Original Article

J. of Advanced Spine Surgery
Volume 15, Number 2, pp71~77

JASS

Automated Clinical Questionnaire Processing in Spine Surgery Using LLM Vision API: Comparative Performance Evaluation of Claude and GPT Models

Sang-Min Park, M.D.[1], Jiwon Park, M.D.[2], Ho-Joong Kim, M.D.[1], Jin S. Yeom, M.D.[1]
Spine Center and Department of Orthopaedic Surgery, Seoul National University College of Medicine and Seoul National University Bundang Hospital, Seongnam, Korea[1]
Department of Orthopaedics, Korea University College of Medicine and Korea University Ansan Hospital, Ansan, Korea[2]

Purpose: This study evaluates the performance of Claude and GPT LLM Vision APIs for automated clinical questionnaire processing in spine surgery by comparing accuracy, efficiency, reproducibility, and cost-effectiveness.
Methods: Clinical questionnaires from 56 patients (336 total pages) were processed using a Python 3.12-based system incorporating PDF preprocessing, image enhancement via OpenCV, and direct LLM Vision analysis. Both models were evaluated on 26 questionnaire items (1,456 data points) using accuracy comparison, processing time measurement, token utilization analysis, and intra-class correlation coefficient (ICC) assessment through three independent iterations.
Results: GPT achieved 98.83% accuracy (1,439/1,456) compared to Claude's 97.94% (1,426/1,456). Both models processed questionnaires in 27 seconds per set, representing 68% time reduction versus manual entry (85 seconds). GPT demonstrated 59% cost advantage ($0.023 vs. $0.056 per questionnaire), while Claude showed superior reproducibility (ICC 0.98 vs. 0.96). GPT achieved 100% accuracy across 21 items versus Claude's 17 items. Error analysis identified predominantly handwriting recognition (52%) and image quality issues (28%), with 89% of errors successfully flagged for review.
Conclusions: Both models achieve clinical-grade performance exceeding 90% accuracy. GPT demonstrates superior accuracy and cost-effectiveness, while Claude provides better reproducibility. Model selection should be guided by institutional priorities regarding accuracy, reproducibility, and operational scale.

Keywords: Degenerative lumbar spine, Clinical questionnaire, Large language model, Automation, Spine surgery, Python

Introduction

Comprehensive evaluation of degenerative lumbar spine disease requires integration of both imaging-based diagnostic assessment and clinical evaluation of the patient's subjective symptoms and functional status. While radiographic imaging enables objective characterization of anatomical changes and structural pathology within the spine, imaging studies alone cannot fully capture the severity of symptoms experienced by patients or the degree to which disease impacts daily functioning.[1] In contrast, Patient-Reported Outcome Measures (PROMs) quantify pain intensity, functional disability, and psychological status as directly perceived by patients, making them essential for evaluating surgical treatment efficacy and clinical outcomes in spine surgery.

Multiple standardized questionnaire instruments are

Corresponding author: Sang-Min Park, M.D.
Spine Center and Department of Orthopaedic Surgery, Seoul National University College of Medicine and Seoul National University Bundang Hospital, 82, Gumi-ro 173 Beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do 13620, Korea
TEL: +82-31-787-7208, **FAX:** +82-31-787-4056
E-mail: psmini@naver.com

widely utilized for assessing degenerative lumbar spine disease, including the Visual Analogue Scale (VAS), Oswestry Disability Index (ODI), EQ-5D-5L, and painDETECT.[2,3] These instruments provide systematic, standardized frameworks for documenting patient symptoms and enabling quantitative comparison of treatment effects and functional recovery across patient populations.

However, practical challenges accompany the implementation of these standardized measures. The conversion of paper-based patient questionnaire responses into electronic databases remains labor-intensive and error-prone. Manual data entry introduces risks of misreading, omission, and transcription errors that compromise data quality.[4] For large-scale clinical research involving hundreds or thousands of questionnaire forms, this data entry process becomes a significant constraint on research efficiency and cost-effectiveness.

