Comparative Accuracy of Large Language Models for CPT Coding Assignments from Surgical Procedure Notes

Abdalrahman Katranji

Abdalrahman@simplifyaio.com

Northwestern University

Aisa De Vries

Simplify AI

Abdalmajid Katranji

Katranji Hand Center

Mohammad Zalzaleh

Coastline Orthopaedic Associates

Article

Keywords: Medical Coding Automation, Orthopedic Surgery, Artificial Intelligence, Healthcare Informatics, Code Hallucination, HCPCS, CPT coding, Large Language Models

Posted Date: January 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8475390/v1

Additional Declarations: Competing interest reported. Abdalrahman Katranji, Aisa De Vries, and Dr. Abdalmajid Katranji are affiliated with Simplify AI, an AI medical scribe platform. Simplify AI utilizes API services from Anthropic, OpenAI, and Google in its commercial products. Dr. Zalzaleh is a client of Simplify AI. No company reviewed or influenced the study design, analysis, or manuscript preparation. The authors received no funding or compensation from any of the three companies evaluated in this study. Benchmark labeling was performed independently by each surgeon without input from Simplify AI staff or AI tools.

Abstract

Background: Medical procedure coding is time-intensive and error-prone, with direct implications for reimbursement accuracy and operational efficiency. Large Language Models (LLMs) show promise for automating CPT code assignment, yet their accuracy on surgical procedure notes compared to physician-defined benchmarks remains understudied.

Objective: To evaluate and compare the CPT-code assignment performance of some of the most popular LLMs capable of reasoning (Anthropic Claude Opus 4.5, OpenAI GPT-5.2, and Google Gemini 3 Pro) against a surgeon-labeled benchmark for orthopedic procedure notes.

Methods: Thirty-three publicly available, de-identified orthopedic procedure notes were obtained from MTSamples and Medical Transcription Sample Reports. Two surgeons, blinded to AI outputs, independently assigned benchmark CPT codes to notes within their specialty scope (28/33 notes labeled). Three frontier-class LLMs (Claude Opus 4.5, GPT-5.2, and Gemini 3 Pro) were selected based on LMArena performance and configured with extended reasoning at maximum settings. Each model was queried three times per note using identical prompts (n=297 total queries). A code was considered "predicted" if it appeared in at least 2 of 3 runs. Predicted codes were validated against the 2025 CMS HCPCS/CPT database. Performance metrics included precision, recall, F1 score, hallucination rate, invalid code rate, and consistency rate.

Results: Of 33 orthopedic procedure notes evaluated (28 with valid benchmark labels), Claude Opus 4.5 achieved the highest accuracy (F1: 65.9%, precision: 66.7%, recall: 65.2%), followed by Gemini 3 Pro (F1: 62.1%) and GPT-5.2 (F1: 56.8%). Consistency did not correlate with accuracy: Gemini demonstrated the highest run-to-run consistency (72.7% identical outputs across runs) despite lower benchmark alignment, while Claude showed greater variance (63.6%) yet superior accuracy. No model produced hallucinated or invalidly formatted codes (0% hallucination rate, 0% invalid rate). Performance varied substantially by procedural complexity: simple single-code procedures achieved near-perfect consistency across models, while complex multi-component procedures were more likely to show F1 scores below 40% and greater inter-run variance.

Conclusion: Current frontier LLMs demonstrate moderate accuracy in CPT code assignment for orthopedic procedures but are not yet suitable for autonomous clinical use. These models may offer value as first-pass tools within human-in-the-loop workflows, particularly for straightforward procedures. Future research should evaluate prompting optimization, modifier assignment, and prospective human-AI collaborative coding in real billing environments.

I. INTRODUCTION

Current Procedural Terminology (HCPCS Level I) codes function as the standardized nomenclature used to describe medical procedures and services performed by providers in the United States healthcare system. CPT code assignment accuracy is pivotal for reimbursement, resource allocation, and regulatory

compliance (Dotson, 2013).[1] Each year, healthcare insurers in the United States process over 5 billion claims for payment (Centers for Medicare & Medicaid Services [CMS], 2025).[2] However, the coding process remains complex and labor-intensive, contributing to error rates in professional coding encounters and operational costs from claim denials, resubmissions, and compliance failures (Hou et al., 2025).[3]

Traditional automation approaches have achieved limited success. Computer Assisted Coding (CAC) systems employ rule-based natural language processing to suggest codes from clinical text but are constrained by rigid pattern-matching, extensive institutional customization requirements, and high false-positive rates requiring manual review (Campbell and Giadresco, 2020).[4] Supervised machine learning approaches show promise for narrow coding tasks but demand large labeled training datasets, lack interpretability, and struggle to generalize across documentation styles and procedure types (Dong et al., 2022).[5]

