Surpassing GPT-4 Medical Coding with a Two-Stage Approach

Zhichao Yang zhichaoyang@umass.edu
University of Massachusetts, Amherst, MA, USA

Sanjit Singh Batra sanjit.batra@optum.com
Joel Stremmel joel_stremmel@optum.com
Eran Halperin eran.halperin@uhg.com
Optum AI, Minnetonka, MN, USA

Abstract

Recent advances in large language models (LLMs) show potential for clinical applications, such as clinical decision support and trial recommendations. However, the GPT-4 LLM predicts an excessive number of ICD codes for medical coding tasks, leading to high recall but low precision. To tackle this challenge, we introduce LLM-codex, a two-stage approach to predict ICD codes that first generates evidence proposals using an LLM and then employs an LSTM-based verification stage. The LSTM learns from both the LLM's high recall and human expert's high precision, using a custom loss function. Our model is the only approach that simultaneously achieves state-of-the-art results in medical coding accuracy, accuracy on rare codes, and sentence-level evidence identification to support coding decisions without training on human-annotated evidence according to experiments on the MIMIC dataset.

Keywords: Natural Language Processing, Large Language Models, Generative Models, Semi Supervised Learning, Explainability, Interpretability

1. Introduction

Clinical text encompasses a vast array of essential information that extends beyond the structured data fields obtained from electronic health records (EHRs) (Zweigenbaum et al., 2007; Uzuner et al., 2010; Wang et al., 2018; Yao et al., 2022; Li et al., 2022; Jiang et al., 2023). A critical task in EHR analysis is the assignment of International Classification of Diseases (ICD) codes (Larkey and Croft, 1996), which entails attributing zero, one, or multiple ICD codes to a given note.

Computational methods have been employed to automate the task of ICD coding. Ideally, such computational methods should overcome the following challenges: (1) The first challenge is the scarcity of training data since labeling EHRs is an expensive process (Wei et al., 2018; Willemink et al., 2020), often resulting in a scarcity of sufficient training data. (2) The second challenge is achieving high precision and recall for all ICD codes, including rare ones, as they may hold equal clinical importance for patients as common codes (Atutxa et al., 2019; Dong et al., 2021). (3) The third challenge is explainability since it is crucial in the medical field to ensure trust in the classifier's decisions. Consequently, computational methods should be capable of providing sentence-level evidence to support their coding decisions.

Unfortunately, existing computational methods for medical coding fail to address all three critical issues concurrently. In particular, state-of-the-art medical coding models are unable to provide *sentence-level* evidence for their coding decisions due to their black-box nature (Yuan et al., 2022; Jain and Wallace, 2019). While some models do offer such sentence-level evidence, they necessitate training on annotated evidence, which requires substantial human annotation costs (Cheng et al., 2023).

Recent studies have demonstrated that large language models (LLMs) can serve as effective few-shot learners when training examples are limited (Zhao et al., 2021; Min et al., 2022; Chen et al., 2022). Furthermore, LLMs can be directly prompted for evidence to support their medical coding decisions, making them well-suited for this task (Agrawal et al., 2022). However, we observe that state-of-the-art LLMs, such as GPT-4, exhibit low precision in medical coding tasks, as depicted in Figure 1. As a result, there is currently no method that effectively addresses all three of these challenges simultaneously.

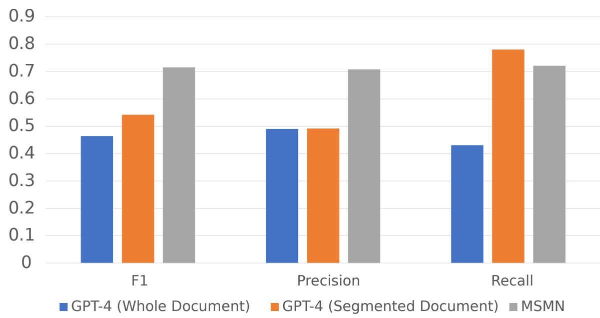


Figure 1: An initial experiment on the accuracy of ICD coding compared between few-shot ICL GPT4 (OpenAI, 2023) and fine-tuned MSMN

In this paper, we propose a two-stage approach, **LLM-codex**, that addresses all three challenges simultaneously. This approach attains state-of-the-art medical coding accuracy even with limited training data and rare codes. Additionally, LLM-codex furnishes precise sentence-level evidence for coding decisions without necessitating training on annotated evidence.

LLM-codex is a two-stage approach consisting of an LLM in the first stage and a *Verifier* model in the second stage. In the first stage, we segment long EHRs into smaller segments and feed each segment into the LLM. While this strategy substantially improves recall, it leads to lower precision due to the over-prediction of ICD codes. Consequently, in the second stage, we introduce an additional filter—a *Verifier* model—which verifies the predicted ICD codes (Zaidan et al., 2007). Our Verifier model is an LSTM trained with a custom loss function leveraging dual labels: the LLM-assigned ICD code as a sentence-level, silver-label (high recall), and the expert-assigned ICD code as the document-level, gold-label (high precision). The Verifier is designed to assign scores to each sentence based on its ability to predict the corresponding ICD code.

Incorporating the LLM in the first stage and the Verifier model in the second stage, LLM-Codex attains a substantial improvement of over 10% in F1 score for rare codes relative to state-of-the-art medical coding models. Furthermore, it exhibits about a 5% increase in F1 score on limited training data. Additionally, without requiring training on annotated evidence, LLM-Codex boosts evidence accuracy by over 10% when compared to the top-performing sentence-level evidence model for coding decisions.

As a result, LLM-Codex presents a comprehensive solution that tackles all three aforementioned issues concurrently, positioning itself as a promising framework for medical coding. We believe LLM codex can potentially be used on classification tasks beyond the medical domain that require providing supporting evidence for classification decision (Samek et al., 2017).

2. Related work

Automated ICD coding employs natural language processing (NLP) models to predict expert-labeled ICD codes using EHRs as input. This problem has traditionally been formulated as a multi-label classification task. Early approaches, such as CAML (Mullenbach et al., 2018), utilized a convolutional neural network to encode medical documents, followed by a label-wise attention mechanism to focus on the labeled ICD codes of the input notes during training. More recently, state-of-the-art methods have incorporated various techniques, such as incorporating synonyms of clinical concepts (Yuan et al., 2022; Yang et al., 2022b), exploring the discourse structure within EHRs (Zhang et al., 2022), and utilizing data augmentation (Falis et al., 2022) to enhance performance. Additionally, advancements in the field have emerged from exploring alternative architectures, such as pretrained bidirectional language models (Huang et al., 2022; Michalopoulos et al., 2022) and pretrained autoregressive language models combined with prompts (Yang et al., 2022a). In this paper, we propose a novel method, LLM-codex, to address the limitations of existing methods in automated ICD coding by leveraging a two-stage approach that significantly improves performance on rare coding labels.