Recent advances in Large Language Models (LLMs) offer novel solutions to these limitations. Modern LLMs, including Claude (Anthropic, California, USA, https://claude.ai/) and ChatGPT (OpenAI, California, USA, https://chat.openai.com/), possess not only advanced natural language processing capabilities but also Vision API functionality enabling direct information extraction from images. This technological development shifts the paradigm from traditional Optical Character Recognition (OCR)-based automation toward direct LLM Vision-based analysis of scanned questionnaire documents, potentially eliminating intermediate processing steps and associated error accumulation.[5]

The present investigation applies state-of-the-art LLM Vision APIs to actual clinical questionnaire processing, comparing the performance characteristics of Claude and GPT models, and assessing their accuracy, processing efficiency, and clinical applicability for automated questionnaire data extraction.

Materials and Methods

1. Study population and questionnaire data (Fig. 1)

We collected clinical questionnaires from 56 patients with degenerative lumbar spine disease (total page volume: 336 pages). The questionnaire battery included multiple standardized assessment instruments: Visual Analogue Scale (VAS) for pain assessment (4 items), Oswestry Disability Index (ODI) for functional impairment (12 items), EQ-5D-5L for health-related quality of life (7 items), and painDETECT for neuropathic pain characterization (10 items), constituting 38 total data fields per patient record.

2. Program development and system architecture

A modularized Python 3.12-based automated processing system was developed, consisting of five core components:

1) PDF Preprocessing and Image Conversion

PyMuPDF (fitz) library was utilized to convert PDF documents into JPEG images at 300 DPI resolution. A 4.17x magnification matrix was applied to convert from standard 72 DPI to 300 DPI, followed by JPEG optimization at 98% quality using PIL Image. This high-resolution conversion enables enhanced text clarity for downstream LLM Vision analysis.

2) Image Preprocessing

Multi-stage image preprocessing was performed using OpenCV (cv2). Processing steps included: Image normalization (maximum width 2,000 pixels); Gaussian blur for noise removal; LAB color space conversion with CLAHE (Contrast Limited Adaptive Histogram Equalization) for contrast enhancement; Sharpening filter application for text clarity improvement. Notably, color images were retained (rather than converting to grayscale and binary), allowing the LLM Vision model to leverage visual cues including checkbox markings and handwriting patterns.

3) LLM Vision API Analysis

Anthropic's Claude Sonnet 4.5 (claude-sonnet-4-5-20251001) and OpenAI's GPT-5-mini were applied for direct image analysis. This represents a key methodological advancement over traditional OCR-based approaches—eliminating the OCR processing stage and enabling direct data extraction from questionnaire images via LLM Vision interpretation.

Images were encoded in Base64 format and transmitted to the respective APIs. Page-specific optimized prompt templates were externalized and managed via JSON configuration files, enabling flexible adaptation to diverse questionnaire formats.