The emergence of frontier Large Language Models (LLMs) including OpenAI's GPT-5.2, Anthropic's Claude Opus 4.5, and Google's Gemini 3 Pro offers a potential alternative to traditional coding automation. Unlike rule-based CAC systems, these models can interpret unstructured clinical documentation contextually and operate with minimal task-specific training through zero-shot prompting (Lee et al., 2025).[6] Recent model generations incorporate extended reasoning modes that enable deliberate, step-by-step analysis before generating outputs, a capability particularly relevant for complex multi-component procedures (Maity and Saikia, 2025).[7] Additionally, these models can produce structured outputs (e.g., JSON-formatted code lists) and provide reasoning for their selections, enabling auditability that traditional machine learning approaches lack (Garcia-Carmona et al., 2025).[8]

Despite growing commercial interest in LLM-based coding automation, rigorous peer-reviewed evaluations remain scarce, particularly for surgical procedure documentation where coding complexity is highest. Critical questions remain unanswered: How accurately can LLMs assign CPT codes compared to physician-defined benchmarks? How consistent are outputs across repeated queries? Do models produce hallucinated codes that do not exist in the CPT database? How do different frontier models compare under controlled conditions?

This pilot study addresses these gaps through a systematic comparative evaluation of three frontier LLMs for orthopedic CPT code assignment. We selected orthopedic procedures due to their procedural diversity, well-defined documentation patterns, and clinical significance. Our methodology emphasizes four key design features: (1) surgeon-defined benchmark standards reflecting real-world subjective coding judgment, (2) identical prompting across models to isolate performance differences, (3) repeated queries to quantify output variance, and (4) comprehensive error categorization including hallucination detection, precision, recall, and consistency scoring. These findings provide preliminary evidence regarding current LLM capabilities for surgical coding and establish a hopeful methodological framework for future validation studies.

II. METHODS

We conducted a prospective comparative analysis evaluating the performance of three Large Language Models (LLMs) in assigning CPT codes to orthopedic surgical procedure notes. This pilot study aimed to assess (1) the coding accuracy of LLMs compared to surgeon-labeled benchmarks, (2) the consistency of outputs across repeated queries to LLMs, and (3) the rate of code hallucinations of LLMs.

2.1 Study Design

Model Selection. Three frontier-class LLMs were selected based on their consistent top-tier performance on the LMArena leaderboard as of December 2025: Anthropic Claude Opus 4.5, OpenAI GPT-5.2, and Google Gemini 3 Pro. Selection criteria included: (1) availability of extended reasoning/thinking capabilities, (2) commercial API access, and (3) support for structured JSON outputs.

Experimental Protocol. Each model was queried three times per procedure note (n=33 notes × 3 runs × 3 models = 297 total queries) to assess run-to-run variance. All queries used identical prompting: a static system prompt containing labeling instructions (Appendix A) and a user prompt containing the procedure note (example provided in Appendix B). Queries were executed programmatically via each provider's API on December 13, 2025.

Blinding. Surgeon labeling was performed independently to AI testing. Neither surgeon viewed AI outputs before completing their benchmark labels. Both surgeons were given the same modified AI prompt instruction set that omitted mandatory JSON outputs and asked only to label notes based on comfortability to the underlying procedure. This resulted in only 28/33 tested notes being given a benchmark to be tested against.

Primary Outcome. The primary outcome measure was F1 score (harmonic mean of precision and recall) comparing AI-predicted codes to surgeon-labeled benchmarks. Secondary outcomes included consistency rate, hallucination rate, and invalid code rate.

2.2 Data Source & Testing Standard

Data Source. Procedure notes were obtained from both MTSamples (mtsamples.com) and MedicalTranscriptionSampleReports (medicaltranscriptionsamplereports.com), both were publicly available repositories of anonymized medical transcription samples.[9-15] Both contain no protected health information (PHI) and both commonly are used for NLP research. [18-19]

Inclusion Criteria. Notes were included if they: (1) described an orthopedic surgical procedure, (2) were written in English, (3) contained sufficient operative detail to reasonably assign CPT codes, and (4) were fully de-identified. Notes were excluded if they were ambiguous, incomplete, or described non-surgical encounters (e.g., consultations, imaging reports).

Sample. All orthopedic procedure notes meeting inclusion criteria were reviewed, yielding a final sample of 33 notes. Notes varied in complexity, procedure type, and documentation style (see Appendix F for full note list).

Benchmark Labeling. Two surgeons independently assigned CPT codes to each note following the same labeling instructions provided to the AI models (Appendix A). To reduce labeling burden, notes were divided between surgeons: Dr. Katranji (general and reconstructive hand surgeon) labeled notes in Appendix C; Dr. Zalzaleh (orthopedic surgeon) labeled notes in Appendix D. Surgeons were blinded to AI outputs during labeling. Division of notes between surgeons precluded inter-rater reliability calculation, which is acknowledged as a limitation (see Section 4.2).

