EasyChair Preprint

№ 15731

Can Language Models Reason about ICD Codes to Guide the Generation of Clinical Notes?

Ivan Makohon, Jian Wu, Bintao Feng and Yaohang Li

January 18, 2025

Can Language Models Reason about ICD Codes to Guide the Generation of Clinical Notes?

Ivan Makohon[1[0000-0002-3627-7242]], Jian Wu[1[0000-0003-0173-4463]], Bintao Feng, and Yaohang Li[1[0000-0002-7892-5295]]

[1] Old Dominion University, Norfolk VA 23529, USA

Abstract. In the past decade a surge in the amount of electronic health record (EHR) data in the United States, attributed to a favorable policy environment created by the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 and the 21st Century Cures Act of 2016. Clinical notes for patients' assessments, diagnoses, and treatments are captured in these EHRs in free-form text by physicians, who spend a considerable amount of time entering them. Manually writing clinical notes may take considerable amount of time, increasing the patient's waiting time and could possibly delay diagnoses. Large language models (LLMs), such as GPT-3 possess the ability to generate news articles that closely resemble human-written ones. We investigate the usage of Chain-of-Thought (CoT) prompt engineering to improve the LLM's response in clinical note generation. In our prompts, we incorporate International Classification of Diseases (ICD) codes and basic patient information along with similar clinical case examples to investigate how LLMs can effectively formulate clinical notes. We tested our CoT prompt technique on six clinical cases from the CodiEsp test dataset using GPT-4 as our LLM and our results show that it outperformed the standard zero-shot prompt.

Keywords: Large language models, generative AI, chain-of-thought (CoT), natural language processing, information retrieval, clinical note generation, International Classification of Diseases (ICD) codes.

1 Introduction

In the past decade, there has been a surge in the amount of electronic health record (EHR) data in the United States. In 2008, only 42% of office-based physicians had access to an EHR. This figure has now risen to 88% as reported in 2021 [1]. This increase can be attributed to a favorable policy environment created by the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 [2] and the 21st Century Cures Act of 2016 [3].

Clinical notes for patients' assessments, diagnoses, and treatments are captured in these EHRs in free-form text by physicians, who spend a considerable amount of time entering them into computers. These notes offer valuable insights based on real-time observed data, which have shown to enhance the predictive capabilities of medical decision-making models [4]. Despite the rich information contained in these notes, it

is likely some details are excluded from publicly available LLMs due to restrictions on access to their content, a consequence of the Health Insurance Portability and Accountability Act (HIPAA) of 1996 [5]. HIPAA plays a crucial role in safeguarding the privacy and security of patients' protected health information (PHI) in the context of clinical notes.

Despite the rapid growth of medical advancements, the quality of healthcare has unfortunately fallen behind [6-8]. One significant contributing factor to this decline is physician burnout. Physicians experience emotional exhaustion, demotivation, and detachment from their patients caused by the demanding and stressful nature of their work. A primary culprit for this burnout is the inconvenient and inefficient structure of EHRs, which requires excessive data entry and clinical note-taking [4,6-8].

The medical scribe industry has emerged to handle the burdensome documentation tasks behind the scenes [9], but relying on non-professional scribes poses challenges, because they often lack the necessary medical expertise. To address physician burnout challenge, we focus our attention to large language models (LLMs) given that a remarkable progress has been made in recent years with some observations suggesting that they exhibit more powerful reasoning abilities as the model size increases [10-11].

Our paper makes the following research contributions:

1. We evaluate GPT-4's performance in generating patient current history of present illness (HPI) based on a task instruction, using diagnosis codes and relevant patient information as input.
2. We explore and apply CoT prompting, using clinical cases as examples to guide GPT-4 in generating clinical notes.

2 Related Works

The rapid advancements in LLMs have greatly enhanced their ability to comprehend patterns and relationships between words and phrases more effectively by developing a general understanding of grammar, syntax, and semantic relationships to generate text, bringing their output closer to human-level quality in areas of news compositions, story generation, and code generation [12]. LLMs like GPT-3 [10] and GPT-4 [11] have demonstrated impressive performance on downstream NLP tasks, even in zero-shot and few-shot settings. With its substantial capacity, it possesses the ability to generate news articles that closely resemble human-written ones, making it difficult to distinguish between the two [10]. This poses a particular challenge in detecting LLM-generated text, which is crucial for ensuring responsible AI governance [12]. GPT-4 is said to adhere more closely to guardrails, ensuring a higher level of responsible text generation.

