Evaluating ChatGPT-4o, Gemini and Claude 3.7 in Endodontic Diagnostics:

A Prospective Clinical Study

NCT number: not applicable

Ethical Approval: Date: 26.06.2025, Protocol No: 2025-07-02/38

Record Release Date/Time: November 24, 2025 14:03

Last Updated: December 8, 2025

Study protocol

This observational study evaluated the diagnostic and treatment-planning performance of three large language models (LLMs)—ChatGPT-4o, Claude 3.7 Sonnet, and Gemini Advanced—using real clinical cases collected from 120 patients who presented to an endodontic clinic. For each patient, a standardized case file was prepared that included demographic information, detailed medical and dental history, clinical findings (pain characteristics, percussion/palpation sensitivity, vitality testing, periodontal status, sinus tract presence, discoloration), and a periapical radiograph. All information was anonymized and presented in an identical, fixed order to all models.

Each LLM received the same single prompt ("Determine the diagnosis and propose the appropriate treatment plan") and the same case structure. Radiographs were uploaded as original-resolution JPEG files through each platform's native multimodal interface. Text was entered first, followed by the image upload. ChatGPT-4o, Gemini Advanced, and Claude 3.7 were all used in their default multimodal configurations with no web-browsing, plug-ins, external retrieval tools, OCR, or fine-tuning enabled. "Search with Gemini" was manually disabled.

Each case was submitted to the LLMs using newly created accounts, accessed via Google Chrome on a desktop computer. All questions were delivered within a 24-hour window and in the same predetermined order. No pre-testing, prompt engineering, or clarifying follow-up messages were allowed. If a model requested additional information, no further input was given. Each model generated only one response per case, reflecting real-world first-response performance.

To avoid contextual contamination, each case was entered in a separate chat session with cleared conversation history. All outputs were recorded verbatim and saved in Word format. A panel of three independent endodontic specialists established the gold-standard diagnoses and treatment plans for all cases. Disagreements were resolved by consensus. AI responses were coded as “1” (consistent with gold standard) or “0” (inconsistent).