2.3 LLM configuration

Each model was configured with extended reasoning capabilities enabled at the highest available setting to maximize coding accuracy. Configuration details are summarized in **Table 1.**

TABLE 1 Model versions and configuration parameters. All models accessed December 13, 2025.

Provider	Model Version	Reasoning Setting	Output Format	Max Tokens
Anthropic	claude-opus-4-5-20251101	Thinking enabled, budget_tokens: 63,999	JSON	64,000
OpenAI	gpt-5.2-2025-12-11	reasoning_effort: xhigh	JSON object	Default
Google	gemini-3-pro-preview	Thinking level: HIGH	JSON	Default

Temperature. Temperature parameters were not specified. Anthropic and Google do not accept temperature settings when thinking/reasoning modes are enabled; OpenAI's temperature parameter was omitted to maintain parity across models.

Prompting Strategy. All models received identical prompts: a system prompt containing structured labeling instructions (Appendix A) and a user prompt containing the raw procedure note text (sample in Appendix B). Prompts requested JSON-formatted output with CPT codes and supporting reasoning.

API Access. All queries were executed programmatically via official provider APIs. Testing code is provided in Appendix E.

2.4 Evaluation Metrics

Model performance was evaluated using the metrics defined in **Table 2.**

Code Validation. Predicted codes were validated against the complete CMS HCPCS/CPT code database (2025 release). Codes with valid format (5-digit CPT or alphanumeric HCPCS Level II) but not present in

the database were classified as hallucinated. Codes with invalid format (e.g., incorrect length, invalid characters) were classified as invalid.

Majority Voting. For benchmark comparison, a code was considered "predicted" by a model if it appeared in at least 2 of 3 runs for a given note. This reduces noise from single-run variance.

TABLE 2 Evaluation metrics and definitions.

Metric	Definition	Rationale
Precision	Correct codes / Total AI-predicted codes	Measures over-coding tendency
Recall (Accuracy)	Correct codes / Total benchmark codes	Measures code omission rate
F1 Score	Harmonic mean of precision and recall	Balances precision and recall
Consistency Rate	Notes with identical outputs across 3 runs / Total notes	Measures determinism; higher = more consistent
Hallucination Rate	Hallucinated codes (don't exist) / Total AI-predicted codes	Measures fabrication of invalid codes
Invalid Rate	Invalid format codes / Total AI-predicted codes	Measures malformed code output

2.5 Statistical Analysis

Evaluation metrics were calculated per-note and aggregated per-model. Given the pilot sample size (n=33), results are presented descriptively without inferential statistics. Analysis was performed using custom Node.js scripts; and visualizations generated via an HTML dashboard. All analysis code is available in Appendix E.

III. RESULTS

3.1 Sample Overview

Of the 33 orthopedic procedure notes evaluated, 28 had valid benchmark labels and were included in accuracy analysis. Five notes were excluded from accuracy analysis: one (Superior Labrum Lesion Repair) was deemed ambiguous by the reviewing surgeon, and four (primarily spine-related procedures) fell outside the labeling surgeons' specialty scope. Interestingly, for the ambiguous note, all three models produced identical codes across all runs, suggesting AI confidence does not always reflect clinical certainty. Each model completed all 99 runs (33 notes × 3 runs) with minimal API failures.

3.2 Aggregate Performance

Model performance across all evaluation metrics is summarized in **Table 3**, with detailed counts and timing data provided in **Table 4.**

TABLE 3 Model Performance Summary

Metric	Anthropic	OpenAI	Gemini	Best
Precision	66.7%	59.5%	65.9%	Anthropic
Recall (Accuracy)	65.2%	54.3%	58.7%	Anthropic
F1 Score	65.9%	56.8%	62.1%	Anthropic
Consistency Rate	63.6%	60.6%	72.7%	Gemini
Invalid Rate	0%	0%	0%	All
Hallucination Rate	0%	0%	0%	All

TABLE 4 Detailed Performance Breakdown

Metric	Anthropic	OpenAI	Gemini	Best
True Positives	30	25	27	Anthropic
False Positives	15	17	14	Gemini
False Negatives	16	21	19	Anthropic
Mean Jaccard	84.6%	80.5%	86.2%	Gemini
Identical Notes	21 / 33	20 / 33	24 / 33	Gemini
Average Response Time	45.65s	83.45s	34.77s	Gemini
Average Input Tokens	1497.58	1559.45	1368.73	Gemini
Average Output Tokens	2776.01	4655.98	233.12	Gemini

Anthropic's Claude Opus 4.5 achieved the overall highest accuracy, with an F1 score of 65.9% followed by Gemini with a score of 62.1% and lastly OpenAI with a score of 56.8%. This was driven primarily by Anthropic's superior recall (65.2% vs. 58.7% and 54.4%), indicating fewer missed codes. Precision was also a big determining factor in its overall success with Anthropic again leading with a rate of 66.7% as compared to Gemini's with a rate of 65.9%, and OpenAI with a rate of 59.5%.