The application of LLMs to unstructured clinical data has been a major focus of recent research (Jimenez Gutierrez et al., 2022; Zhou et al., 2022; McInerney et al., 2023). For instance, Agrawal et al. (2022) demonstrated that LLMs can effectively extract information from clinical text, even without training on clinical data. Likewise, Meoni et al. (2023) emphasized the potential of LLMs for information extraction tasks in the clinical domain, particularly when data is scarce due to confidentiality concerns arising from stringent privacy regulations that protect sensitive patient information. However, recent studies have found that LLMs struggle to extract

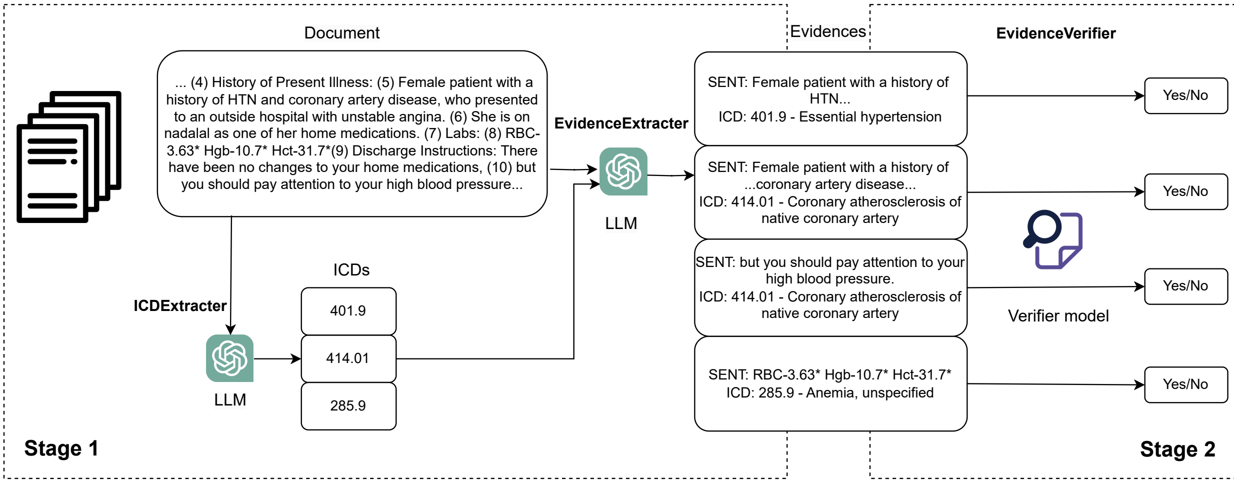


Figure 2: An illustration of LLM-codex where we use an LLM to extract code-evidence pairs and then verify them with a *Verifier* model. The examples are artificial and for demonstration only.

information when tasks necessitate accessing relevant information within lengthy contexts (Liu et al., 2023). We address this challenge by segmenting long documents to enhance their flow and readability.

To elucidate the reasons for assigning an ICD code to a document, previous research has primarily relied on attribution maps, derived either from the salience of individual words or the attention weights of specific tokens (Mullenbach et al., 2018; Lovelace et al., 2020; Dong et al., 2020; Liu et al., 2021; Kim et al., 2022; Wang et al., 2022; Nguyen et al., 2023b). However, these attribution maps exhibit limited explanation accuracy (Sinha et al., 2021; Ivankay et al., 2022). Ivankay et al. (2023) observed that when minor perturbations (modifications to a single task-irrelevant phrase or sentence) were introduced to a medical document, many words with initially positive attributions shifted to negative values, despite the code prediction remaining accurate. This issue resonates with the findings of Jain and Wallace (2019), who argued that attention-based explanations might not provide a complete understanding of model decisions. Their research demonstrated that attention weights can be easily manipulated without significantly affecting model predictions, that the same model with different attention weights could produce identical predictions, and that attention weights might remain unchanged even when input perturbations change the model's output. These findings, along with the limitations of attribution maps, emphasize the need for more reliable interpretability methods in ICD coding and other NLP tasks. In this paper, we address these challenges by proposing LLM-codex, which identifies the most relevant sentence from a long document when predicting an ICD code and subsequently verifies this sentence to produce a final prediction. This strategy is inspired by previous research demonstrating that the verification of LLMs can improve their output (Weng et al., 2022).

3. Datasets

We utilized several datasets to evaluate the model's coding performance and explainability.

MIMIC-III *common*: MIMIC-III (Johnson et al., 2016) is a publicly accessible dataset containing discharge summary documents from an Intensive Care Unit (ICU), with each document associated with ICD codes labeled by medical coding experts. In line with prior work (Mullenbach et al., 2018), we filtered the dataset to retain instances featuring at least one of the top 50 most frequent ICD codes. This results in 8,067 training instances and 1,729 test instances based on the canonical data splits from Mullenbach et al. (2018).

MIMIC-III *few-shot*: To assess the model's performance under limited training data conditions, we randomly selected about one-eighth of the instances from the training data. This resulted in 1,000 training instances and 1,729 test instances. This subset comprises the top 50 most frequent ICD codes and $\sim$ 14 training instances per label (shot) on average, adhering to the few-shot criteria.

MIMIC-III *rare*: To assess the model's performance on predicting rare disease codes, which could be of equal importance as common disease codes, for a given patient, we built a rare code dataset using MIMIC-III. We collected rare diseases defined by medical experts (Pavan et al., 2017; Wakap et al., 2019), and followed the pre-processing steps described in Yang et al. (2022b). This resulted in $\sim$ 5 training instances per label (shot) on average.

MDACE Profee: For evaluating the model's explainability, we used the code evidence dataset from Cheng et al. (2023). Expert annotators labeled a short text span for each ICD code assigned, indicating the rationale behind the assignment. The MIMIC-III dataset was annotated under professional fee billing guidelines, resulting in the *MDACE Profee* datasets. We subsequently mapped each annotated text span to a sentence, serving as evidence for evaluation purposes. There are 172 sentence-ICD pairs in the evaluation dataset.

4. Methods

4.1. Task formulation

ICD coding is typically formulated as a multi-label classification task, wherein the objective is to assign a binary label $y_{c,k} \in \{0, 1\}$ for each ICD code c in the label space Y, given thousands of words from an input EHR document k. A label of 1 indicates that a medical document is positive for a specific ICD code. Candidate ICD codes can be described using a short code description phrase in free text, such as the description "essential hypertension." corresponding to the ICD code 401.9. In addition to assigning the correct code, the goal is to also extract sentence-level evidence m from the document for each c to explain the model's decision.

To address these two tasks, we first employed an LLM to identify sentence-level evidence for all candidate ICD codes (Section 4.2 and Section 4.3). Subsequently, we used the ICD codes predicted by the LLM as *silver* labels to train a *Verifier* model that verifies whether the sentence-level evidence is accurate for the given ICD (Section 4.4).