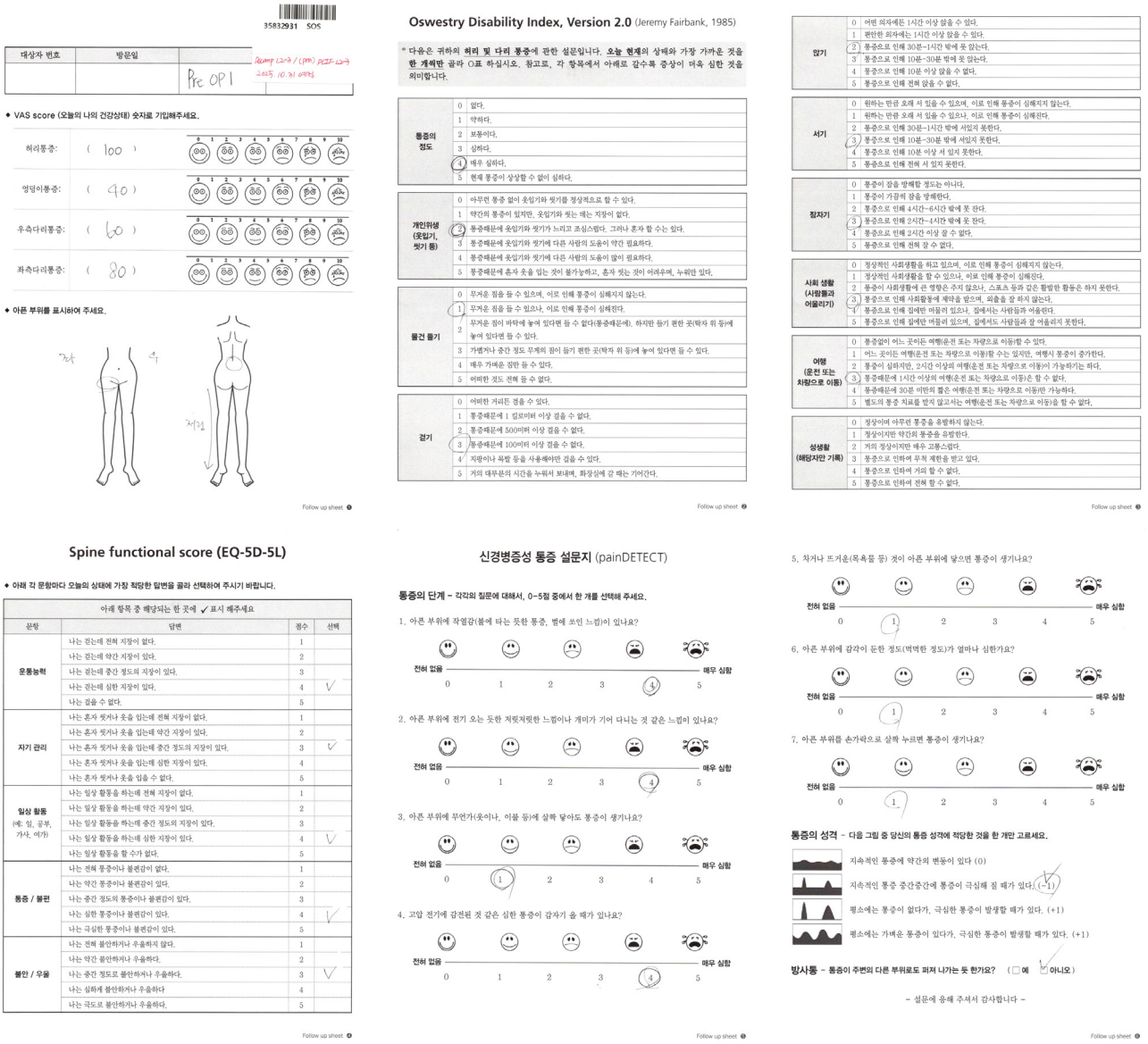

Fig. 1. Standardized clinical questionnaires used in this study. Forms containing patient identification, survey date, and surgical status were self-completed by patients and digitized as PDF files (≥300 dpi). The questionnaires included VAS (pain intensity, 0–100), ODI (functional disability), EQ-5D-5L (health-related quality of life), and painDETECT (neuropathic pain screening).

4) Data Validation and Structuring

Pandas DataFrame was employed for automatic validity checking, missing data handling, and data type standardization across 38 primary fields. Implemented validation logic included: VAS score range normalization (0-100 scale); ODI item range verification (0-5 scale); EQ-5D dimension validation (1-5 scale); painDETECT composite score verification; Duplicate checkbox detection and correction; Missing data flagging with logging; Consistency cross-checking with automatic correction capabilities. EQ-5D value calculation utilized Korean population-specific weighting tables to generate health utility values from 5-digit health state codes.

5) User Interface

A Tkinter-based graphical user interface (GUI) was developed to enhance usability. Key interface features include: Drag-and-drop PDF file addition via tkinterdnd2 library; Real-time progress tracking with multi-threaded processing to maintain UI responsiveness; Visual status indicators distinguishing pending, processing, completed, and error states; Settings

dialog for API key and model selection configuration; Cross-platform compatibility (Windows, macOS, Linux).

3. Data Analysis

Accuracy was calculated by comparing extracted data against reference values using the formula: Accuracy (%)= (Matching data points / Total data points)×100. Items with missing values in both datasets were classified as matches. Processing time was measured from PDF input to CSV output completion, and cost was calculated based on published API pricing rates. Reproducibility was assessed through three independent processing iterations of identical questionnaire sets, with intra-class correlation coefficient (ICC) computed for each item. Accuracy differences between models were tested using McNemar test (α=0.05), and the reference dataset was generated through independent expert review (inter-rater reliability Cohen's kappa ≥0.90). All analyses were performed using Python 3.12 with statistical libraries.