Despite lower accuracy, Gemini demonstrated the highest overall run-to-run consistency (Consistency Score: 72.7%) and mean Jaccard similarity (86.2%), suggesting it having the most deterministic behavior of the three models tested. OpenAI showed the greatest variability, with only 60.6% of notes producing identical outputs across runs (Jaccard of 80.5%). And Anthropic stood in the middle with a consistency rate of 63.6% and a mean Jaccard of 84.6%.

No model produced codes with invalid formatting nor did any model have hallucinated codes outputted, indicating that all predicted codes existed in the HCPCS/CPT database even when incorrect codes or descriptions were given for the underlying procedure.

Gemini was the fastest model (34.77s average) while producing the most concise outputs (233 tokens). OpenAI was slowest (83.45s) with the most verbose reasoning (4,656 tokens). Anthropic was intermediate (45.65s, 2,776 tokens).

3.3 Variance Analysis

To assess run-to-run consistency, each model was queried three times per note. **Table 5** summarizes variance metrics across models.

Gemini produced the most deterministic outputs, with nearly three-quarters of notes (72.7%) yielding identical codes across all runs. Anthropic and OpenAI showed greater variability, with approximately one-third of notes producing different code sets between runs.

Notes with high procedural complexity or multiple billable components were more likely to exhibit variance. For example, "Anterior Cervical Discectomy & Fusion" produced different code combinations across runs for all three models, likely reflecting ambiguity in whether instrumentation and graft codes should be reported separately. Conversely, straightforward single-code procedures such as “Achilles Tendon Repair” and “Biceps Tendon Repair” showed perfect consistency across all models and runs.

Notably, higher consistency did not correlate with higher accuracy. Gemini achieved the highest Consistency Rate (72.7%) but a lower F1 (62.1%) than Anthropic (Consistency Rate: 63.6%, F1: 65.9%), suggesting that deterministic output does not always guarantee correctness.

TABLE 5 Variance metrics by model. Consistency Rate represents the proportion of notes with identical code outputs across all three runs. Jaccard similarity measures mean pairwise overlap between runs (1.0 = perfect agreement).

Metric	Anthropic	OpenAI	Gemini	Best
True Positives	30	25	27	Anthropic
False Positives	15	17	14	Gemini
False Negatives	16	21	19	Anthropic
Mean Jaccard	84.6%	80.5%	86.2%	Gemini
Identical Notes	21 / 33	20 / 33	24 / 33	Gemini
Consistency Rate	63.6%	60.6%	72.7%	Gemini

3.4 Error Analysis

False positives primarily consisted of valid codes that were related to but not appropriate for the documented procedure. Models tended to over-prescribe CPT codes, predicting separate codes for components that should have been bundled into the primary procedure code per CMS guidelines, or assigning codes for procedures not clearly documented in the operative note.

False negatives reflected codes omitted by the models, particularly in complex multi-procedure cases. As procedural complexity increased, so did the error rate and run-to-run variance. Notes with single, straightforward procedures (e.g., Achilles Tendon Repair) achieved more perfect consistency across all models, while notes with multiple billable components (e.g., Bunionectomy & Arthrodesis with 7 benchmark codes) showed substantial disagreement both between models and across runs within the same model.

A good example of this pattern emerged as all models struggled with determining correct code quantity assignment on a per-note basis. In the Bunionectomy case, the benchmark specified CPT 28285 three times to reflect procedures on three distinct surgical sites. Anthropic over-coded (4×), while Gemini and OpenAI under-coded (1× each).

3.5 Illustrative Cases

Two cases are presented below to illustrate patterns observed across the dataset. Automated per-note analysis is provided in Appendix G through the supplementary interactive dashboard.

Case 1: Unanimous Agreement (Achilles Tendon Repair):

All three models correctly and consistently identified CPT 27650 (repair of Achilles tendon) across all runs, matching the benchmark standard. This straightforward single-code case demonstrated that all models are somewhat capable of understanding accurate coding when procedural documentation is completely unambiguous (results as shown in **table 6**).