4.2. Stage 1a: Extracting document-level ICD codes using an LLM

Utilizing an LLM such as GPT-4 (OpenAI, 2023) with in-context learning (ICL) necessitates the specification of:

a) A template for providing input documents via the prompt;

b) An LLM to execute the prompt and generate output text;

c) A parser to convert the output text into the task-specific output space.

Thus, we first used the LLM to extract ICD codes using ICL. To achieve this, we carefully designed our prompt templates using a single ICL example. As depicted in Example 1, the template instructed the LLM to emulate a proficient clinical coding expert and assign a list of ICD codes to the given document. In order to effectively manage long documents, we first split it into multiple segments containing an equal number of sentences and passed each segment individually to the LLM. The LLM then predicted which of the candidate ICD codes are present in each segment, in the form of free text which was then parsed into a Python list of predicted ICD codes. Finally, we aggregated the ICD code predictions obtained from each EHR segment to generate the LLM's document-level ICD code predictions.

4.3. Stage 1b: Extracting sentence-level ICD code evidence using an LLM

Given the document-level ICD code predictions, we used the LLM to identify sentence-level evidences for each predicted ICD code. Similar to the document-level ICD code prediction, we split the EHR into multiple segments containing an equal number of sentences. As demonstrated in Example 2, the template guided the LLM to emulate an evidence extraction expert by scanning each sentence in the segment and assigning one or more of the predicted document-level ICD codes to each sentence.

The LLM's output was subsequently parsed and aggregated across the segments within an EHR[1] to generate a Python list of tuples, wherein each tuple comprised an ICD code and its corresponding evidence sentence index.

1. https://api.python.langchain.com/en/latest/chains/langchain.chains.llm.LLMChain.html

4.4. Stage 2: Verifying sentence-level evidence using a *Verifier* model

Upon extracting pairs of predicted ICD codes c and evidence sentences m using the LLM, we verified the relationship between the pairs with the help of a *Verifier* model (Zaidan et al., 2007). The *Verifier* model assessed the accuracy of a silver label (which consists of an ICD code and its corresponding evidence sentence index) predicted by the LLM. To accomplish this, LLM-codex first split the document into sentences and subsequently ranked these sentences based on their relevance to predicting the document-level *gold* labels y_c on each ICD code c. Additionally, it incorporated supervision from LLM-assigned *silver* labels y'_c for each sentence, during the ranking process.

We denote the set of sentence-level evidences corresponding to the *silver* labels obtained by the LLM for the k-th document as:

$$m_k = [m_{k,1}, ..., m_{k,j}, ..., m_{k,S_k}] \tag{1}$$

where S_k is the total number of sentence-level evidences identified by the LLM in document k.

We then used the *Verifier* model iteratively across each predicted document-level ICD code c, to verify which of the predicted sentence-level evidences truly correspond to c. We therefore represented the predicted *silver* labels for the k-th document as $x_{c,k}$ where:

$$x_{c,k} = [(m_{k,1}, y'_{c,k,1}), ..., (m_{k,S_k}, y'_{c,k,S_k})] \tag{2}$$

where $y'_{c,k} \in \mathbb{R}^{S_k}$ and $y'_{c,k,j}$ was 1 if and only if, $m_{k,j}$ was predicted to have evidence for c in the *silver* labels.

The *Verifier* model consists of a text encoder TE which transforms a sentence-level evidence $m_{k,j}$ into its latent representation, h^m_j, using the following:

$$h^m_j = TE(m_{k,j}) \tag{3}$$

We followed MSMN (Yuan et al., 2022) and used an LSTM (Hochreiter and Schmidhuber, 1997) as text encoder TE. It also transforms the short ICD code description, $c_{description}$, of code c, into its latent representation, h^c, using the following:

$$h^c = TE(c_{description}) \tag{4}$$

The per-label-attention AT then combines the latent representations computed above to obtain label-specific logits $z_{k,j}$ (Mullenbach et al., 2018; Liu et al., 2021; Yuan et al., 2022):

$$z_{k,j} = AT(h^m_j, h^c) \tag{5}$$

where $z_{k,j} \in \mathbb{R}^2$ because each label takes on one of two binary values in the ICD coding task.

The loss function corresponding to the *Verifier* model was designed to consist of two terms, l_{gold} and l_{silver}. Inspired by Clark and Gardner (2018) and Min et al. (2019), the first term, l_{gold}, can be written as the weighted sum of losses corresponding to each sentence-level evidence, as follows:

$$l_{gold} = \sum_{j=1}^{S_k} w_{k,j} l_{k,j} \tag{6}$$

where, $l_{k,j}$ is the cross-entropy loss computed using $z_{k,j}$ and the document-level *gold* label $y_{c,k}$ corresponds to the ICD code c on document k.

In order to compute the weight $w_{k,j}$, we first performed a maximum operation over the two dimensions of $z_{k,j}$ and then normalized them across j using a softmax function. Therefore,

$$w_{k,j} = softmax(max(z_{k,j})) \tag{7}$$

The second term in the loss function, l_{silver} uses the *silver* labels, y'_k, and can be written as:

$$l_{silver} = \sum_{j=1}^{S_k} l'_{k,j} \tag{8}$$

where $l'_{k,j}$ computes the cross-entropy loss between $y'_{c,k,j}$ and confidence score logits $z'_{k,j}$. To obtain $z'_{k,j}$ we first computed a maximum over the two dimensions of $z_{k,j}$:

$$z'_{k,j} = max(z_{k,j}) \tag{9}$$

Finally, we trained the *Verifier* model with the total loss for the k-th document, corresponding to ICD code c as follows,

$$L_{k,c} = l_{gold} + l_{silver} \tag{10}$$

To make predictions for code c in document k, we first select the sentence index j with the highest weight $w_{k,j}$ among all candidate sentences m_k. If the *argmax* over the two-dimensional $z_{k,j}$ corresponds to the positive label, we then output its corresponding value as the prediction score for the code c.

4.5. Baselines for benchmarking

1. CAML (Mullenbach et al., 2018) uses a convolutional layer to extract features from an EHR and an attention mechanism to select the most relevant part of the EHR for predicting each ICD code.

2. MSMN (Yuan et al., 2022) uses code description synonyms with multi-head attention and achieves state-of-the-art performance on the MIMIC-III *common* task.

3. EffectiveCAN with supervised attention (Cheng et al., 2023) employs a convolutional attention network to train on both document-level labels and evidence annotations using supervised attention. Their evidence annotations are generated by clinical coding experts, in contrast to our evidence (*silver*) labels which are obtained from an LLM.

4. Medalpaca (Han et al., 2023) is a 13 billion parameter LLM trained to answer 1.5 million medical question. We replaced GPT-4 with this model to see how different LLMs performs.

5. Results

5.1. Predicting document-level common ICD codes with limited training data

Table 1: Coding performance on MIMIC-III *few-shot*. Mean and standard deviation over 20 experiments are shown.

