NYU Center for Data Science

AI-Driven Reimbursement Code Discovery for HealthTech Startups

Xingyin Xu, Tongyu Zhao, Albert Kong, Erchi Zhang
Mentor: Moshe Bellows, Sophie Hao

Introduction

In the complex landscape of healthcare billing, accurately matching medical services to appropriate CPT (Current Procedural Terminology) codes is a critical yet challenging task. Healthcare professionals and startups often struggle with the time-consuming and error-prone process of identifying the correct codes, which can lead to billing inaccuracies, delayed reimbursements, and compliance issues. The prevailing method for identifying appropriate CPT codes often involves engaging third-party consultants, which incurs significant costs in both labor and time. This challenge is compounded by the complexity and variability of language used in medical documentation and pitch decks, requiring a nuanced understanding to ensure accurate code selection.

In this project, we aim to address this problem by providing an fast, efficient, and reliable tool to streamline CPT code matching, empowering users to focus on delivering quality care and innovation. In this project, we developed a web application that helps healthcare startups match appropriate CPT codes. Users can upload PDF documents through a web interface and choose between two analysis methods: a keyword-based matching system using TF-IDF, or an AI-powered analysis using GPT-4. The application processes the uploaded documents, and returns the top 5 most relevant CPT codes with their descriptions. This tool is particularly useful for healthcare startups and medical professionals who need to quickly identify appropriate medical billing codes based on service descriptions.

Data and Preprocessing

- **CPT Dataset**
  - We extracted over 10,000 CPT codes and their text descriptions from healthcare website (Codify by APPC). After obtaining a CSV file, we cleaned and imputed the raw data. Finally, we converted the cleaned CSV to a pickle file for faster loading.
- **User's Uploaded PDF**
  - When user uploads a PDF file, we extract text from the PDF. Then, if the user selects to use AI analysis, we apply several privacy-preserving transformations to the extracted text. We use regex patterns and spaCy NER package to remove sensitive personal/company informations including email addresses, phone numbers, and company names. This comprehensive preprocessing is essential when leveraging generative AI, since it ensures that sensitive information is removed before the text is sent to the GPT-4 API for CPT code analysis, while preserving the relevant healthcare service descriptions. We also obtained consents from relevant personnels to upload their materials to third-party open-source tools.