TABLE 6 Results of Unanimous Agreement (Achilles Tendon Repair)

Model	Run 1	Run 2	Run 3	Benchmark
Anthropic	27650	27650	27650	27650
OpenAI	27650	27650	27650	27650
Gemini	27650	27650	27650	27650

Case 2: Model Divergence (Bunionectomy & Arthrodesis):

This complex multi-procedure case revealed inter-model disagreement.The benchmark standard [28293, 28285, 28285, 28285, 28286, 28234, 28308] coded the CPT 28285 three times to reflect the procedure that occurred on the three distinct surgical sites. Yet every model coded in variance to this with Anthropic overcoding 28285 for four surgical sites as compared to the benchmark's analysis of three,

and both Gemini and OpenAI only giving 28285 once, representing the procedure was done once on a single surgical site. Further the benchmark codes of: 28293, 28286, 28234, 28308 were all varied in output from run to run across models (results as shown in **table 7**).

TABLE 7 Results of Model Divergence (Bunionectomy & Arthrodesis)

Model	Predicted Codes	P	R	F1
Anthropic	Run 1: [28293, 28285, 28285, 28285, 28285, 28234, 28270, 28110] Run 2: [28293, 28285, 28285, 28285, 28285, 28270, 28110] Run 3: [28293, 28285, 28285, 28285, 28285, 28110, 28270]	50%	29%	36%
OpenAI	Run 1: [28292, 28285, 28270, 28110] Run 2: [28293, 28285, 28270, 28110] Run 3: [28291, 28285, 28270, 28110]	33%	14%	20%
Gemini	Run 1: [28291, 28285, 28110] Run 2: [28291, 28285, 28110, 28270] Run 3: [28293 28285, 28110, 28270]	25%	14%	18%
Benchmark	28293, 28285, 28285, 28285, 28286, 28234, 28308	--	--	--

This case illustrates that complex procedures with multiple billable components still remain somewhat challenging for current LLMs, with models both missing codes (false negatives) and suggesting inappropriate codes (false positives).

IV. DISCUSSION

4.1 Main Findings

This pilot study compared three frontier-class LLMs in their ability to assign CPT codes to orthopedic surgical procedure notes. Anthropic's Claude Opus 4.5 achieved the highest overall accuracy (F1: 65.9%), outperforming both Gemini 3 Pro (62.1%) and OpenAI GPT-5.2 (56.8%). However, accuracy did not correlate with consistency: Gemini produced the most deterministic outputs (72.7% of notes identical

across runs) despite lower accuracy, while Anthropic showed greater run-to-run variance yet achieved better benchmark alignment. OpenAI lagged behind both competitors across all primary metrics with a precision score of 59.5%, a recall score of 54.3%, while offering no advantage in consistency. Notably, no model produced hallucinated codes (codes that do not exist in the HCPCS/CPT database), suggesting that current LLMs have internalized valid code structures even when applying them incorrectly. All errors represented valid codes misapplied to the clinical context rather than fabricated codes, although a key limitation of this claim is that descriptions of CPT codes were not tested nor factored into our hallucination rate, had they been, this may have altered our understanding of these results. One unexpected finding was Gemini's repeated failure to produce valid JSON output for a single procedure (Arthroplasty), requiring approximately 15 re-runs to obtain three valid responses, the only API reliability issue observed across 297 total queries.

4.2 Interpretation

The performance gap between these models likely reflects differences in reasoning architecture and capacity rather than medical-specific training. Anthropic's superior accuracy may have stemmed from more exploratory reasoning during its extended thinking process, allowing it to consider and reject alternative code assignments before settling on a final answer. This exploration appears to come at the cost of consistency (the same note processed multiple times may traverse different reasoning paths, arriving at different conclusions). Conversely, Gemini's high consistency but lower accuracy suggests that its architecture may be more deterministic in internal processing. Once Gemini identifies a plausible code, it appears to commit to that interpretation with less second-guessing, producing identical outputs across runs, even when those outputs are incorrect. This "confidently wrong" pattern may reflect lower effective temperature or less stochastic sampling in Gemini's thinking architecture. OpenAI's underperformance was unexpected given GPT-5.2's strong benchmark results in other domains. Despite producing the most verbose reasoning (nearly 5,000 tokens per response), this additional output did not translate to accuracy gains. The model appeared to "explore wrong" more frequently, generating extensive justifications for incorrect code assignments rather than reconsidering its initial interpretation.

A consistent pattern emerged across all models: simple, unambiguous procedures yielded high accuracy regardless of model, while complex multi-component procedures produced substantial errors and variance (such as the tested Open Plantar Fasciotomy note). For single-code procedures like Achilles Tendon Repair, all models achieved unanimous correct coding across all runs. However, even straightforward procedures revealed a troubling pattern: when models erred, they erred consistently. In the Condylectomy example (benchmark: 28288), all three models consistently predicted 28110 across all runs; incorrect, yet deterministic. This 'confidently wrong' pattern illustrates that consistency does not guarantee correctness. This suggested to our team that the models may lack the contextual judgment that experienced physicians bring to code selection. Practicing surgeons draw on knowledge beyond the operative note itself (institutional conventions, payer requirements, anatomical implications, etc.) that current LLMs cannot access without fine-tuning or specialized prompting. When benchmark labelers reviewed our sample notes, their clinical experience informed much of their code selection in ways not

explicitly documented in the procedure text. One example of this was code quantity determination. In the Bunionectomy case requiring CPT code 28285 to be reported three times for three surgical sites, Anthropic over-reported (4× on each run) while Gemini and OpenAI under-reported (1× each). This pattern (difficulty recognizing when a procedure is performed on multiple sites and should be billed multiple times) represents a systematic weakness that, although not tested in this study, prompting alone might be able to resolve.