Model	ROCAUC		F1	
	MACRO	MICRO	MACRO	MICRO
CAML	0.665 ±0.003	0.729 ±0.004	0.258 ±0.007	0.364 ±0.014
MSMN	0.833 ±0.012	0.874 ±0.007	**0.489** ±0.010	0.561 ±0.006
EffectiveCAN	0.802 -	0.871 -	0.434 -	0.556 -
Medalpaca	0.435 ±0.011	0.636 ±0.009	0.189 ±0.023	0.224 ±0.015
LLM-codex	**0.834** ±0.006	**0.911** ±0.005	0.468 ±0.017	**0.611** ±0.015
LLM-codex /wo silver	0.511 ±0.011	0.737 ±0.004	0.169 ±0.003	0.382 ±0.011

First, we benchmarked LLM-codex on the medical coding task using the MIMIC-III few-shot dataset described in Section 3. LLM-codex achieved a micro F1 of 0.611, which represents $\sim 5\%$ (absolute) improvement over existing approaches (Table 1), and a micro ROCAUC of 0.911, $\sim 3\%$ (absolute) improvement over all existing methods (Table 1). Similar performance improvements were observed for ICD-10 prediction with limited training data (Table A.1).

We also found that removing the silver labels obtained using the LLM from LLM-codex's training process led to a significant decline in ICD coding prediction metrics, highlighting their crucial role in its performance.

We also found that different LLMs performed very differently, GPT-4 performed the best among all baselines while Medalpaca performed the worst in ROCAUC and F1 score.

To investigate the impact of training data quantity on LLM-codex's performance, we benchmarked it on the MIMIC-III *common* dataset with 3 different size of training data. When trained with all 8066 instances, LLM-codex performed on par with existing methods in terms of coding accuracy (Table A.3). When trained as few as 1000 and 500 instances, LLM-codex outperformed existing methods. This robustness highlights LLM-codex's potential with constrained data resources.

5.2. Predicting document-level rare ICD codes

Table 2: Coding performance on the MIMIC-III *rare*

Model	ROCAUC		F1	
	MACRO	MICRO	MACRO	MICRO
CAML	0.574 ±0.004	0.602 ±0.003	0.072 ±0.006	0.083 ±0.004
MSMN	0.755 ±0.002	0.761 ±0.002	0.169 ±0.002	0.173 ±0.003
LLM-codex	**0.825** ±0.003	**0.832** ±0.002	**0.279** ±0.004	**0.302** ±0.005

To evaluate LLM-codex's performance on rare ICD codes, we assessed it using the MIMIC-III *rare* dataset described in Section 3.

We found that LLM-codex achieved an absolute improvement of $\sim 12\%$ in micro F1 and $\sim 5\%$ in micro ROCAUC compared to existing approaches (Table 2).

These results further support the notion that LLMs are effective few-shot learners, capable of outperforming existing classification models fine-tuned for rare ICD code prediction (Lewis et al., 2020; Taylor et al., 2022; Shyr et al., 2023).

5.3. Ablation study of LLM-codex

Table 3: Ablation study of LLM-codex on MIMIC-III *few-shot*. Micro scores are reported.

Model	F1	Precision	Recall
Blackbox CAML	0.365 ±0.014	0.349 ±0.005	0.383 ±0.009
Blackbox MSMN	0.561 ±0.006	0.545 ±0.010	0.581 ±0.019
LLM-codex	**0.611** ±0.015	0.587 ±0.015	0.638 ±0.015
LLM-codex (stage 1) + random forest	0.339 ±0.016	**0.648** ±0.024	0.230 ±0.013
LLM-codex (stage 1)	0.493 ±0.010	0.388 ±0.011	0.674 ±0.011
LLM-codex (stage 1a)	0.360 ±0.009	0.233 ±0.007	**0.792** ±0.009

To better understand the factors contributing to LLM-codex's predictive performance, we compared it to three variants: one using the LLM only for ICD code extraction (Stage 1a), another without the *Verifier* model (Stage 1), and a third where the *Verifier* model was replaced by a random forest classifier. In the last variant, we counted the occurrence of evidence sentences per ICD code for the ICD code and evidence sentence index pairs extracted by the LLM and then used the occurrence matrix as features to train the random forest for ICD code verification.

We present the results on the MIMIC-III few-shot dataset in Table 3 and make the following observations:

1. Implementing only Stage 1a of LLM-codex resulted in a significant decline in F1 score for ICD code prediction.

2. Including Stage 1b, which extracts sentence-level evidence for predicted ICD codes, improved the F1 score by $\sim 13\%$ (absolute) compared to LLM-codex with just Stage 1a.

3. Substituting the Verifier model with a random forest model led to a reduction in the F1 score by $\sim 27\%$ (absolute) compared to LLM-codex.

In summary, both stages of LLM-codex significantly contributed to its ICD coding predictive performance.

5.4. Ablation study of EHR segmentation on GPT-4

We examined two distinct methodologies for prompting GPT-4 to perform ICD coding prediction: one involved presenting the entire document, while the other presented 10 equal-sized sentence segments of the document and aggregated the results across these segments. The latter approach (GPT4-seg) significantly increased recall (while maintaining comparable precision) compared to using the whole document as input (GPT4-doc) (Table 4). This finding aligns with literature reports that LLMs face challenges in extracting information from the middle of long contexts (Liu et al., 2023). Despite this increase in recall, LLM-codex outperformed both methods in terms of F1 score on ICD code prediction.

Table 4: Ablation of EHR segmentation on MIMIC-III *few-shot* ICD code prediction with GPT-4

Model	F1	Precision	Recall
GPT4-seg	0.582 ±0.010	0.482 ±0.011	**0.730** ±0.011
GPT4-doc	0.484 ±0.015	0.500 ±0.009	0.471 ±0.010
LLM-codex	**0.611** ±0.015	**0.587** ±0.015	0.638 ±0.015

5.5. Predicting sentence-level evidence for common ICD codes

To assess LLM-codex's explainability capabilities, we utilized the MDACE Profee dataset Cheng et al. (2023) outlined in Section 3, which comprises ICD code evidence annotations created by professional medical coders. For each ICD code, LLM-codex provides a single sentence-level evidence if the predicted score exceeds a threshold optimized for that ICD code based on the F1 score of a validation dataset. LLM-codex selects the sentence-level evidence with the highest confidence score generated by the *Verifier* model. We considered the evidence for each ICD code in an EHR as a true positive if a method captured at least one of its expert-annotated sentence-level evidences from the MDACE study (Glockner et al., 2020).

Our findings indicated that LLM-codex yielded sentence-level evidences with the highest precision compared to existing methods like EffectiveCAN, which were trained on evidence annotations (Table 5). This result is consistent with prior literature that highlights LLMs as proficient medical evidence extractors (McInerney et al., 2023; Gero et al., 2023). While GPT4-seg exhibited the highest recall, a detailed analysis of its outputs uncovered an over-prediction of sentences as evidence for predicted ICD codes, leading to reduced precision (Table 5). As a result, LLM-codex surpasses existing methods in F1 score, striking a better balance between the precision and recall of its provided sentence-level evidences.