Prompt engineering (or In-Context Prompting) [13-14] emerged as a recent field focused on crafting and refining prompts to effectively harness techniques aimed at interacting with LLM to guide its behavior towards specific goals, without making changes to the model weights. Since the recent releases of LLMs, Google researchers recently revolutionized a prompting strategy in solving word problems across five different LLMs [15]. Several prompt engineering techniques [16-17] have emerged and

significantly improves the performance of LLMs on many natural language generation tasks. Recent studies, such as Chain-of-Thought (CoT) [14-15], Tree-of-Thought (ToT) [18-19], and Graph-of-Thought (GoT) [20-21] have shown to improve the reasoning and accuracy performance of LLMs by providing rationales for a given word or phrase [14-15, 18-21]. Although self-verification [22] and self-consistency [23] have enhanced performance in CoT prompting, recent prompting techniques such as ToT [18-19] and GoT [20-21] have shown improvement, though their effectiveness is still being assessed. CoT [24] has demonstrated that LLMs are capable of reasoning through multiple-choice questions on medical board exams. For our purposes, can it reason about ICD codes along with some patient information to generate clinical notes?

Previous endeavors have demonstrated that employing an attention mechanism in a multi-label classification task can effectively yield ICD codes from clinical notes [25] and shows that numerous prior research endeavors have revolved around classifying ICD codes using clinical notes as their primary input data. Our work in this paper is to reverse this process by generating comprehensive clinical notes, guided by provided ICD codes and supplemented basic patient information using instructional prompting techniques. In a recent study [26], LLMs were investigated using zero-shot prompting to predict ICD-10 codes. ICD-10 codes were provided in their prompt: "Predict these ICD-10 codes to the best of your ability" without any patient information or a clear instruction task to generate clinical notes. Based on their outputs, the LLMs outputs predicted just the ICD codes titles, not actual patient clinical notes.

Additionally, recent studies have explored the use of LLMs for generating clinical notes with the use of prompt engineering. These include leveraging LLMs to convert transcribed interactions into structured notes through structured prompting and integration of supplementary data for improved quality [27], developing a specialized medical LLM to understand and summarize medical conversations using zero-shot prompt for note generation [28], and providing rapid access to medical information via a chatbot that utilizes a predefined system prompt to perform contextual searches and Retrieval-Augmented Generation (RAG) techniques [29]. Some of the challenges in clinical note generation through use of LLMs are captured in this study [30], which highlights the feasibility of training efficient open-source LLMs for clinical note generation, with opportunities for further exploration in domain adaptation, data selection, and reinforcement learning from human feedback (RLHF). RLHF helps align LLMs with human preferences and can be applied in two ways: outcome-supervised, which focuses on improving the overall quality of the text, and process-supervised, which provides more detailed guidance on specific text components, such as reasoning steps, as seen in approaches like InstructGPT [16].

We conduct experiments on the closed-source GPT-4 using semantic searches and the CoT prompting technique to query similar clinical cases based on the given ICD codes or text references. To our knowledge, we are the first to perform experiments of this kind using diagnosis codes (ICD codes) as input along with basic patient information to generate clinical notes using LLMs and CoT prompting instructions. We seek to answer our research question: Can LLMs reason about ICD codes to guide the generation of clinical notes using instruction prompting?

3 Methodology

This paper explores a method for guiding the generation of clinical notes by using an LLM (GPT-4) while providing a task instruction, ICD codes and patient information utilizing CoT instruction prompting as rationale prompts with examples of clinical notes diagnosis with similar ICD codes.

3.1 Semantic Search & Clinical Cases

CodiEsp, introduced during the CodiEsp track for CLEF eHealth 2020, is recognized as a gold-standard annotated data source [31]. The dataset is comprised of 1000 clinical cases, where the clinical notes are translated from Spanish to English. It encompasses both ICD-10 CM and PCS codes, distributed across three randomly sampled datasets: the Training set contains 500 clinical cases. The Development (validation) and the Test set each contains 250 clinical cases. The text-reference column consists of text used during the annotated process using the Brat visualization tool [32]. Hereafter, we will refer to text-reference column as the Text Reference.