4.3 Comparison to prior studies

Our findings build upon a small but growing body of literature evaluating LLMs for medical coding tasks. Direct comparison across studies is complicated by differences in input complexity, model versions, and evaluation methodology, but several patterns emerge.

Prior studies using simpler inputs have reported higher accuracy rates. The LWW orthopedic study using ChatGPT-4o achieved 86.7% accuracy (improving to 93.9% after re-prompting), but notably used procedure names alone (e.g., "ACL reconstruction") rather than full operative notes.[20] This represents a fundamentally easier task (extracting codes from standardized procedure terminology versus interpreting narrative surgical documentation). Our use of complete operative notes, with their inherent ambiguity and varying documentation styles, likely accounts for our lower overall accuracy despite using more advanced the models.

Studies using more complex inputs align more closely with our findings. A study evaluated GPT-4, Gemini, and Copilot on plastic surgery operative note templates and found that only 7.7-19.2% of responses were fully correct (Carrarini et al, 2025).[16] Another study tested multiple LLMs on hand surgery procedure descriptions and reported 0% accuracy on complex procedures for ChatGPT models, with simple procedures achieving only 40-75% accuracy. (Isch et al, 2025).[17] Our results with frontier thinking models (56.8-65.9% F1) suggest meaningful improvement over these earlier findings, though the difference in evaluation metrics and procedure types limits direct comparison. Unlike prior studies that relied on manual code validation, we validated all predicted codes against the complete 2025 CMS HCPCS/CPT database, enabling automated detection of hallucinated codes. The finding that no model produced non-existent codes, despite frequent misapplication, has not been systematically reported in prior work and suggests current frontier models have reliably internalized valid code structures.

A consistent finding across studies, including ours, is that procedural complexity drives error rates. The LWW study found hand procedures harder than sports medicine; Isch et al. found complex procedures dramatically harder than simple ones. Our data similarly showed that the number of billable components, rather than anatomical subspecialty, predicted both error rate and run-to-run variance. Single-code procedures achieved more perfect consistency on a run-to-run basis while multi-code procedures (e.g., Bunionectomy with 7 benchmark codes) commonly showed F1 scores below 40%.

4.4 Clinical implications

These findings suggest that current LLMs are not ready to autonomously assign CPT codes in clinical practice, but may have value as first-pass tools that generate candidate codes for human review. The 65.9% F1 achieved by the best-performing model means roughly one-third of codes were either incorrect or missing, an error rate unacceptable for autonomous billing. However, a workflow where LLMs generate initial code suggestions that trained coders or physicians then verify and correct could reduce cognitive burden without sacrificing accuracy.

Such human-in-the-loop workflows would need to account for the variance patterns observed in this study. For complex multi-component procedures, LLM suggestions should be treated with particular skepticism, as these cases showed both the highest error rates and greatest run-to-run inconsistency. Conversely, simple single-code procedures may benefit most from LLM assistance, as models demonstrated reliable (if not always correct) performance on straightforward cases. Importantly, the consistent errors observed, where all models predicted the same wrong code across multiple runs, highlight a risk in over-trusting LLM confidence. A model that consistently outputs the same incorrect code may appear reliable while systematically introducing billing errors. Human reviewers must maintain independent clinical judgment rather than defaulting to LLM suggestions simply because they appear deterministic.

4.5 Limitations

Several limitations should be considered when interpreting these results. First, the sample size was small (n=33 notes, 28 with benchmark labels), limiting statistical power and generalizability. Five notes were excluded from accuracy analysis: one (Superior Labrum Lesion Repair) was deemed ambiguous by the reviewing surgeon, and four (primarily spine-related procedures) fell outside the labeling surgeons' specialty scope. Interestingly, for the ambiguous note, all three models produced identical codes across all runs, suggesting AI confidence does not always reflect clinical certainty. Second, benchmark labels were assigned by a single surgeon per note without cross-validation, precluding an inter-rater reliability calculation. CPT coding often involves legitimate interpretive differences between experts; our benchmark represents one surgeon's reasonable interpretation rather than definitive ground truth. A multi-reviewer consensus approach would strengthen future studies. Third, the publicly available notes used in this study may differ systematically from real clinical documentation. Both reviewing surgeons noted that sample notes appeared more vague and less detailed than typical operative reports encountered in clinical practice. Performance on actual clinical notes (which tend to be more comprehensive) may differ. Fourth, this evaluation represents a point-in-time snapshot. LLM capabilities evolve rapidly with model updates; results obtained in December 2025 may not reflect current or future performance. Additionally, our prompting strategy, while standardized across models, may not have been optimal for each. The decision to omit temperature parameters for parity (since Anthropic and Google do not accept temperature with thinking modes enabled) may have disadvantaged OpenAI, whose default temperature settings could have introduced additional variance. Finally, our hallucination analysis validated only whether predicted codes existed in the HCPCS/CPT database, not whether code descriptions matched the model's stated reasoning. A model could output a valid code while providing