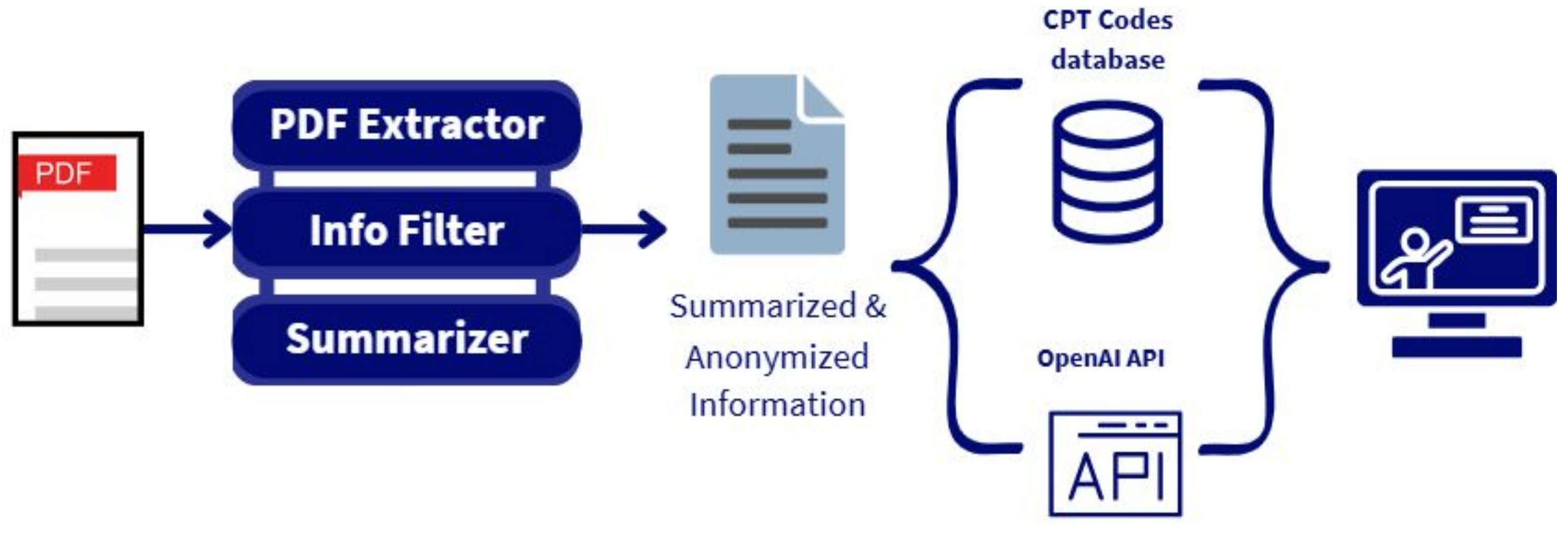


Figure 1. Overall Workflow

Methodologies

Approach I: Similarity Based Retrieval

- The extracted text from PDF is preprocessed using TF-IDF vectorization which removes English stop words, converts text to lowercase, then creates a term-frequency inverse-document-frequency matrix. This preprocessing pipeline ensures that the text comparison is done on cleaned, standardized data that can be effectively compared using cosine similarity.
- We experimented with several different matching algorithms, including direct text fuzzy match, Word2Vec embedding with Word2Vec similarity, BERT embedding with cosine similarity, named entity matching, topic modelling with Latent Dirichlet Allocation (LDA), etc., the results are stored in json format and can be printed out as tables on our web app for better visualization.

Approach II: AI Analysis

- Since we are using LLMs in this approach, we do not need to prepare any dataset. The PDF has already been preprocessed and anonymized.
- The processed text is sent to the GPT-4 API for CPT code analysis. By using self-designed well-constructed prompt, we are able to command GPT to identify key service information among the messy text, and summarize these key information into complete sentences and paragraphs.
- Afterward, the generated summary will be used as a short-term memory and further processed through prompt to be matched with corresponding CPT codes. The prompt asks for the top 5 matched CPT codes sorted by relevance to the startup's offerings, and output the code result and corresponding code descriptions in JSON format.

Results

- **Evaluation Metrics**
  - The performance of our tool is evaluated using recall and precision within the top 5 results, complemented by qualitative feedback from key stakeholders.

Collecting CPT codes for health product companies is challenging due to proprietary information. Thanks to the companies who voluntarily provided us the CPT codes their services are currently using, we are able to utilize them as ground truths to compare to our methods' results.

	Similarity Based Retrieval	AI-analysis
Covered Percentage	18.2%	80%
Precision@5	0.12	0.52
Recall@5	0.18	0.80

Table 1. Experiment Results

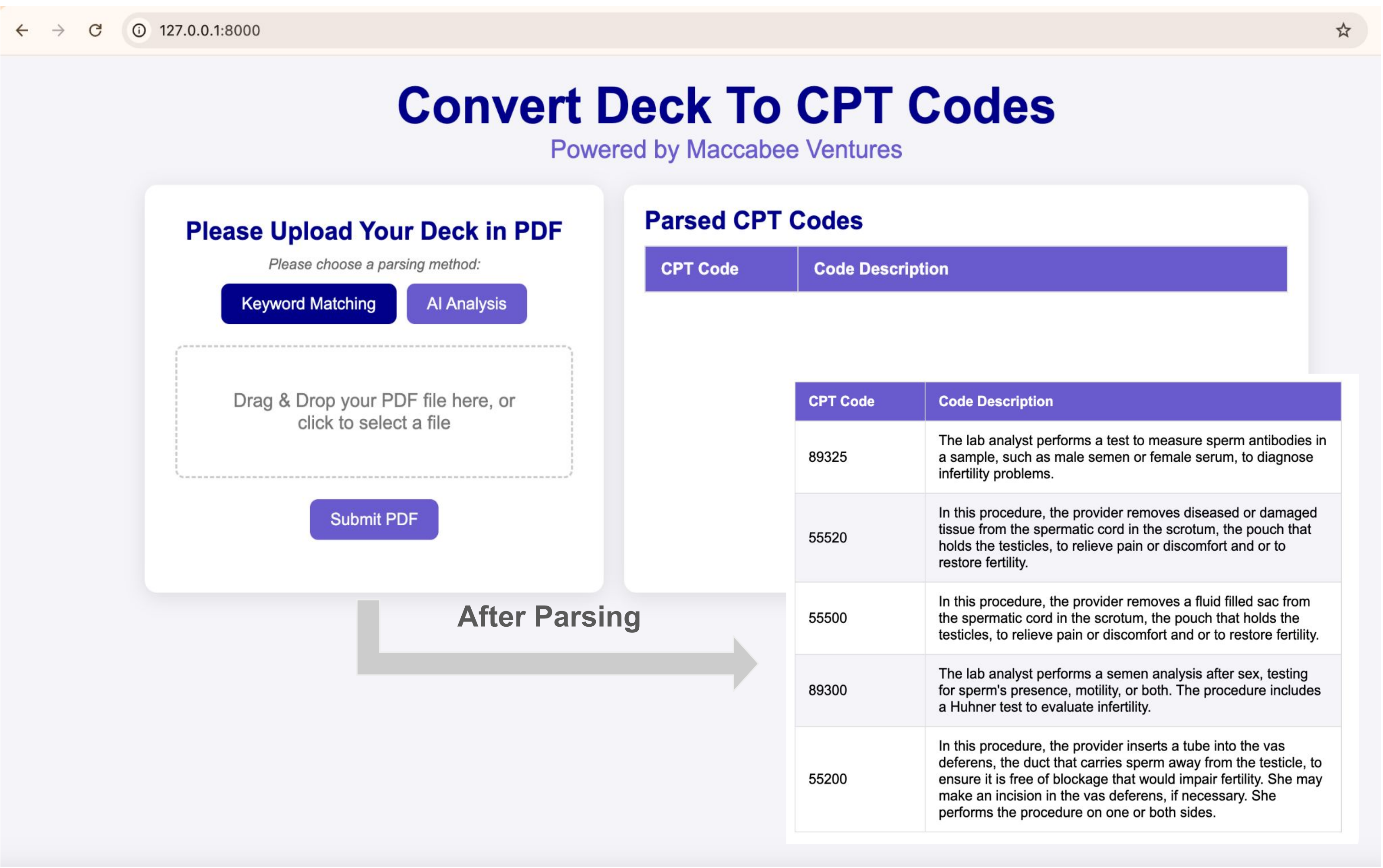
According to Table 1, Generative LLM significantly exceeded the performance of similarity based retrieval across all metrics, making it more effective for identifying relevant CPT codes accurately and consistently.

