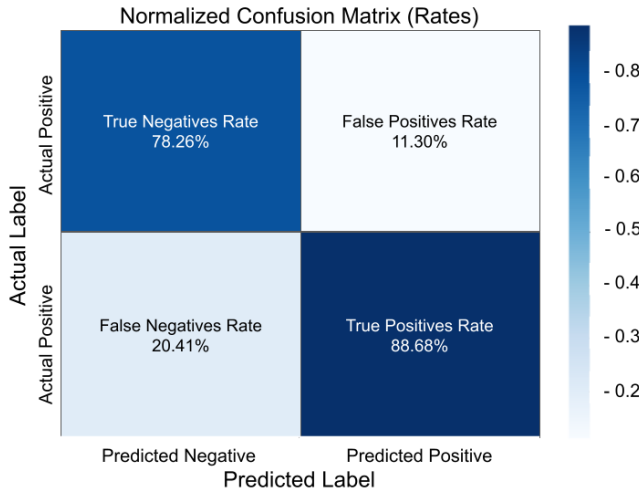


Figure 2. The confusion matrix visualizes agreement between human-annotated ICF labels and the model predictions across all narrative spans.

These results indicate strong agreement between automated and human annotation. Qualitative review of the confusion matrix shows that performance was strongest for high-frequency, clinically salient aspects such as pain (b2), mobility (d4), and mental functions (b1). Precision (0.887) exceeded recall (0.825), suggesting the model was conservative in assigning labels, more likely to miss subtle spans than to hallucinate incorrect ones. Cohen’s κ of 0.673 reflects substantial agreement. Most false negatives arose from indirect metaphors (e.g.,

“numbers danced” → cognitive dysfunction), identity-oriented language (e.g., “old me”), and social-stigma cues embedded in humor or comparison. False positives primarily reflected over-generalization of environmental factors and over-labeling of ambiguous emotional spans.

System-Level Evaluation: Longitudinal Composite Wellness Index Performance. To assess whether the system could produce a clinically meaningful longitudinal wellness signal, the full pipeline was applied to 50 entries from a single synthetic FM patient over a six month period. When applied to the 50-entry single-patient cohort, the system produced stable, interpretable values across all three components of the Wellness Index: Aspect Sentiment Subscore (A_t), Numeric Scale Subscore (Q_t), and the Composite Index ($Index_t$). Across the 50 entries, $Index_t$ values ranged from 5.1 to 82.1, capturing periods of both high symptom burden and intervals of relative improvement. Mean values for the three components were highly consistent: $A_t = 51.3$, $Q_t = 39.2$, and $Index_t = 46.8$, indicating that the narrative-derived sentiment signal and the symptom-scale signal contributed comparably to overall wellness estimates. As expected in FM-like trajectories, entries displayed substantial within-subject variability, with alternating periods of symptom exacerbation and recovery (Figure 3). The system successfully preserved these fluctuations without introducing volatility, demonstrating that the weighting structure and β -mixing parameter ($\beta = 0.7$) produced a smooth but responsive longitudinal curve. These trends are visualized, which plots A_t , Q_t , and $Index_t$ over time and illustrates the alignment, and occasional divergence, between narrative sentiment, numeric symptom ratings, and the resulting composite assessment. High-impact episodes corresponded with the lowest index scores (e.g., Entry 17: $Index_t = 5.1$), while entries with positive emotional framing and improved numeric indicators produced the highest values (e.g., Entry 25: $Index_t = 82.1$). Collectively, these results demonstrate that the composite index behaves in a clinically interpretable manner, integrates narrative and numeric data without dominance by either modality, and captures the complex structure central to chronic pain trajectories.