an incorrect description of what that code represents, constituting a form of hallucination not captured by the metrics we used in this study.

4.6 Future directions

Several avenues warrant further investigation. First, a larger validation study using real clinical operative notes across multiple institutions would establish whether these findings generalize beyond public sample databases. Such a study should include multi-reviewer benchmark labeling with inter-rater reliability assessment to establish more robust ground truth.

Second, comparative evaluation of thinking versus non-thinking model configurations could clarify whether extended reasoning capabilities improve coding accuracy or simply increase variance. Our study used maximum reasoning settings for all models; whether simpler, faster configurations achieve comparable results remains unknown. Third, the role of prompting optimization deserves systematic study. Our finding that models struggled with code quantity determination (e.g., reporting a procedure performed on multiple sites) may be addressable through prompt engineering that explicitly instructs models to consider laterality and multiplicity. Specialty-specific prompts incorporating domain conventions could potentially improve performance. Fourth, modifier assignment (not evaluated in this study) represents a logical next step. Given the variance observed in base code assignment, we anticipate that modifier selection (which adds another layer of interpretive complexity) would show even greater inconsistency. However, modifiers are essential for accurate billing and warrant dedicated evaluation. Finally, prospective studies of human-AI collaborative workflows could quantify whether LLM-assisted coding improves coder efficiency without compromising accuracy. Measuring time savings, error rates, and coder satisfaction in real billing environments would provide practical guidance for implementation.

V. CONCLUSION

This pilot study evaluated three frontier-class Large Language Models (Anthropic Claude Opus 4.5, OpenAI GPT-5.2, and Google Gemini 3 Pro) on the task of assigning CPT codes to orthopedic surgical procedure notes. Claude Opus 4.5 achieved the highest overall accuracy (F1: 65.9%), followed by Gemini 3 Pro (62.1%) and GPT-5.2 (56.8%). Notably, accuracy did not correlate with consistency: Gemini produced the most deterministic outputs despite lower benchmark alignment, while Claude demonstrated greater run-to-run variance yet achieved superior accuracy. No model generated hallucinated codes based on our narrow scope of this definition, indicating that current LLMs have internalized valid CPT code structures even when misapplying them to clinical contexts.

Performance was strongly influenced by procedural complexity. Simple, unambiguous procedures achieved the most near-unanimous consistent coding across all models, while multi-component procedures produced the most substantial errors and variance. These findings suggest that current LLMs might not yet be suitable for autonomous CPT code assignment in clinical practice. However, they may offer value as first-pass tools within human-in-the-loop workflows, generating candidate codes for

subsequent review by trained coders or physicians. Such implementations must account for the observed variance patterns and guard against over-trusting deterministic but incorrect outputs.

This study was limited by its small sample size, single-surgeon benchmark labeling, and reliance on publicly available notes that may differ from typical clinical documentation. Future research should include larger, multi-institutional validation studies with consensus-based benchmarking, systematic evaluation of prompting optimization, and prospective assessment of human-AI collaborative coding workflows in real billing environments.

Declarations

Disclaimers & Acknowledgments

AI Assistance. AI-assisted tools were used during manuscript preparation and code generation. AI tools were not used in CPT code assignment by the benchmark surgeons. The authors reviewed and edited all AI-generated outputs and take full responsibility for the content of this publication.

Conflicts of Interest. Abdalrahman Katranji, Aisa De Vries, and Dr. Abdalmajid Katranji are affiliated with Simplify AI, an AI medical scribe platform. Simplify AI utilizes API services from Anthropic, OpenAI, and Google in its commercial products. Dr. Zalzaleh is a client of Simplify AI. No company reviewed or influenced the study design, analysis, or manuscript preparation. The authors received no funding or compensation from any of the three companies evaluated in this study. Benchmark labeling was performed independently by each surgeon without input from Simplify AI staff or AI tools.

Funding. This research received no external funding.

Author Contributions. Abdalrahman Katranji conceived the study, conducted API testing & statistical analysis, and drafted the manuscript. Aisa De Vries assisted with data analysis and manuscript preparation. Dr. Katranji and Dr. Zalzaleh provided benchmark labeling and clinical interpretation.