Table 5: Benchmarking sentence-level evidence extraction in the MDACE Profee evaluation dataset

Model	F1	Precision	Recall
EffectiveCAN	0.542	0.408	0.806
GPT4-seg	0.123	0.066	**0.944**
GPT4-doc	0.675	0.596	0.778
LLM-codex	**0.713**	**0.608**	0.861

5.6. Error case analysis

LLM-codex tends to overlook some ICD codes when the length of the sentence is long, as shown in row 2 and 3 in the Table A.4. Lengthier sentences typically have more ICD codes to assign, which can reduce GPT-4's accuracy. Additionally, LLM-codex tends to assign ICD V-codes excessively. V-codes are used to indicate non-diagnostic information, such as preventive services, routine check-ups, and administrative encounters. Since fee-for-service payment systems do not incentivize coding V-codes, they are rarely utilized (Torres et al., 2017; Guo et al., 2020). Hence, the ground truth may be under-labeled. The over-prediction of ICD V-codes using GPT-4 may further support this line of research in healthcare.

6. Discussion

In this paper, we present LLM-codex, a two-stage model that leverages LLMs for predicting document-level ICD codes and their corresponding sentence-level evidence. Our results show that LLM-codex significantly outperforms prior state-of-the-art models in predicting common document-level ICD codes, particularly when faced with limited training data. Additionally, LLM-codex demonstrates superior performance in predicting document-level rare ICD codes. When a single sentence-level evidence suffices to justify predicted ICD codes, LLM-codex notably achieves higher precision compared to existing approaches.

Our work has several limitations. First, We found that when comprehensively extracting all available sentence-level evidence for a predicted ICD code is essential, GPT-4 with segmentation outperforms LLM-codex (Table A.7). This is due to LLM-codex's current constraint of generating only one sentence-level evidence per predicted ICD code. To boost LLM-codex's recall, one could increase the number of evidence sentences returned by the *Verifier* model for each predicted ICD code. Although the impact on its precision remains unclear, exploring this modification could be part of future work. Moreover, incorporating explainability methods like the Masked Sampling Procedure (MSP) (Stremmel et al., 2022) into the *Verifier* model could further enhance LLM-codex's explainability by more comprehensively identifying sentence-level evidence for each predicted ICD code. Second, case studies show limited accuracy in long sentences, as GPT-4 can be misled by many medical keywords in a sentence. Finally, LLM-codex requires GPT-4 during inference. We estimate it costs \$0.50 per discharge summary to run LLM-Codex on the MIMIC-III dataset with a latency of about 10 seconds per document. To reduce cost and latency, future work could distill GPT-4 ICD coding performance into other large language models such as Llama2.

LLM-codex constitutes a substantial advancement in ICD code prediction and explainability by accurately predicting ICD codes, even with limited training data and for rare codes, while providing sentence-level explanations for coding decisions—capabilities not concurrently demonstrated by existing approaches. We believe this versatile method has the potential to extend to various classification tasks with limited annotations which require explanations for model decisions, such as medication abuse detection (Fleming et al., 2008; Kwon et al., 2023) and social determinants of health identification (Davidson and McGinn, 2019; Mitra et al., 2022), thus paving the way for promising future research.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.130. URL https://aclanthology.org/2022.emnlp-main.130.

Aitziber Atutxa, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, and Olatz Perez de Viñaspre. Interpretable deep learning to map diagnostic texts to icd-10 codes. *International journal of medical informatics*, 129:49–59, 2019. URL https://www.sciencedirect.com/science/article/abs/pii/S1386505618310670?via%3Dihub.

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.260. URL https://aclanthology.org/2022.naacl-main.260.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. MDACE: MIMIC documents annotated with code evidence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7534–7550, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.416. URL https://aclanthology.org/2023.acl-long.416.

Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1078. URL https://aclanthology.org/P18-1078.

Karina W. Davidson and Thomas McGinn. Screening for Social Determinants of Health: The Known and Unknown. *JAMA*, 322(11):1037–1038, 09 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.10915. URL https://doi.org/10.1001/jama.2019.10915.

Hang Dong, V'ictor Su'arez-Paniagua, William Whiteley, and Honghan Wu. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, page 103728, 2020. URL https://api.semanticscholar.org/CorpusID:225103119.

Hang Dong, V'ictor Su'arez-Paniagua, Huayu Zhang, Minhong Wang, Emma Whitfield, and Honghan Wu. Rare disease identification from clinical notes with ontologies and weak supervision. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2294–2298, 2021. URL https://api.semanticscholar.org/CorpusID:233739818.

Matúš Falis, Hang Dong, Alexandra Birch, and Beatrice Alex. Horses to zebras: Ontology-guided data augmentation and synthesis for ICD-9 coding. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 389–401, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.39. URL https://aclanthology.org/2022.bionlp-1.39.

Michael Francis Fleming, James Davis, and Steven D. Passik. Reported lifetime aberrant drug-taking behaviors are predictive of current substance use and mental health problems in primary care patients. *Pain medicine*, 9 8:1098–106, 2008. URL https://api.semanticscholar.org/CorpusID:22904673.

Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. Self-verification improves few-shot clinical information extraction. *ArXiv*, abs/2306.00024, 2023. URL https://api.semanticscholar.org/CorpusID:258999642.

Max Glockner, Ivan Habernal, and Iryna Gurevych. Why do you think that? exploring faithful sentence-level rationales without supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1080–1095, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.97. URL https://aclanthology.org/2020.findings-emnlp.97.

Yi Guo, Zhaoyi Chen, Ke Xu, Thomas J. George, Yonghui Wu, William R. Hogan, Elizabeth A. Shenkman, and Jiang Bian. International classification of diseases, tenth revision, clinical modification social determinants of health codes are poorly used in electronic health records. *Medicine*, 99, 2020. URL https://api.semanticscholar.org/CorpusID:229350312.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997. URL https://api.semanticscholar.org/CorpusID:1915014.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Seattle, WA, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.clinicalnlp-1.2. URL https://aclanthology.org/2022.clinicalnlp-1.2.

Adam Ivankay, Ivan Girardi, Chiara Marchiori, and Pascal Frossard. Fooling explanations in text classifiers. *ArXiv*, abs/2206.03178, 2022. URL https://api.semanticscholar.org/CorpusID:249431362.

Adam Ivankay, Mattia Rigotti, and Pascal Frossard. DARE: Towards robust text explanations in biomedical and healthcare applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11499–11533, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.644. URL https://aclanthology.org/2023.acl-long.644.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357.

Lavender Yao Jiang, Xujin Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard A. Riina, Ilya Laufer, Paawan Punjabi, Madeline Miceli, Nora C. Kim, Cordelia Orillac, Zane Schnurman, Christopher Livia, Hannah Weiss, David Kurland, Sean Neifert, Yosef Dastagirzada, Douglas Kondziolka, Alexander T M Cheung, Grace Yang, Mingzi Cao, Mona G. Flores, Anthony B Costa, Yindalon Aphinyanaphongs, Kyunghyun Cho, and Eric Karl Oermann. Health system-scale language models are all-purpose prediction engines. *Nature*, 619:357 – 362, 2023. URL https://api.semanticscholar.org/CorpusID:259112211.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.329. URL https://aclanthology.org/2022.findings-emnlp.329.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database.