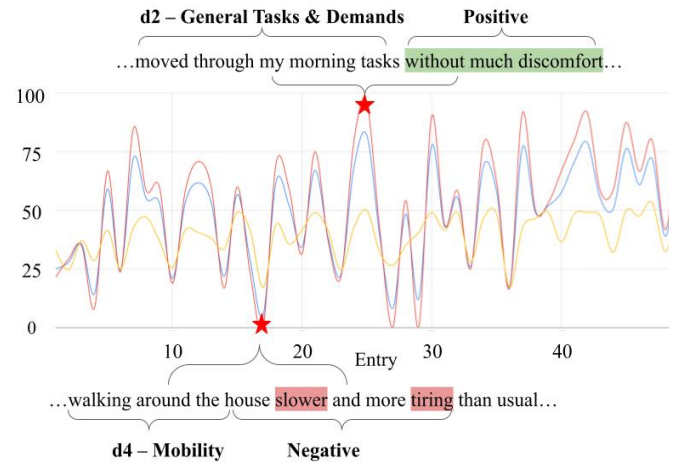


Figure 3. This figure plots the Aspect Sentiment Subscore (red), Numeric Scale Subscore (yellow), and the Composite Wellness Index (blue), across 50 entries over 6 months. The trajectories illustrate how narrative sentiment and symptom ratings interact over time.

DISCUSSION.

This study developed and evaluated a computational framework that integrates aspect-based sentiment analysis of patient narratives with numeric symptom scales to produce a longitudinal Composite Wellness Index. Across a five-patient synthetic FM cohort, the system achieved strong ICF classification performance (precision = 0.887, F1 = 0.855, $\kappa = 0.673$), and when applied to a single patient’s 50-entry record, the Composite Wellness Index produced clinically plausible

trajectories. Narrative-based computational assessment can be systematically integrated with traditional symptom scaling to produce a longitudinal, interpretable representation of patient wellness. Importantly, the system performed particularly well for clinically salient, high-frequency domains such as pain, fatigue, mobility, and mental functions. Patient narratives contain consistent and extractable signals relevant to chronic pain management. Subjective text does not need to be treated as inherently unreliable when analyzed computationally. It becomes measurable, interpretable, and clinically useful data [5,10,6,11]. Equally meaningful is the system-level behavior of the longitudinal Wellness Index across the single patient's 50-entry record. The index successfully reflected expected chronic pain trajectories, distinguishing periods of exacerbation and relative relief without over smoothing or producing unrealistic volatility. The close alignment between the narrative-derived sentiment subscore (A_t) and numeric symptom subscore (Q_t) confirms that each modality contributes complementary information. Emotional tone, coping behaviors, and contextual stressors emerged in the narrative signal. Integrating narrative sentiment with numerical scales captures aspects of the patient experience that neither modality alone can fully represent.

This novel framework also identifies important directions for future work. The Composite Wellness Index holds potential as an outcomes measurement tool in research settings and as a structured evidence base for insurance documentation and treatment authorization. While this initial evaluation used synthetic FM narratives to establish feasibility and methodological soundness, clinical deployment will require validation on real-world patient data. The next step is therefore to test the system in a real clinical setting, ideally through an observational pilot in a pain clinic or interdisciplinary rehabilitation program. Deployment would allow assessment of actual patient uptake, clinician interpretability, and the degree to which the index meaningfully alters decision-making or communication during care encounters. Evaluating the index's sensitivity to change in response to interventions (e.g., medication adjustments, physical therapy, behavioral treatments) will also be critical.

Expanding the dataset beyond fibromyalgia-like narratives will be necessary to assess the generalizability of the pipeline across conditions where subjective experience plays a central diagnostic role, including chronic fatigue syndrome, depression, anxiety, and trauma-related disorders [3]. Increased data diversity will also enable stronger performance on metaphor, indirect language, and culturally variable descriptions of suffering. Although classification performance was strong, performance varied by aspect frequency and conceptual complexity. Future work should incorporate a per-label confusion matrix, which will allow more detailed inspection of misclassifications and help determine where fine-tuning or re-weighting are needed. This level of granularity is especially important for low-frequency ICF categories that capture social participation, environmental barriers, and identity-related constructs. These domains are clinically meaningful yet linguistically subtle.

There is substantial opportunity to enhance the clinical usefulness of the system through additional interpretability tools. This may include real-time visualization dashboards, per-aspect trajectory plots for clinicians, patient-facing summaries that contextualize changes in mood or function, or adaptive weighting schemes tailored to individual presentation patterns. Computational approach leveraging ICF-based aspect extraction and sentiment analysis can meaningfully transform subjective narratives into structured, longitudinal insight. The results highlight both the feasibility and clinical promise of integrating patient language into routine symptom tracking. With further validation in real patient populations and refinement of per-label performance characteristics, this approach has the potential to reduce recall bias, enhance shared understanding between patients and clinicians, and improve the monitoring and management of complex conditions.