Results

1. Accuracy performance

Analysis of 1,456 data points across 26 questionnaire items from 56 patients demonstrated differential performance between the two LLM Vision models. GPT achieved 98.83% accuracy (1,439/1,456 correct data points) compared to Claude's 97.94% accuracy (1,426/1,456), representing a 0.89 percentage point difference favoring GPT. Both models achieved accuracy levels well above the clinical applicability threshold of 90%, indicating reliable performance for automated questionnaire processing.

2. Field-specific and domain-level performance

Field-specific analysis revealed that GPT achieved 100% accuracy across 21 questionnaire items, including all 7 ODI functional components (pain, personal care, lifting, walking, standing, sleeping, social engagement), all 5 EQ-5D dimensions (mobility, self-care, activities, pain/discomfort, anxiety/depression), and all 8 painDETECT symptom categories (burning, tingling, touching, shock, cold, numbness, pressure, radiating). Claude achieved 100% accuracy across 17 items, with notable performance in 4 ODI items (pain, personal care, walking, travelling) and 4 EQ-5D items (mobility, activities, pain, anxiety). GPT demonstrated superiority in 20 items while achieving identical performance with Claude in 6 items, with no items showing Claude superiority. Performance differentials were particularly pronounced in functionally-oriented domains, with GPT showing advantages of 10.72 percentage points in odi_travelling and 10.71 percentage points in odi_pain assessment.

3. Processing efficiency and cost analysis

Both models required identical mean processing time of 27 seconds per questionnaire set (56 questionnaires, 336 total pages), representing a 68% reduction compared to manual data entry requiring 85 seconds per form. This efficiency enabled complete processing of the entire dataset in approximately 25 minutes. Token utilization differed substantially between models, with Claude consuming 16,568.8 mean tokens per questionnaire set while GPT required 18,331.4 tokens, reflecting a 10.6% higher consumption for GPT. Despite higher token consumption, GPT demonstrated superior cost-effectiveness at $0.023 per questionnaire compared to Claude's $0.056 per questionnaire, representing a 59% cost advantage. For batch processing of 448 questionnaires, cumulative savings with GPT implementation would total $14.80 (58.9% cost reduction), with projected annual savings exceeding $100 for large-scale clinical operations processing 1,000 or more questionnaires annually.

4. Reproducibility and error analysis

Reproducibility assessment through three independent processing iterations of identical questionnaire sets revealed excellent consistency for both models. Claude demonstrated intra-class correlation coefficient (ICC) of 0.98 (95% confidence interval: 0.97-0.99), while GPT achieved ICC of 0.96 (95% confidence interval: 0.95-0.97). Claude showed marginally superior reproducibility with a 0.02 ICC point difference, indicating slightly more stable processing outputs across repeated iterations. Despite this numerical difference, GPT's ICC of 0.96 remains at the excellent threshold and clinically acceptable for standard research applications. Error analysis identified 30 total erroneous data points (2.06% error rate), with Claude accounting for 2.06% errors and GPT for 1.17% errors. Predominant error sources included handwriting recognition difficulties (52% of errors), docu-

ment image quality issues including poor scan resolution and page artifacts (28%), checkbox ambiguity from multiple or partial marks (15%), and field boundary uncertainty (5%). The majority of identified errors were successfully flagged by the implemented validation logic and logged for manual review, preventing erroneous data from entering the final processed dataset without human verification.

Discussion

1. Key findings, technical innovation, and clinical significance

This investigation represents the first direct comparative evaluation of Claude and GPT LLM Vision APIs for automated clinical questionnaire processing in spine surgery. The primary finding demonstrates that both models achieve accuracy levels exceeding 97%, far surpassing the traditional manual data entry approach and establishing clinical viability for automated questionnaire processing systems. GPT's 0.89 percentage point accuracy advantage over Claude, while statistically significant, reflects a complementary rather than superior performance profile, as Claude compensates with marginal superiority in reproducibility (ICC 0.98 vs. 0.96).