Scientific Data, 3, 2016. URL https://api.semanticscholar.org/CorpusID:33285731.

Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine*, 65 2:155–66, 2015.

Byung-Hak Kim, Zhongfen Deng, Philip Yu, and Varun Ganapathi. Can current explainability help provide references in clinical notes to support humans annotate medical codes? In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 26–34, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.louhi-1.3. URL https://aclanthology.org/2022.louhi-1.3.

Sunjae Kwon, Xun Wang, Weisong Liu, Emily Druhl, Minhee L. Sung, Joel Reisman, Wenjun Li, Robert D. Kerns, William Becker, and Hongfeng Yu. Odd: A benchmark dataset for the nlp-based opioid related aberrant behavior detection. *ArXiv*, abs/2307.02591, 2023. URL https://api.semanticscholar.org/CorpusID:259360903.

Leah S. Larkey and W. Bruce Croft. Combining classifiers in text categorization. In *SIGIR*, page 289–297, 1996.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.clinicalnlp-1.17. URL https://aclanthology.org/2020.clinicalnlp-1.17.

Raymond Li, Ilya Valmianski, Li Deng, Xavier Amatriain, and Anitha Kannan. Oslat: Open set label attention transformer for medical entity retrieval and span extraction. In Antonio Parziale, Monica Agrawal, Shalmali Joshi, Irene Y. Chen, Shengpu Tang, Luis Oala, and Adarsh Subbaswamy, editors, *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pages 373–390. PMLR, 28 Nov 2022. URL https://proceedings.mlr.press/v193/li22a.html.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *ArXiv*, abs/2307.03172, 2023. URL https://api.semanticscholar.org/CorpusID:259360665.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.481. URL https://aclanthology.org/2021.emnlp-main.481.

Justin Lovelace, Nathan C. Hurley, Adrian D. Haimovich, and Bobak J. Mortazavi. Dynamically extracting outcome-specific problem lists from clinical notes with guided multi-headed attention. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 245–270. PMLR, 07–08 Aug 2020. URL https://proceedings.mlr.press/v126/lovelace20a.html.

Denis Jered McInerney, Geoffrey S. Young, J.-W. van de Meent, and Byron C. Wallace. Chill: Zero-shot custom interpretable feature extraction from clinical notes with large language models. *ArXiv*, abs/2302.12343, 2023. URL https://api.semanticscholar.org/CorpusID:257205986.

Simon Meoni, Eric De la Clergerie, and Theo Ryffel. Large language models as instructors: A study on multilingual clinical entity extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bionlp-1.15. URL https://aclanthology.org/2023.bionlp-1.15.

George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. ICDBigBird: A contextual embedding model for ICD code classification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages

330–336, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.32. URL https://aclanthology.org/2022.bionlp-1.32.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. A discrete hard EM approach for weakly supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2851–2864, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1284. URL https://aclanthology.org/D19-1284.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.201. URL https://aclanthology.org/2022.naacl-main.201.

Avijit Mitra, Richeek Pradhan, Rachel D. Melamed, Kun Chen, David C. Hoaglin, Katherine Louise Tucker, Joel Reisman, Zhichao Yang, Weisong Liu, Jack Tsai, and Hongfeng Yu. Associations between natural language processing–enriched social determinants of health and suicide death among us veterans. *JAMA Network Open*, 6, 2022. URL https://api.semanticscholar.org/CorpusID:254564500.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL https://aclanthology.org/N18-1100.

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Ramesh Kashyap, Stefan Winkler, Shao-Syuan Huang, Jie-Jyun Liu, and Chih-Jen Lin. Mimic-iv-icd: A new benchmark for extreme multilabel classification. *ArXiv*, abs/2304.13998, 2023a. URL https://api.semanticscholar.org/CorpusID:258352403.

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Ramesh Kashyap, and Stefan Winkler. A two-stage decoder for efficient ICD coding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4658–4665, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.285. URL https://aclanthology.org/2023.findings-acl.285.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

Sonia Pavan, Kathrin Rommel, María Elena Mateo Marquina, Sophie Höhn, Valérie Lanneau, and Ana Rath. Clinical practice guidelines for rare diseases: The orphanet database. *PLoS ONE*, 12, 2017.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

Cathy Shyr, Yan Hu, P. A. Harris, and Hua Xu. Identifying and extracting rare disease phenotypes with large language models. *ArXiv*, abs/2306.12656, 2023. URL https://api.semanticscholar.org/CorpusID:259224453.

Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. Perturbing inputs for fragile interpretations in deep natural language processing. In *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2021. URL https://api.semanticscholar.org/CorpusID:236976089.

Joel Stremmel, Brian L. Hill, Jeffrey Hertzberg, Jaime Murillo, Llewelyn Allotey, and Eran Halperin. Extend and explain: Interpreting very long language models. In Antonio Parziale, Monica Agrawal, Shalmali Joshi, Irene Y. Chen, Shengpu Tang, Luis Oala, and Adarsh Subbaswamy, editors, *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pages 218–258. PMLR, 28 Nov 2022. URL https://proceedings.mlr.press/v193/stremmel22a.html.

Niall Taylor, Yi Zhang, Dan W. Joyce, Alejo J. Nevado-Holgado, and Andrey Kormilitzin. Clinical prompt learning with frozen language models. *ArXiv*, abs/2205.05535, 2022.

Jacqueline M. Torres, John Lawlor, J. D. Colvin, Marion R. Sills, Jessica L. Bettenhausen, Amber Davidson, Gretchen J. Cutler, Matt Hall, and Laura M. Gottlieb. Icd social codes: An underutilized resource for tracking social needs. *Medical Care*, 55:810–816, 2017. URL https://api.semanticscholar.org/CorpusID:13590045.

Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 17 5:514–8, 2010. URL https://api.semanticscholar.org/CorpusID:20264071.

Stéphanie Nguengang Wakap, Deborah M. Lambert, Annie Olry, Charlotte Rodwell, Charlotte Gueydan, Valérie Lanneau, Daniel Murphy, Yann le Cam, and Ana Rath. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. *European Journal of Human Genetics*, 28:165 – 173, 2019.

Tao Wang, Linhai Zhang, Chenchen Ye, Junxi Liu, and Deyu Zhou. A novel framework based on medical concept driven attention for explainable medical code prediction via external knowledge. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1407–1416, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.110. URL https://aclanthology.org/2022.findings-acl.110.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. *Journal of biomedical informatics*, 77:34–49, 2018. URL https://api.semanticscholar.org/CorpusID:3632923.