- **User Feedback**
  - "Yes 100% these codes are appropriate or relevant to our offerings!"
  - "This is a great solution that would provide value to digital health companies."
  - "I absolutely love this idea."
  - "I think this is great..."

Web Interface

Web Design

- **Interface Layout**
  - The webpage provides two options for parsing PDF documents: "Keyword Matching" (Approach I) and "AI Analysis" (Approach II).
  - Users can select "Keyword Matching" (default option), upload a file via drag-and-drop or file selection, and click "Submit PDF" to parse it using the TF-IDF function. Results appear on the right side.
  - For "AI Analysis", users must input a company name (to prevent data leakage) and upload their file similarly. Clicking "Submit PDF" invokes LLM-based parsing, with results displayed on the right.
  - The key difference is that AI Analysis requires a company name input and has a slightly longer wait time.



Convert Deck To CPT Codes

Powered by Maccabee Ventures

Please Upload Your Deck in PDF

Please choose a parsing method:

Drag & Drop your PDF file here, or click to select a file

Parsed CPT Codes

CPT Code	Code Description

CPT Code	Code Description
89325	The lab analyst performs a test to measure sperm antibodies in a sample, such as male semen or female serum, to diagnose infertility problems.
55520	In this procedure, the provider removes diseased or damaged tissue from the spermatic cord in the scrotum, the pouch that holds the testicles, to relieve pain or discomfort and or to restore fertility.
55500	In this procedure, the provider removes a fluid filled sac from the spermatic cord in the scrotum, the pouch that holds the testicles, to relieve pain or discomfort and or to restore fertility.
89300	The lab analyst performs a semen analysis after sex, testing for sperm's presence, motility, or both. The procedure includes a Huhner test to evaluate infertility.
55200	In this procedure, the provider inserts a tube into the vas deferens, the duct that carries sperm away from the testicle, to ensure it is free of blockage that would impair fertility. She may make an incision in the vas deferens, if necessary. She performs the procedure on one or both sides.

Figure 2. Before & After submission (Keyword Matching)

Future Work

- **Expand the Pitch Deck Repository**
  - Collaborate with healthcare startups, medical organizations, and billing consultants to collect a more diverse and extensive set of pitch decks.
- **Curating Annotated Ground Truth**
  - Work with domain experts to establish a benchmark dataset with annotated ground truth CPT codes for accurate evaluation and continuous improvement.
- **New Version Releases of the Web Application**
  - Enhance the webpage with additional functionalities, including user sign-up and login, the ability to retrieve CPT checking histories, providing detailed explanations for why the CPT codes are generated, and so forth.
- **Interactive Feedback Mechanism**
  - Enable users to validate or reject suggested CPT codes. Use this feedback to improve both AI and keyword-based systems over time.