Artificial Intelligence Assisted Extraction of Patient-Reported Pain Outcomes in Osteoarthritis Using Prompt Engineering of Large Language Models

Jainesh Doshi[1], Stephen Batter[1], Yiyuan Wu[2], Alice Santilli[1] Sandhya Kannayiram[1] Susan Goodman[1] Bella Mehta[1]
[1]Hospital For Special Surgery, New York, NY [2]Weil Cornell Medicine, New York, NY

Disclosures: Jainesh Doshi (N), Stephen Batterl (N), Yiyuan Wu (N), Alice Santilli (N) , Sandhya Kannayiram (N), Susan Goodman (3B- Johnson and Johnson, Pfizer. 5- Novartis), Bella Mehta (9- Janssen, Tonix Pharmaceutical)

ABSTRACT INTRODUCTION: Unstructured data in physician notes can be incredibly valuable, especially for understanding patient-reported outcomes (PROs), which are often mentioned in these notes. In this study, we explore an Artificial Intelligence technique to efficiently extract patient-reported pain in Osteoarthritis (OA) patients by prompt engineering Large Language Models (LLMs). Prompt engineering is the process of designing inputs in LLMs to help to get insights from unstructured data. Our objective is to apply prompt engineering techniques to extract/predict patient-reported pain in physician notes in OA patients.

METHODS: We used notes from patients with osteoarthritis at Internal Medicine screening (IM) and Orthopedic visits(ortho) enrolled in a registry before Total Knee Arthroplasty. A trained abstractor extracted pain scores (0-10) where they were explicitly stated. For notes without explicit scores, a board-certified IM physician estimated it based on the content. Notes were processed using the Llama 3.2-90B model, with a prompt engineered to extract 0-10 pain scores. (Figure 1). Prompt engineering techniques, such as Chain of Thought and Few Shot examples, were employed. Outside of the visits, patients reported a 0–10 pain score (Table 1). We calculated Kendall's Tau-b to measure ordinal correlation between the LLM predictions, abstractor scores when present and patient reported pain. Physician predictions were analyzed for notes without scores.

RESULTS SECTION: There were 159 patients in this study with corresponding ortho and IM notes. We identified 133 IM notes and 66 ortho notes that explicitly stated pain scores (Figure 2). In comparisons between pain scores extracted via a trained abstractor and the LLM, we observed a strong correlation in IM notes ($\tau = 0.81$) and an even stronger correlation in ortho notes ($\tau = 0.99$). In IM notes, 18 had a pain score written as 0 because the question was unanswered by patients, causing the EMR system to default to zero. In most cases, the LLM recognized and flagged these abnormalities. We then excluded the default zero and had a perfect correlation ($\tau = 1$).

For notes lacking explicit pain scores, the LLM predicted scores for 26 IM notes and 93 ortho notes. When comparing LLM-predicted pain scores to patient-reported pain scores obtained outside of the visit, the correlations were weak (IM: $\tau = 0.17$, $p = 0.30$) (Ortho: $\tau = 0.09$, $p = 0.30$). When a board-certified physician was asked to predict the scores, the correlation remained weak (IM: $\tau = 0.12$, $p = 0.46$) (Ortho: $\tau = 0.13$, $p = 0.12$).

DISCUSSION: When patient pain is explicitly documented in the chart, large language models (LLMs) can accurately extract this information, offering a potential time-saving tool for abstracting data from clinical notes. In contrast, when pain scores are not explicitly documented, LLM predictions are less accurate but still comparable to the performance of physicians attempting to infer scores from the same unstructured text. LLMs represent an efficient method for extracting information from clinical notes which may have variability in documentation when relevant data exists in unstructured form, opening new pathways for leveraging this information in both clinical care and research.

SIGNIFICANCE/CLINICAL RELEVANCE: Patient-reported outcomes (PROs), such as pain, are critical measures for evaluating disease burden and treatment effectiveness in osteoarthritis but are often buried in unstructured clinical notes, making them difficult to systematically capture. Developing efficient AI-based methods to extract these outcomes can enhance clinical decision-making, support research, and reduce the burden of manual chart abstraction.

