Automated Medical Coding: An AI-Powered Approach

Our Approach

1. Vector Database for Large Code Sets

One of the primary challenges in medical coding is dealing with extensive code sets. We had 74,209 ICD codes and 11,241 CPT codes, which together amount to approximately 2.5 million tokens when loaded into a single prompt. This far exceeds even the most advanced LLMs' context windows, including Gemini Pro's 2 million token capacity. Not only would this overwhelm the model's processing capabilities, but it's simply not possible to input such a large amount of data in a single prompt. To address this, we employ a vector database (LanceDB) that supports hybrid search, combining full text search and semantic similarity.

For modifiers, however, we don't use LanceDB. There are only 426 modifiers, which is a manageable number that can easily fit within the LLM's context window. This allows us to include the full set of modifiers directly in our LLM call, eliminating the need for an additional database query step and potentially improving processing speed for this particular code set.

Justification: This approach allows us to efficiently search and retrieve relevant codes without overwhelming the LLM or compromising on performance, while also optimizing the handling of the smaller modifier set. By using a vector database, we can work with our extensive code sets in a way that's both computationally feasible and effective for accurate code selection.

2. Keyword Extraction from Patient Charts

To enhance the search process, we use an LLM to extract keywords and phrases from patient charts. This step helps in focusing the search on the most relevant information.

Justification: Extracted keywords provide a more targeted input for the vector database search, improving the relevance of retrieved codes. Compared to sending the chart content as-is, this approach offers several advantages:

1. **Noise Reduction:** Patient charts often contain extraneous information. Keyword extraction filters out this noise, focusing on the most pertinent medical terms and conditions.
2. **Improved Search Efficiency:** By using targeted keywords, we can perform more precise searches in our vector database, potentially reducing the number of irrelevant codes retrieved.
3. **Contextual Understanding:** LLMs can extract not just individual words but also meaningful phrases, capturing the context that might be missed in a simple word-matching approach.

4. **Standardization:** Different doctors may use varying terminology. An LLM can help standardize these terms to match our code descriptions more closely.
5. **Handling of Synonyms:** LLMs can recognize and extract synonymous terms, improving the chances of matching with the correct codes even when the exact terminology in the code description isn't used in the chart.

While sending the chart content as-is might seem simpler, it could lead to less accurate code retrieval, especially for complex cases where the relevant information is buried within lengthy narratives. Our keyword extraction step strikes a balance between leveraging the full context of the chart and providing focused input for efficient code matching.

3. Targeted Code Retrieval

Using the extracted keywords and chart content, we query the vector database to retrieve a shortlist of 100 ICD and 100 CPT codes. This shortlist is much more manageable for the LLM to process.

Justification: This step significantly reduces the number of codes the LLM needs to consider, making the process more efficient and accurate.

4. Structured Data Extraction with Pydantic

We use Pydantic models to define the structure of the output XML schema. This approach, combined with the Instructor library, forces the LLM to respond in a fixed schema rather than free-form text.

Justification: Structured output ensures consistency and adherence to the required XML format, reducing the need for post-processing and error correction.

5. Final Code Assignment

The LLM is provided with the chart content, shortlisted ICD and CPT codes, and modifier codes. It then extracts the relevant information and returns a structured Claim object.

Advantages of Our Approach

1. **Scalability:** Can handle large code sets efficiently.
2. **Accuracy:** Reduces human error through AI-powered code selection.
3. **Speed:** Significantly faster than manual coding.
4. **Consistency:** Ensures uniform coding practices across different charts.
5. **Cost-effectiveness:** Reduces the need for human coders, lowering operational costs.

Alternative Approaches and Why Ours is Optimal

1. **Full LLM Processing:** Using an LLM to process the entire code set would be inefficient and prone to errors due to context window limitations.
2. **Rule-Based Systems:** While potentially faster, these lack the flexibility to understand nuanced medical language and context.
3. **Keyword Matching Only:** This might miss important context and lead to inaccurate code assignments.

Our hybrid approach combines the strengths of vector databases for efficient search, LLMs for understanding medical context, and structured data models for accurate output. This balance makes our solution optimal for handling the complexities of medical coding while maintaining efficiency and accuracy.

Challenges and Future Work

- Limited sample data: Currently working with 3 AI-generated sample charts and XMLs. Real-world testing with actual patient data will be crucial for validation.
- Continuous updates: Medical codes are regularly updated. Our system is already built to accommodate these changes efficiently.
- Fine-tuning for specialties: Different and rare medical specialties may require specialized models or additional training data.

Conclusion

Our AI-powered approach to medical coding offers a promising solution to automate and improve the accuracy of the coding process. By leveraging advanced technologies like vector databases and large language models, we can significantly reduce the time and cost associated with medical coding while potentially improving accuracy. As we move forward, real-world testing and continuous refinement will be key to ensuring the system's effectiveness across various medical scenarios.