Qiang Wei, Amy Franklin, Trevor A. Cohen, and Hua Xu. Clinical text annotation - what factors are associated with the cost of time? *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2018:1552–1560, 2018. URL https://api.semanticscholar.org/CorpusID:73482002.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. 2022. URL https://api.semanticscholar.org/CorpusID:258840837.

Martin J. Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, D. Rubin, and Matthew P. Lungren. Preparing medical imaging data for machine learning. *Radiology*, page 192224, 2020. URL https://api.semanticscholar.org/CorpusID:211160137.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hongfeng Yu. Multi-label few-shot icd coding as autoregressive generation with prompt. In *AAAI Conference on Artificial Intelligence*, 2022a. URL https://api.semanticscholar.org/CorpusID:254018044.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1767–1781, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.127. URL https://aclanthology.org/2022.findings-emnlp.127.

Zonghai Yao, Jack Tsai, Weisong Liu, David Levy, Emily Druhl, Joel Reisman, and Hongfeng Yu. Automated identification of eviction status from electronic health record notes. *Journal of the American Medical Informatics Association : JAMIA*, 2022. URL https://api.semanticscholar.org/CorpusID:254275327.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.

91. URL https://aclanthology.org/2022.acl-short.91.

Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York, April 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-1033.

Shurui Zhang, Bozheng Zhang, Fuxin Zhang, Bo Sang, and Wanchun Yang. Automatic ICD coding exploiting discourse structure and reconciled code embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2883–2891, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.254.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231979430.

Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang. Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 29:1208 – 1216, 2022. URL https://api.semanticscholar.org/CorpusID:247677346.

Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin Bretonnel Cohen. Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics*, 8 5:358–75, 2007. URL https://api.semanticscholar.org/CorpusID:5689080.

7. Appendix

7.1. Implementation Details

For the experiments in this study, the LLM we employ is GPT4-8k version 0314 (OpenAI, 2023). It is accessed securely through the Azure OpenAI API under the responsible use requirement[2]. We set the sampling temperature to 0.1 and truncate the EHRs to satisfy the 8k token constraint. Additionally, we define and evaluate the number of candidate codes N_c as 50; in theory, N_c could vary depending on the specific application. The LSTM architecture of our Verifier is the same as MSMN. Detailed hyperparameters are reported in Table A.5.

7.2. Empirical results on the number of segments to split

In Table 4, we showed that breaking down the input patient record into multiple equal-size sentence segments and then aggregating results across these segments increased the F1 of GPT-4. To find the best number of segments *segn*, we tuned the *segn* as hyperparameter and observed the best F1 when $segn = 10$ in Table A.6.

7.3. Benchmarking comprehensive evidence extraction

In Table 5, an ICD code in an EHR is considered positively predicted by a method if it predicts at least one expert-annotated sentence-level evidence from the MDACE Profee dataset corresponding to that ICD code. However, for some applications, it may be beneficial to assess the differences between methods that capture more than one sentence-level evidence for a given ICD code in an EHR. Consequently, we introduce an evaluation of comprehensive sentence-level evidence extraction, where each expert-annotated sentence-level evidence for an ICD code in an EHR is treated as an individual data point, allowing a method to predict multiple positives for that ICD code in the EHR.

We observe that while LLM-codex maintains superior precision compared to existing methods, it exhibits the lowest recall (Table A.7). This is due to the fact that the average number of gold evidence labels per ICD code in the MDACE Profee dataset is 3, while LLM-codex outputs at most one sentence-level evidence for each ICD code (providing exactly

2. https://physionet.org/news/post/415

one sentence-level evidence for all ICD codes whose predictions exceed a threshold optimized for the ICD code based on the F1 score of a validation dataset). Notably, we find that GPT4-seg achieves the highest recall, consistent with Table 5, but has low precision. Thus, the method demonstrating the optimal balance between precision and recall, and achieving the highest F1 score, is GPT4-doc, which outperforms EffectiveCAN in terms of F1, precision, and recall.

7.4. ICD-10 accuracy evaluation

We also tested coding accuracy on ICD-10 codes. We followed Nguyen et al. (2023a) and filtered the dataset to include instances with at least one of the top 50 most frequent ICD-10 codes. We limited the number of training instances to 1000 and named this dataset MIMIC-IV *few-shot*. The result is shown in Table A.1.

7.5. Disease-specific ablation study of LLM-codex

We investigate whether LLM-codex's performance varies across individual ICD codes and which of its components are critical for its performance. In order to do so, we first locate mentions of ICD-9 codes with an NER tool MedCat (Kavuluru et al., 2015). We then evaluate ICD coding accuracy on codes that were not explicitly mentioned in the documents.

We observe that on anemia prediction, LLM-codex with stage 1 and 2 achieves an F1 score of 0.567, outperforming MSMN which only scores 0.252 F1 (Figure A.2). Among the documents that had the hypertension code assigned, $\sim$ 8% of the EHRs were missing mentions of hypertension. In comparison, $\sim$ 55% of EHRs were missing mentions of anemia, thereby making the task of predicting anemia harder as it would require inference without it being explicitly mentioned. LLM-codex performs on par with MSMN in predicting hypertension. Furthermore, on the task of predicting anemia, LLM-codex achieves an AUPRC of, significantly outperforming MSMN's AUPRC of 0.208.

7.6. Prompt template

Example 1 *As a proficient clinical coding professionals, it is your responsibility to assign ICD 9 codes given the CLINICAL NOTE from the CANDIDATE LIST provided below.*

—

CLINICAL NOTE (or partial):
[text note]

—

Here is a CANDIDATE LIST of 50 ICD 9 codes and their associated descriptions to assign:
[candidates]

—

For each disease/procedure based on the context in CLINICAL NOTE, you must generate a list of strings containing the ICD 9 codes you assigned.

Example 2 *As a proficient clinical coding professional, it is your responsibility to extract evidence when assigning ICD code. Given the list of ICD 9 CANDIDATE codes (diseases/procedures) to assign, you need to verify each code by extracting associated evidence sentence from CLINICAL NOTE. You could inference based on basic medical commonsense, such as prescription of metformin is evidence to type 2 diabetes.*

—

ICD 9 CANDIDATE codes and descriptions: [diseases].

—

Here is the CLINICAL NOTE split by sentence, each sentence starts with an index number surrounded by parentheses: [text note]

—

When assigning ICD code, you should:
1. Carefully assign ICD code to each sentence as evidence even ICD code is already assigned in the previous sentence;
2. If multiple ICD code found in one sentence, label them all and separate them by semicolon;
3. Do not assign ICD code if it is negated or ruled out in the CLINICAL NOTE, for example you should not assign "287.5" if "No leukemia or thrombocytopenia";
4. Include ICD code only, not the associated English description.