IMAGES AND TABLES:

Table 1

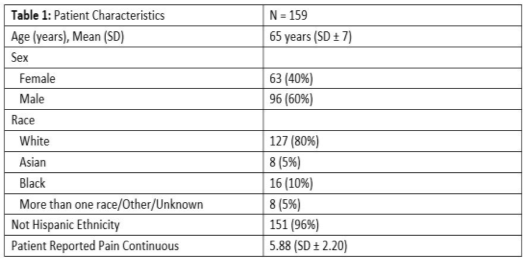
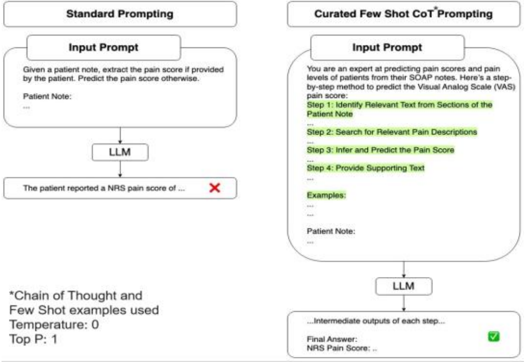


Table 1: Patient Characteristics	N = 159
Age (years), Mean (SD)	65 years (SD ± 7)
Sex	
Female	63 (40%)
Male	96 (60%)
Race	
White	127 (80%)
Asian	8 (5%)
Black	16 (10%)
More than one race/Other/Unknown	8 (5%)
Not Hispanic Ethnicity	151 (96%)
Patient Reported Pain Continuous	5.88 (SD ± 2.20)

Figure 1



Given a patient note, extract the pain score if provided by the patient. Predict the pain score otherwise.

Patient Note:
...

The patient reported a NRS pain score of ... ✗

You are an expert at predicting pain scores and pain levels of patients from their SOAP notes. Here's a step-by-step method to predict the Visual Analog Scale (VAS) pain score:
Step 1: Identify Relevant Text from Sections of the Patient Note
...
Step 2: Search for Relevant Pain Descriptions
...
Step 3: Infer and Predict the Pain Score
...
Step 4: Provide Supporting Text
...

Examples:
...
...

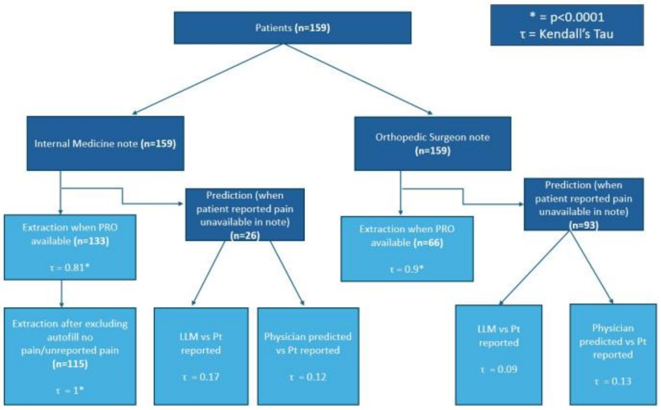
Patient Note:
...

...Intermediate outputs of each step...

Final Answer:
NRS Pain Score: .. ✓

*Chain of Thought and
Few Shot examples used
Temperature: 0
Top P: 1

Figure 2



Patients (n=159)
* = p<0.0001
τ = Kendall's Tau
Internal Medicine note (n=159)
Orthopedic Surgeon note (n=159)
Extraction when PRO available (n=133) τ = 0.81*
Prediction (when patient reported pain unavailable in note) (n=26)
Extraction when PRO available (n=66) τ = 0.9*
Prediction (when patient reported pain unavailable in note) (n=93)
Extraction after excluding autofill no pain/unreported pain (n=115) τ = 1*
LLM vs Pt reported τ = 0.17
Physician predicted vs Pt reported τ = 0.12
LLM vs Pt reported τ = 0.09
Physician predicted vs Pt reported τ = 0.13