References

1. P. Dotson, "CPT® codes: What are they, why are they necessary, and how are they developed?," Advances in Wound Care, vol. 2, no. 10, pp. 583–587, 2013, doi: 10.1089/wound.2013.0483.
2. Centers for Medicare & Medicaid Services. "Healthcare Common Procedure Coding System." CMS.gov. https://www.cms.gov/medicare/coding-billing/healthcare-common-procedure-system (accessed Dec. 24, 2025).
3. Z. Hou, H. Liu, J. Bian, X. He, and Y. Zhuang, "Enhancing medical coding efficiency through domain-specific fine-tuned large language models," npj Health Systems, vol. 2, no. 1, p. 14, May 2025, doi: 10.1038/s44401-025-00018-3.
4. S. Campbell and K. Giadresco, "Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals," Health

Information Management Journal, vol. 49, no. 1, pp. 5–18, Jan. 2020, doi: 10.1177/1833358319851305.

5. H. Dong et al., "Automated clinical coding: What, why, and where we are?," npj Digital Medicine, vol. 5, no. 1, p. 159, Oct. 2022, doi: 10.1038/s41746-022-00705-7.
6. R. Y. Lee et al., "Assessment of a zero-shot large language model in measuring documented goals-of-care discussions," medRxiv (Preprint), Sep. 2025, doi: 10.1101/2025.05.23.25328115.
7. S. Maity and M. J. Saikia, "Large language models in healthcare and medical applications: A review," Bioengineering, vol. 12, no. 6, p. 631, 2025, doi: 10.3390/bioengineering12060631.
8. *A. Garcia-Carmona, M. Prieto, E. Puertas, and J. Beunza, "Leveraging large language models for accurate retrieval of patient information from medical reports: Systematic evaluation study," JMIR AI, vol. 4, p. e68776, 2025, doi: 10.2196/68776.*
9. Anthropic. "Claude API Documentation." Anthropic. https://docs.anthropic.com (accessed Dec. 13, 2025).
10. OpenAI. "OpenAI API Documentation." OpenAI. https://platform.openai.com/docs (accessed Dec. 13, 2025).
11. Google. "Gemini API Documentation." Google AI for Developers. https://ai.google.dev/gemini-api/docs (accessed Dec. 13, 2025).
12. *LMSYS Org. "LMArena: Chatbot Arena Leaderboard." LMArena. https://lmarena.ai (accessed Dec. 13, 2025).*
13. *Anthropic. "Claude Opus 4.5." San Francisco, CA, USA: Anthropic, Nov. 2025. [Online]. Available: https://www.anthropic.com/news/claude-opus-4-5*
14. *OpenAI. "GPT-5.2." San Francisco, CA, USA: OpenAI, Dec. 2025. [Online]. Available: https://openai.com/index/introducing-gpt-5-2/*
15. *Google DeepMind. "Gemini 3: Introducing the latest Gemini AI model from Google." Mountain View, CA, USA: Google, Nov. 2025. [Online]. Available: https://blog.google/products/gemini/gemini-3*
16. *M. J. Carrarini, H. Y. Liu, C. K. Perez, and F. M. Egro, "Evaluating large language model's accuracy in current procedural terminology coding given operative note templates across various plastic surgery sub-specialties," Journal of Plastic, Reconstructive & Aesthetic Surgery, vol. 106, pp. 50–52, Jul. 2025, doi: 10.1016/j.bjps.2025.04.025.*
17. *E. L. Isch et al., "Bridging the coding gap: Assessing large language models for accurate modifier assignment in craniofacial operative notes," Journal of Craniofacial Surgery, vol. 36, no. 7, pp. 2260–2263, Oct. 2025, doi: 10.1097/SCS.0000000000011390.*
18. *MTSamples. "Medical Transcription Samples." MTSamples.com. https://mtsamples.com (accessed Dec. 25, 2025).*
19. *Medical Transcription Sample Reports. "Medical Transcription Sample Reports." MedicalTranscriptionSampleReports.com. https://www.medicaltranscriptionsamplereports.com (accessed Dec. 25, 2025).*

20. *D. E. Fulkerson, A. A. Haider, and D. E. Pereira, "Evaluating the accuracy and reliability of a large language model in coding common orthopaedic procedures," Current Orthopaedic Practice, 2025, online ahead of print. [Online]. Available: https://journals.lww.com/c-orthopaedicpractice/abstract/9900/evaluating_the_accuracy_and_reliability_of_a_large.210.*

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixD.pdf
- AppendixF.pdf
- AppendixB.pdf
- AppendixC.pdf
- AppendixA.pdf
- AppendixG.html
- AppendixE.zip