We combine CodiEsp's training and validation datasets (750) for our semantic search embedding query, while reserving the test dataset (250) for selecting six clinical cases samples for evaluating ground-truth against generative text. We converted the texts in the combined dataset of 750 clinical cases into numerical vector representations with OpenAI's text embedding model (text-embedding-3-small). Our objective is to leverage the embedding-driven retrieval to tap into the rich semantic features present in other clinical cases. This is achieved through the use of query ICD codes or text references, which facilitates the provision of clinical case examples for use as "thoughts" in our CoT prompt. As we will demonstrate, these semantic searches provide an efficient approach to identifying examples resembling the examples in the prompts. The embedded query (ICD code or text reference) is used to pinpoint the most relevant clinical cases by assessing their proximity within the embedding space, utilizing document similarity to rank and present the top-n most suitable clinical cases. For each query, the cosine similarity is used to identify the top-n most similar clinical cases. We randomly selected six clinical cases with less than 1000 words from the test dataset that contain 1 or more ICD codes or text references. The breakdown of the ICD codes, text references and word count for each clinical case is shown in Table I.

Table I. Clinical Case Samples (counts).

Clinical Case	CodiEsp ArticleID	ICD Code	Text Reference	Clinical Note
A	S0213-12852003000600002-1	2	2	634
B	S1130-05582017000100031-1	1	1	764
C	S1130-01082008001000008-1	4	5	855
D	S1130-01082009000500011-1	9	9	745
E	S1130-01082008000100009-1	10	11	767
F	S1130-01082006001000017-1	9	9	633

3.2 Prompt Format

Our standard prompt, referred to as the baseline, is formatted as task instructions to generate the HPI clinical notes based on the given ICD codes with zero-shot prompt. Our CoT semantic search (CoT prompting) is formatted as instructions to guide the output of language model by controlling its generated text. In the CoT instruction prompt, each experiment contains the ground-truth clinical case's ICD codes along with basic patient information. In addition, the top-10 similar clinical cases are provided by semantic search query (based on the ground-truth ICD codes or text references), which uses contextualized word embeddings and the cosine similarity function to find related clinical cases, are provided as prompts. These inputs act as rationales, enabling the LLM to learn and generate the intended clinical notes based on the provided ICD codes.

Our CoT prompting takes inputs, such as:

- Task instruction.
- ICD codes for the diagnosis and/or procedure.
- Examples of similar clinical cases using the semantic search (ICD codes or text references) query.
- Basic patient information (age and gender).

3.3 Metrics

Cosine distance is the complement of cosine similarity, which measures the angular difference between two vector representations in a multi-dimensional space. It is a mathematical function that quantifies the degree of dissimilarity between two vectors based on their orientation rather than their magnitude. The cosine distance formula is defined as:

$$Cosine\ Distance = 1 - \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

where, $A \cdot B$ is the dot product of the sentence vectors A and B, $\|A\|$ and $\|B\|$ are the magnitudes of the vectors, and the result gives a measure of the angular distance between the vectors.

Transformer-based models, like Bidirectional Encoder Representations from Transformers (BERT) [33], capture both syntactic and semantic relationships between words by generating contextualized word embeddings. To assess the similarity between machine-generated and ground-truth documents, we use BERT. Both documents are processed through the BERT model (bert-large-cased) to obtain embeddings, which are then used to calculate sentence similarity:

- Using the special "classification" [CLS] token of each sentence. The [CLS] token in BERT serves as a means to gather a holistic representation of the input sequence. The output of [CLS] is inferred by all other words in this sentence. This implies that the [CLS] contains all information in other words, which makes [CLS] a representation for sentence-level classification.
- Calculating the MEAN of the sentence embeddings provides a way quantify the overall cosine distance between the sentences based on semantic meaning.