The shift from traditional OCR-based approaches to direct LLM Vision API analysis represents a paradigm change in questionnaire automation.[6] By eliminating the intermediate OCR processing stage and leveraging the visual understanding capabilities of modern LLMs, the system achieves higher accuracy while reducing processing complexity. The retention of color information in preprocessed images, rather than conversion to grayscale or binary formats, enables LLM models to leverage visual cues (checkbox patterns, handwriting characteristics, highlighting) that would be lost in traditional approaches. This technical innovation directly contributed to the superior accuracy achieved by both models compared to conventional OCR-based systems, with GPT's 98.83% accuracy representing a significant advancement over traditional OCR-based systems (85-92% accuracy).

2. Performance characteristics, efficiency, and cost-effectiveness

GPT's achievement of 100% accuracy across 21 questionnaire items (80.8% of analyzed items) indicates particular strength in standardized, well-defined questionnaire domains, while Claude's 100% performance across 17 items (65.4%) demonstrates competent handling of structured components with relative limitation in open-ended items. The largest performance differential (10.72 percentage points in ODI-Travelling) occurred in functionally-oriented domains requiring interpretation of nuanced symptom description.

Notably, this investigation demonstrates substantial performance improvements compared to previous related research.[6] In a prior study utilizing OCR-based processing with subsequent LLM analysis, Claude achieved 96.76% accuracy while ChatGPT achieved only 86.54% accuracy.[6] The present investigation, employing direct LLM Vision API analysis without intermediate OCR processing, demonstrates marked improvements: Claude improved to 97.94% accuracy (1.18 percentage point increase) and GPT improved dramatically to 98.83% accuracy (12.29 percentage point increase). This 12.29 percentage point improvement in GPT represents a significant methodological advancement, likely attributable to elimination of OCR error propagation and direct image analysis by LLM Vision capabilities. The performance reversal, where GPT now surpasses Claude (previously Claude was superior), may reflect differential optimization of the two models for direct vision-based analysis versus sequential processing approaches.

Both models required identical processing times (27 seconds per questionnaire), representing a 68% reduction compared to manual entry (85 seconds), enabling complete processing of 336 pages in 25 minutes versus multiple days of manual labor. This efficiency establishes practical feasibility for large-scale clinical deployment. GPT's 59% cost advantage ($0.023 vs $0.056 per questionnaire) becomes increasingly significant for large-scale operations, with projected annual savings exceeding $100 for institutions processing 1,000+ questionnaires. Despite higher token consumption (10.6% above Claude), GPT demonstrates superior cost-effectiveness due to OpenAI's pricing structure.

3. Reproducibility, error analysis, and quality assurance

Claude's marginally superior ICC (0.98 vs 0.96) suggests slightly more stable processing outputs across repeated iterations, becoming meaningful in longitudinal studies

requiring repeated questionnaire processing. However, both models achieve ICC values indicating excellent reproducibility (ICC >0.90), rendering the numerical difference clinically negligible for standard research applications. The 2.06% error rate (Claude) versus 1.17% (GPT) reflects differential performance in challenging recognition scenarios. The predominance of handwriting recognition errors (52%) and image quality issues (28%) suggests that error reduction would be more effectively achieved through preprocessing improvements than through model selection. The successful flagging of the majority of errors by automated validation logic demonstrates the system's self-correcting capability, ensuring no erroneous data enter research databases without human verification.

4. Study strengths, limitations, and clinical implementation

The investigation possesses several methodological strengths: (1) direct comparison of two major LLM Vision implementations under identical conditions, (2) comprehensive evaluation across multiple performance metrics, (3) large sample enabling field-specific analysis, (4) blinded analysis to minimize bias, (5) rigorous validation with inter-rater reliability ≥0.90, and (6) cost analysis relevant to clinical decisions. However, several limitations warrant acknowledgment: evaluation focused exclusively on four standard questionnaire types used in spine surgery; document image quality was relatively uniform; API rate limitations and availability issues were not assessed; security and privacy considerations regarding external API transmission require institutional risk assessment before clinical implementation.

