Evaluating the Accuracy and Reliability of a Large Language Model in Coding Common Orthopaedic Procedures

Daniel E Pereira, Daniel E Fulkerson[1], Ameer A Haider[1]
[1]Department of Orthopaedic Surgery

INTRODUCTION: Current Procedural Terminology (CPT) codes are critical for accurate billing and documentation in orthopaedic surgery, yet coding can be complex and costly. Large language models (LLMs), which utilize artificial intelligence (AI) to produce human-like responses to written text, have shown promise in automating CPT coding. However, the current orthopaedic literature has not evaluated LLM coding using only procedure names, nor has it compared LLM coding performance across subspecialties. This study evaluates the coding accuracy and reliability of a LLM when prompted and iteratively trained with the names of common orthopaedic procedures from the subspecialties of sports, foot and ankle, and hand.

METHODS: Ten orthopaedic procedures and the corresponding CPT codes were selected for the subspecialties of sports, foot and ankle, and hand. An institutional, Health Insurance Portability and Accountability Act (HIPAA)-compliant, ChatGPT-4o-based LLM was prompted to generate the appropriate CPT code based on a procedure name. Three trials in independent sessions were performed for each procedure. Outputs were categorized as “correct” (provided only the target code), “partially correct” (provided the target code and at least one other code), or “incorrect” (failed to provide the target code). These categories were assigned point values of 1, 0.5, and 0, respectively. Model performance was assessed as a percentage of the maximum possible score. An iterative prompting approach was employed for incorrect responses, after which the scoring process was repeated. Model performance was compared across subspecialties before and after iterative prompting. A post hoc trial was performed for one procedure that had consistently generated incorrect CPT codes.

RESULTS:

Ninety trials for 30 procedures were performed. The LLM generated the correct CPT code in 86.7% of trials, a partially correct CPT code in 4.4% of trials, and an incorrect CPT code in 8.9% of trials, corresponding to an overall model performance of 88.9%. The LLM consistently produced the correct CPT code for 22 procedures and an incorrect code for one procedure (“Open Reduction and Internal Fixation (ORIF) of intra-articular distal radius fracture”). Iterative prompting was performed for the eight trials that initially produced incorrect responses, after which model performance improved to 93.9%. LLM performance varied between the subspecialties before and after iterative prompting (p=0.018; p=0.012), with significantly lower scores for hand compared to sports procedures (p=0.019; p=0.017) and a trend towards worse performance for hand compared to foot and ankle procedures (p=0.055; p=0.075). The LLM had the most difficulty distinguishing between closely related CPT codes, particularly within the hand subspecialty. Preliminary and iterative trials for “ORIF of intra-articular distal radius fracture” failed to generate the target CPT code, instead providing a code corresponding to a different number of fracture fragments. A post hoc trial specifying the number of fragments and stating the LLM’s error successfully generated the target code.

DISCUSSION AND CONCLUSION: The present study demonstrates the accuracy and reliability of a LLM in generating CPT codes based on procedure names for common orthopaedic operations across multiple subspecialties. Further prompt refinement is needed to optimize model performance for complex and closely related procedures. However, our findings suggest the potential for LLMs to function as an effective tool for orthopaedic coding.

Figure 1. Representative LLM inputs and responses. (A) Sample LLM response that produced the correct CPT code. (B) Sample LLM response that produced an incorrect CPT code, with a prompt to try again.

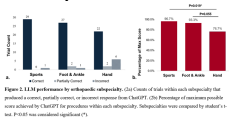

Figure 2. LLM performance by orthopaedic subspecialty. (A) Counts of trials within each subspecialty that produced a correct, partially correct, or incorrect response from ChatGPT. (B) Percentage of outcomes possible score achieved by ChatGPT for procedures within each subspecialty. Subspecialties were compared by student’s t-test. P<0.05 was considered significant (*).

Figure 3. LLM inputs and responses for post hoc trial for “Open Reduction and Internal Fixation of distal radius fracture”. The trial prompt was modified from the standard prompting approach by specifying that two fracture fragments were present. It was then informed of its specific error and prompted to try again.

Table 1. List of procedures and corresponding CPT codes by orthopaedic subspecialty.

Sports		Foot & Ankle		Hand	
Elbow Medial Collateral Ligament Ligament Reconstruction Using Tendon Graft	24346	Arthroplasty, Ankle with Implant	27702	Finger Tendon Transfer, Each Tendon (Original Finger Rotator)	26855
Shoulder Rotator Cuff Repair Using Arthroscopy (Medial or Lateral)	29882	Triple Arthrodesis	28715	Tenosynovitis with Tendon Interposition	25447
Arthroscopically Assisted Anterior Cruciate Ligament Reconstruction	29888	Arthrodesis, Great Toe, Metatarsophalangeal Joint Fusion	28750	Open Reduction and Internal Fixation, Intra-articular Distal Radius Fracture	25608
Arthroscopically Assisted Meniscectomy (Medial or Lateral)	29881	Open Reduction and Internal Fixation, Tarsal Fracture (Individual)	28415	Primary Flexor Tendon Repair at the Wrist or Forearm	25260
Arthroscopically Assisted Hip Labral Repair	29914	Bunion Correction with Double Osteotomy	28299	Neuroplasty or Transposition of the Median Nerve at the Carpal Tunnel	64721
Shoulder Arthroscopy with Debridement/Subacromial Decompression	29806	Open Ankle Arthrodesis	27870	Cubital Tunnel Decompression and Ulnar Nerve Transposition	64718
Arthroscopic Distal Clavicle Excision	29824	Open Primary Subtalar Arthrodesis	28725	Arthrodesis of the Metacarpophalangeal Joint	26850
Arthroscopic Rotator Cuff Repair	29827	Primary Achilles Tendon Repair	27650	Open Reduction and Internal Fixation, Phalanx Shaft Fracture	26755
Open Distal Bicep Repair with Autograft Tendon Graft	24342	Percutaneous Single Toe Tenotomy	28030	Fascial Release or Fasciectomy and Release of a Single Finger	26123
Latarjet Procedure, Anterior Shoulder Capsule Repair with Coracoid Transfer	23462	First Metatarsophalangeal Joint Cheilectomy (Hallux Rigidus Cheilectomy)	28289	Primary Ganglion Cyst Removed from Wrist	25111

Table 2. Repeat trials for incorrect LLM responses after iterative prompting.

Procedure	Target CPT Code	Initial Response	Repeat Response
Elbow Medial Collateral Ligament Reconstruction using Tendon Graft	24346	24345	**24346**
Open Reduction and Internal Fixation, Lisfranc Fracture Dislocation	28615	28485	**28615**
Open Reduction and Internal Fixation, Intra-articular Distal Radius Fracture	25608	25609	25609
Open Reduction and Internal Fixation, Intra-articular Distal Radius Fracture	25608	25609	25609
Open Reduction and Internal Fixation, Intra-articular Distal Radius Fracture	25608	25609	25607, **25608,** 25609
Fascial Release or Fasciectomy and Release of a Single Finger	26123	26055	26045
Fascial Release or Fasciectomy and Release of a Single Finger	26123	26043	**26123**
Primary Ganglion Cyst Removed from Wrist	25111	26500	**25111**