3.4 Experiment

For our experiment, we use OpenAI's closed-source GPT-4 (gpt-4) model as the foundational LLM for all experiments, with the following parameters: *seed* (123), *temperature* (0), *top_p* (0.000001), *frequency_penalty* (0), and *presence_penalty* (0). We establish a baseline for our results using standard zero-shot prompt and compare it against the results from our CoT prompts, which utilize a semantic search query based on the provided ICD codes or text references from ground-truth clinical case samples. Additionally, basic patient information from the ground-truth data is provided as supplementary prompts. Our semantic search query introduces an extra prompt, which includes the top-10 most similar clinical cases based on the given ICD codes or text references.

For each clinical case sample, we collect results from 100 API calls to GPT-4, with each call treated as an independent interaction. This ensures there is no memory or history from previous interactions, making each response independent. The clinical case's top-10 relatedness scores from the semantic search query are presented in Table II. These scores are calculated using cosine distance, which evaluates spatial proximity to identify the top-10 most similar clinical cases based on the provided ICD codes or text references.

Table II. Semantic Search Query (Top-10 Relatedness Scores).

Clinical Case	ICD Code Relatedness	Text Reference Relatedness
A	0.762, 0.754, 0.729, 0.724, 0.718, 0.715, 0.708, 0.695, 0.687, 0.683	0.563, 0.483, 0.478, 0.469, 0.462, 0.458, 0.433, 0.431, 0.429, 0.417
B	0.720, 0.717, 0.711, 0.701, 0.677, 0.672, 0.653, 0.644, 0.635, 0.630	0.469, 0.434, 0.411, 0.387, 0.380, 0.361, 0.359, 0.338, 0.336, 0.334
C	0.812, 0.780, 0.775, 0.768, 0.768, 0.760, 0.759, 0.754, 0.754, 0.751	0.601, 0.553, 0.545, 0.522, 0.521, 0.520, 0.520, 0.517, 0.512, 0.511
D	0.803, 0.802, 0.798, 0.796, 0.787, 0.783, 0.783, 0.782, 0.782, 0.782	0.630, 0.613, 0.606, 0.568, 0.565, 0.551, 0.547, 0.537, 0.524, 0.522
F	0.846, 0.807, 0.805, 0.795, 0.786, 0.774, 0.770, 0.769, 0.767, 0.757	0.637, 0.632, 0.623, 0.613, 0.610, 0.609, 0.601, 0.598, 0.595, 0.594
G	0.797, 0.775, 0.764, 0.763, 0.759, 0.752, 0.744, 0.741, 0.739, 0.739	0.654, 0.646, 0.645, 0.629, 0.627, 0.625, 0.624, 0.617, 0.615, 0.610

4 Results and Discussions

We evaluate our prompting technique using cosine distance as the primary metric, comparing generated text to ground truth. The results are visualized with a Kernel Density Estimation (KDE) plot, which includes Bootstrap Confidence Intervals (BCIs) for both sentence-level [CLS] and Mean scores. This comparison contrasts out CoT prompting against the baseline zero-shot prompt. The distribution analysis of CoT prompting, which incorporates semantic search queries (such as ICD codes and text

references) for similar clinical cases, consistently shows that our method enhances the language model's ability to capture the underlying patterns and reasoning of the ICD codes and text references, as well as basic patient information, compared to the baseline zero-shot prompt.

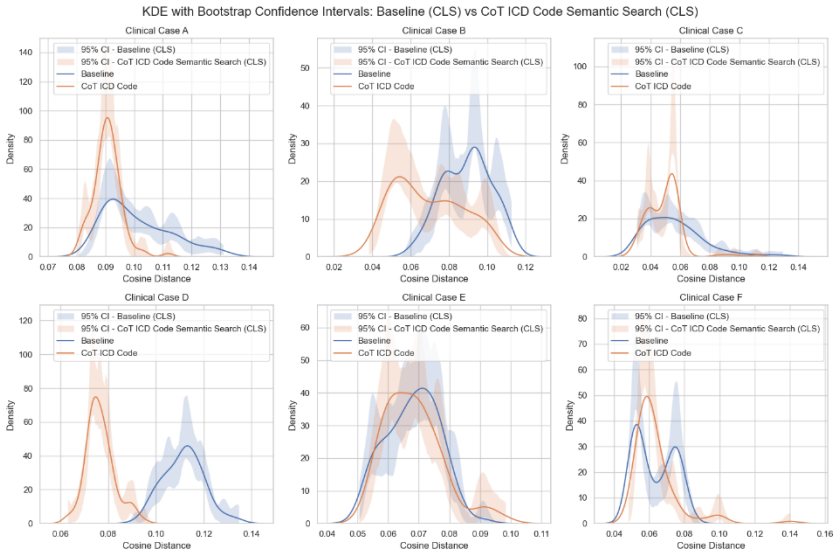

Fig. 1. Illustration of the six clinical cases using KDE with BCIs to compare the Baseline and CoT ICD code semantic search, based on cosine distance score of sentence-level [CLS].