Table A.1: Results on the MIMIC-IV *few-shot* benchmark on ICD-10 codes

Model	ROCAUC		F1	
	MACRO	MICRO	MACRO	MICRO
MSMN	**0.840** ±0.015	0.883 ±0.010	**0.508** ±0.017	0.577 ±0.015
LLM-codex	0.837 ±0.005	**0.906** ±0.007	0.497 ±0.009	**0.604** ±0.074

Table A.2: Results on the MIMIC-III *common* benchmark

Model	ROCAUC		F1	
	MACRO	MICRO	MACRO	MICRO
Blackbox CAML	0.884 -	0.916 -	0.576 -	0.633 -
Blackbox MSMN	0.928 ±0.004	**0.947** ±0.003	**0.683** ±0.007	**0.725** ±0.008
EffectiveCAN	0.920 -	0.945 -	0.668 -	0.717 -
LLM-codex	**0.929** ±0.003	0.948 ±0.002	0.674 ±0.006	0.715 ±0.005
LLM-codex /wo silver label	0.555 ±0.004	0.762 ±0.003	0.218 ±0.007	0.435 ±0.004

Table A.3: Results on the MIMIC-III benchmark with varying number of training instances.

Train size (shot)	Model	F1	Precision	Recall
8066 (all)	LLM-codex	0.715	0.704	0.726
	MSMN	0.725	0.713	0.738
1000 (14)	LLM-codex	0.611	0.587	0.638
	MSMN	0.561	0.545	0.581
500 (10)	LLM-codex	0.498	0.475	0.523
	MSMN	0.413	0.394	0.434

Table A.4: Examples of qualitative evaluations on LLM-Codex

Sentence	ICD code	Status
CT abdomen showed colitis.	556.8: Other ulcerative colitis	correct
The patient is a 46 year old female with a history of hypertension, OSA, and depression who was transferred from [**Hospital1 **] after presenting to the ED there with 4 days of nausea, vomiting, diarrhea, and worsening jaundice.	401.9: Essential Hypertension	correct
The patient is a 46 year old female with a history of hypertension, OSA, and depression who was transferred from [**Hospital1 **] after presenting to the ED there with 4 days of nausea, vomiting, diarrhea, and worsening jaundice.	782.4: Jaundice, unspecified, not of newborn	missed
Patient is a 56 y/o M s/p Whipple resection + SMV reconstruction [**5-14**] for pancreatic ca with SMV thrombosis ([**Doctor Last Name **] and [**Doctor Last Name **]), with post-operative course marked by delayed gastric emptying, requiring NGT reinsertion on POD 7 until POD 11.	536.3: Gastroparesis	missed
3 packs a day	V15.82: Personal history of tobacco use	overpredicted
Post-operative course was characterized by fever	V45.81: Post-surgical aorto-coronary bypass status	overpredicted

Table A.5: Hyperparameters used to train *Verifier* model

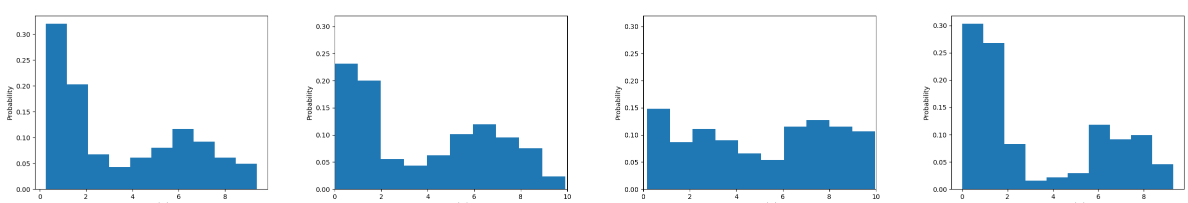
Parameters	Value
Emb. dim.	100
Emb. dropout	0.2
LSTM Layer	1
LSTM hidden dim.	512
LSTM output dim.	512
Rep. dropout	0.2
learning rate	5e-4
Batch size	1 doc
Weight decay	0.02

Table A.6: Best *segn* for GPT-4 model on MIMIC-III *few-shot* ICD code prediction task

Model	F1	Precision	Recall
GPT4 (segn=50)	0.346 ±0.009	0.225 ±0.008	**0.749** ±0.011
GPT4 (segn=25)	0.511 ±0.010	0.389 ±0.011	0.744 ±0.011
GPT4 (segn=10)	**0.582** ±0.010	0.482 ±0.011	0.73 ±0.011
GPT4 (segn=1)	0.484 ±0.015	**0.5** ±0.009	0.471 ±0.010

Table A.7: Benchmarking sentence-level *comprehensive* evidence extraction in the MDACE Profee evaluation dataset

Model	F1	Precision	Recall
EffectiveCAN	0.480	0.393	0.616
GPT4-seg	0.103	0.055	**0.860**
GPT4-doc	**0.550**	0.455	0.698
LLM-codex	0.453	**0.608**	0.360



(a) Human (b) LLM-codex (c) GPT4-seg (d) GPT4-doc

Figure A.1: Bin plot on the location of evidences, where x-axis is the sentence position from the start of document and y-axis is the occurrence density of each bin.

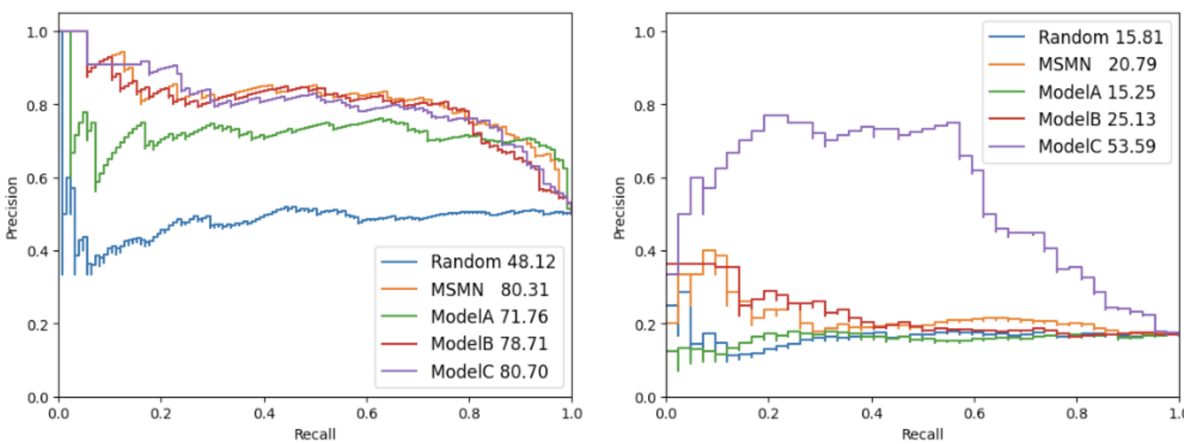


Figure A.2: The precision-recall curve and PRAUC for two example diseases. Left: a) hypertension with a limited amount of missing mentions in the medical note; Right: b) anemia with many missing mentions. ModelA is the LLM Stage 1a with the *Verifier* model, ModelB is the LLM Stage 1 with the *Verifier* model and ModelC is LLM-codex with the LLM Stages 1 and 2 along with the *Verifier* model.