SUPPORTING INFORMATION.

Table S1. Complete list of all second-level International Classification of Functioning (ICF) labels used in the study.

REFERENCES.

- 1 T. Merten, The self-report fallacy: When diagnosis predominantly relies on subjective symptom report. *Curr. Opin. Psychol.* **60**, 102096 (2025).
- 2 M. Rosendal, R. Jarbøl, K. Pedersen, P. Thorsen, Multiple perspectives on symptom interpretation in primary care research. *BMC Fam. Pract.* **14**, 167 (2013).
- 3 G. Franssen, Narratives of undiagnosability: Chronic fatigue syndrome life-writing and the indeterminacy of illness memoirs. *Philos. Psychiatry Psychol.* **27**, 403–416 (2020).
- 4 S. Van Rysewyk, G. Williams, L. K. Jones, Understanding the lived experience of chronic pain: A systematic review and synthesis of qualitative evidence syntheses. *Br. J. Pain* **17**, 592–605 (2023).
- 5 A. Nunes, A. T. Lopes, M. A. T. Figueiredo, Chronic pain patient narratives allow for the estimation of current pain intensity. arXiv:2210.17473 (2022).
- 6 A. D. Dave et al., Automated extraction of pain symptoms: A natural language approach using electronic health records. *Pain Physician* **25**, E245–E254 (2022).
- 7 T. H. Wideman et al., The Multimodal Assessment Model of Pain: A Novel Framework for Further Integrating the Subjective Pain Experience Within Research and Practice. *Clin J Pain*, **35**(3), 212–221 (2019).
- 8 F. Van Hout, A. Van Rooden, J. Slatman, Chronicling the chronic: Narrating the meaninglessness of chronic pain. *Med. Humanit.* **49**, e1–e8 (2023).
- 9 H. Elyazori et al., Capturing patients' lived experiences with chronic pain through motivational interviewing and information extraction. In *Proc. Second Workshop on Patient-Oriented Language Processing (CL4Health)*, Association for Computational Linguistics, 321–330 (2025).
- 10 D. A. Ettlin, A. R. Gallo, A. M. Lutz, "In patients' words": Natural language processing of reports from patients experiencing orofacial pain and dysfunction. *J. Headache Pain* **26**, 94 (2025).
- 11 V. Venerito, F. Iannone, Large language model-driven sentiment analysis for facilitating fibromyalgia diagnosis. *BMJ Innov.* **10**, e000994 (2024).
- 12 World Health Organization, International Classification of Functioning, Disability and Health (ICF). WHO, Geneva (2001).
- 13 B. Bicknell, A. J. DeGrauw, P. J. Krause, ChatGPT-4 Omni performance in USMLE disciplines and clinical skills: Comparative analysis. *JMIR Med. Educ.* (2024).
- 14 L. Zhang, S. Tashiro, M. Mukaino, S. Yamada, Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: A comparative test case. *J. Rehabil. Med.* **55**, jrm13373 (2023).
- 15 O. Hellström et al., A phenomenological study of fibromyalgia: Patient perspectives. *Scand. J. Prim. Health Care* **17**, 11–16 (1999).
- 16 P. Juuso, L. Skär, M. Olsson, S. Söderberg, Living with a double burden: Meanings of pain for women with fibromyalgia. *Int. J. Qual. Stud. Health Well-being* **6**, 7184 (2011).
- 17 H. K. Lempp, S. L. Hatch, S. F. Carville, E. H. Choy, Patients' experiences of living with and receiving treatment for fibromyalgia syndrome: A qualitative study. *BMC Musculoskelet. Disord.* **10**, 124 (2009).
- 18 S. C. Ashe et al., A qualitative exploration of the experiences of living with and being treated for fibromyalgia. *Health Psychol. Open* **4**, 2055102917724336 (2017).



Ishaan Sinha is a student at The Pingry School in Basking Ridge, New Jersey.