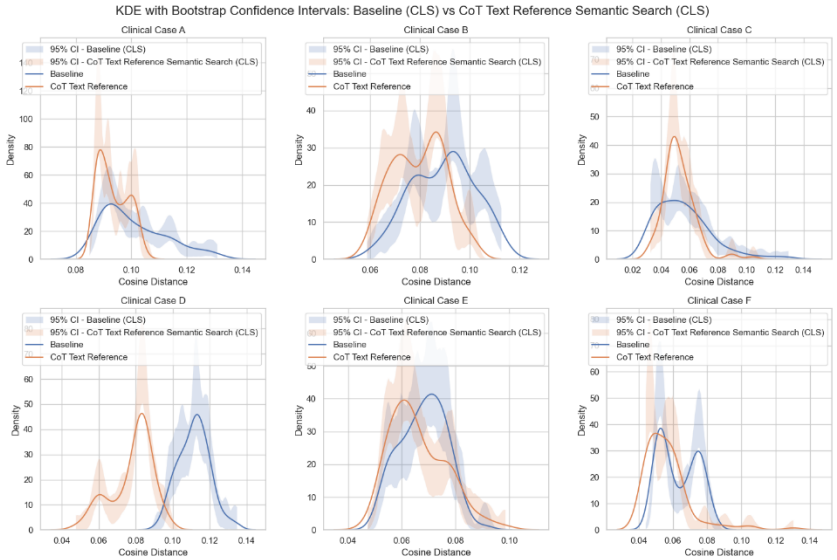

Fig. 2. Illustration of the six clinical cases using KDE with BCIs to compare the Baseline and CoT text reference semantic search, based on cosine distance score of sentence-level [CLS].

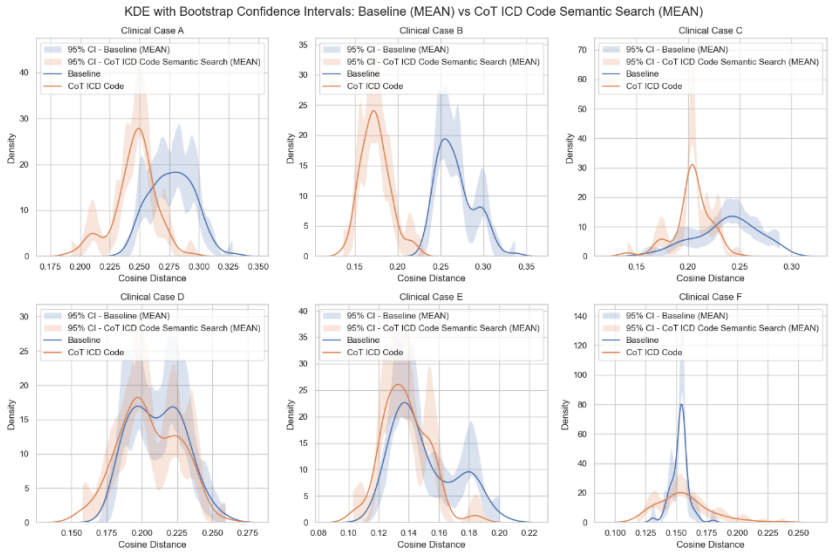

Fig. 3. Illustration of the six clinical cases using KDE with BCIs to compare the Baseline and CoT ICD code semantic search, based on cosine distance score of sentence-level MEAN.