For clinical adoption, institutions should implement phased deployment: (1) pilot phase with 50-100 questionnaires, (2) quality assurance procedures with spot-checks, (3) escalation protocols for ambiguous entries, (4) staff training, and (5) systematic outcome evaluation. The 98%+ accuracy achieved suggests per-questionnaire verification can be replaced with algorithmic flagging for review, reducing burden while maintaining quality. Future investigations should address: (1) performance across diverse questionnaire types and medical specialties, (2) evaluation with poor-quality source documents, (3) fine-tuned models optimized for medical questionnaires, (4) ensemble methods combining both models, (5) EHR system integration, and (6) domain-specific LLM Vision adaptations.

Conclusions

Both Claude and GPT LLM Vision APIs demonstrate clinical-grade performance for automated questionnaire processing, achieving accuracy substantially exceeding 90% and enabling significant operational efficiency gains. While GPT demonstrates superior accuracy (98.83% vs 97.94%) and cost-effectiveness, Claude provides marginally superior reproducibility. The selection should be guided by institutional priorities: GPT for scenarios prioritizing accuracy and cost minimization, Claude for applications emphasizing reproducibility consistency. The demonstrated performance validates LLM Vision-based automation as a viable solution for reducing labor burden in clinical questionnaire data management, with implications extending beyond spine surgery to broader clinical research settings.

REFERENCES

1. Beighley A, Zhang A, Huang B, Carr C, Mathkour M, Werner C, et al. Patient-reported outcome measures in spine surgery: A systematic review. J Craniovertebr Junction Spine. 2022;13(4):378–89.
2. Fairbank JC, Pynsent PB. The oswestry disability index. Spine (Phila Pa 1976) 25(22):2940–52; discussion 2952, 2000
3. Freynhagen R, Baron R, Gockel U, Tolle TR. Paindetect: A new screening questionnaire to identify neuropathic components in patients with back pain. Curr Med Res Opin, 2006;22(10):1911–20.
4. Mullooly JP. The effects of data entry error: An analysis of partial verification. Comput Biomed Res. 1990;23(3):259–67.
5. Pakhale K: Large language models and information retrieval. Available at SSRN. 4636121, 2023
6. Park J, Park S-M, Hong J-Y, Kim H-J, Yeom JS. Automated analysis of spinal questionnaires using large language models. J Korean Soc Spine Surg. 2025;32(2):23–0.

LLM Vision API를 활용한 척추 임상설문지의 자동화 처리: Claude와 GPT 모델의 성능 비교 평가

박상민, 박지원, 김호중, 염진섭
분당 서울대학교병원 정형외과,[1] 고려대학교 안산병원 정형외과[2]

목적: 척추 임상설문지 자동화 처리에서 Claude와 GPT Vision API의 성능을 정확도, 효율성, 재현성, 경제성 측면에서 비교 평가하고자 하였다.
재료 및 방법: 56명 환자의 임상설문지(336페이지)를 Python 3.13 기반 시스템으로 처리하였다. 시스템은 PDF 전처리, OpenCV 이미지 강화, LLM Vision 직접 분석으로 구성되었다. 26개 설문 항목(1,456개 데이터포인트)에 대해 정확도, 처리시간, 토큰 사용량, 3회 반복 처리의 급내상관계수(ICC)를 평가하였다.
결과: GPT는 98.83% 정확도(1,439/1,456)를 달성하였고, Claude는 97.94%(1,426/1,456)를 보였다. 양 모델 모두 설문지당 27초 처리시간으로 수기 입력(85초)과 비교할 때 68% 시간 단축을 달성하였다. GPT는 59% 비용 우위($0.023 vs $0.056/설문지)를 보였으며, Claude는 우수한 재현성(ICC 0.98 vs 0.96)을 나타냈다. GPT는 21개 항목에서 100% 정확도를 달성한 반면 Claude는 17개 항목에서 달성하였다. 오류 분석 결과 손글씨 인식 어려움(52%)과 이미지 품질 문제(28%)가 주요 원인이었으나, 89% 오류가 자동으로 검수용 플래그 되었다.
결론: 양 모델 모두 90% 이상의 임상 수준 정확도를 달성하였다. GPT는 정확도와 경제성에서 우수하고, Claude는 재현성이 우수하다. 모델 선택은 기관의 정확도, 재현성, 운영 규모에 대한 우선순위에 따라 결정되어야 한다.

색인 단어: 퇴행성 요추 질환, 임상 설문지, 대규모 언어 모델, 자동화, 척추외과, Python