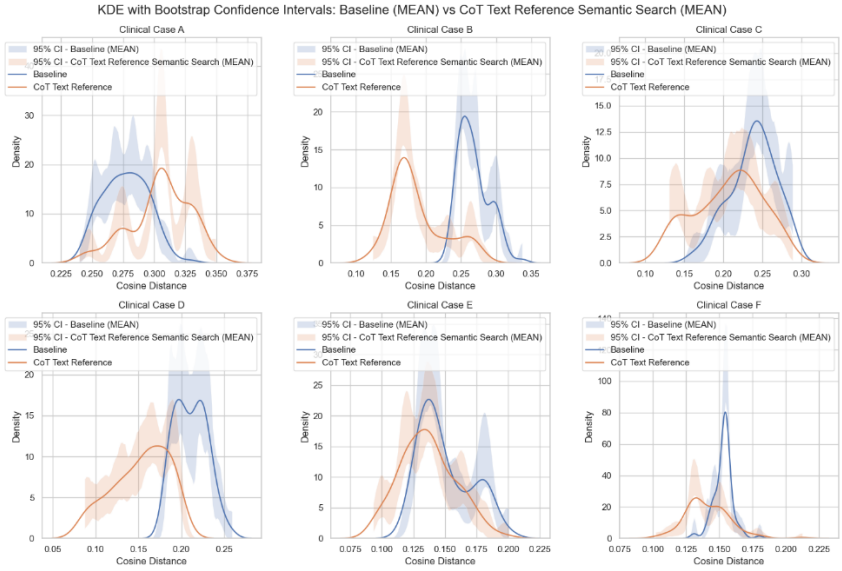

Fig. 4. Illustration of the six clinical cases using KDE with BCIs to compare the baseline and CoT text reference semantic search, based on cosine distance score of sentence-level MEAN.

The KDE with BCI plots (Figures 1-4) reveal a leftward shift in the peaks for the CoT semantic search prompting technique, indicating that its distribution has lower values compared to the baseline prompts. This shift highlights notable differences in semantic alignment with ground-truth clinical cases. The inclusion of BCIs provides statistical validation, enhancing the robustness and interpretability of these findings. However, as shown in Figure 4, we observe that in clinical case A, the baseline prompt outperforms the CoT text reference prompt. This may be due to the presence of two text references (pain, toothache). In contrast, in Figure 3, the two ICD codes (K08.89, R52) clinical case A perform better than the baseline prompt. The word "pain" for the text reference could be too general or nonspecific to accurately capture the clinical details needed for generating the HPI or guiding the model's output.

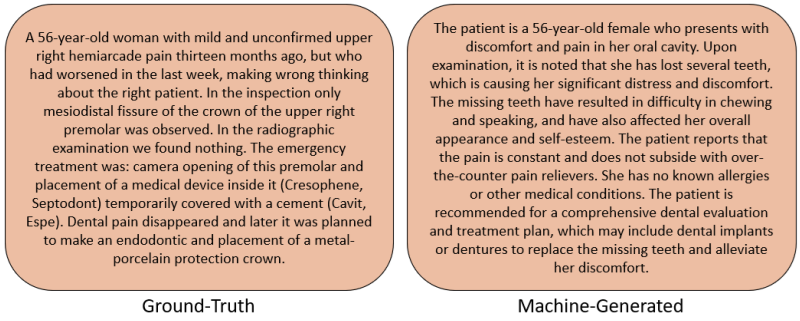

Fig. 5. A sample from Clinical Case A showing ground-truth and the machine-generated text.

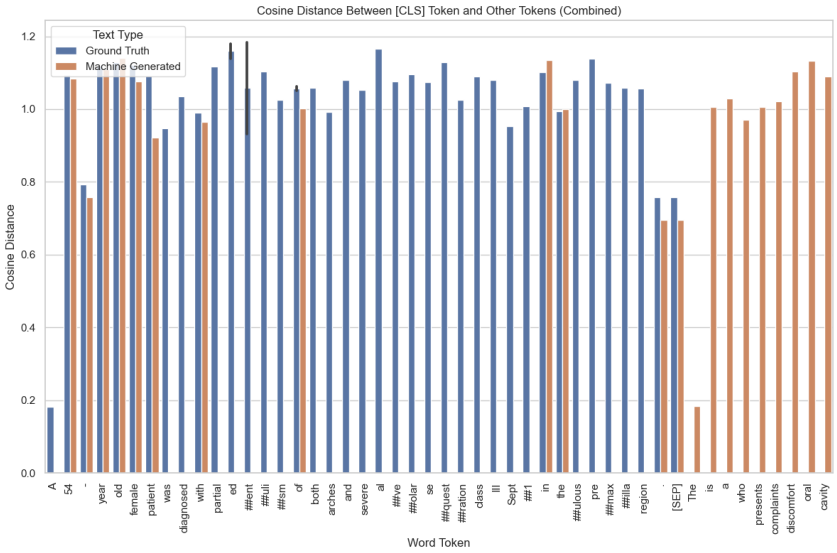

Fig. 6. Illustration of a sample from Clinical Case A, showing the cosine distance scores of the [CLS] tokens for both ground-truth and machine-generated text.

Figure 6 shows that some words are present in the ground-truth but missing from the machine-generated text. It also reveals that ICD-10 codes K08.89 (pain) and R52 (toothache) are linked to terms like "oral," "cavity," and "discomfort," which appear in the generated text, all related to oral health. Both texts include "A 54-year-old female," confirming alignment in basic patient information. However, the cosine distance between the [CLS] tokens is slightly over 1.0, likely due to contextual differences in the surrounding text. As seen in Figure 4, the ground-truth [CLS] embedding captures a more specific medical context, while the generated text is broader, which could explain the higher cosine distance.

Overall, by using visual comparisons, such as overlaying KDE plots for all clinical cases and prompting techniques, we were able to assess the extent of the observed shifts. These plots, along with the cosine distance measurements, confirm that the differences in semantic similarity performance between the techniques are both statistically significant and practically meaningful. For instance, clinical cases A and B exhibit a significant leftward shift in the peaks (Figures 1 and 3) for the ICD code semantic search, indicating that their distributions have lower values compared with the baseline. The ICD code prompting technique appears to outperform the Text Reference, likely due to higher relatedness scores for the clinical case examples (Table II). From observation, these visual insights combined with the precision offered by the BCIs show that both CoT prompting technique results distributions differ from the baseline standard prompt, which suggest that this technique can guide the generation of clinical notes through instruction prompt using similar clinical cases.

5 Conclusion and Future Work

Our study constructs the HPI clinical notes using CoT prompting with ICD codes, clinical case examples, and basic patient information. Experiments were conducted across various clinical cases (clinical cases with 1 ICD code, 2 ICD codes, and several cases with multiple ICD codes), comparing results obtained from a baseline zero-shot prompt and two CoT prompting templates. Through cosine distance analysis, we compared the generated text with ground-truth text, addressing whether LLMs can effectively reason about ICD codes to produce clinical notes using CoT prompting.

Our analysis concludes that the GPT-4 LLM is capable of reasoning about ICD codes using our CoT semantic search prompting techniques over the baseline zero-shot prompt to produce clinical notes. Also, comparing an EHR to a single ground-truth may not be effective, as different doctors write EHRs differently. Thus, human evaluation should be considered to ensure alignment with automatic metrics.

5.1 Future Work

To enhance the prompting techniques, we propose some follow up studies not only in areas of using other instruction prompting techniques, but in these areas as well to enrich LLMs reasoning:

1. CoT Prompting using Patient's Past Medical History

The Medical Information Mart for Intensive Care (MIMIC-III) [34] offers clinical data on 30-day ICU readmissions, allowing past medical history and admission notes to guide model predictions for future visits.

2. Fine-Tune an LLM to become more biased towards the Physician's output
   LLMs are prone to biases from training data, but they can be fine-tuned for individual physicians using personalized data. This adjustment, achieved through instruction prompting, allows the model to better meet specific needs. Physician notes can be extracted for this fine-tuning from the MIMIC-III dataset.
3. Use Retrieval Augmented Generation (RAG) in conjunction with CoT prompting
   RAG enables retrieval of relevant information, such as patient data from medical databases, to inform the instruction prompting process. Our semantic search embeddings identify the most relevant documents based on query similarity (e.g., patients with similar ICD codes). This helps guide LLM text generation, while RAG minimizes "hallucinations" by feeding relevant facts into the model, improving the accuracy and relevance of clinical note generation.

References

1. Office of the National Coordinator for Health Information Technology. "Office-based Physician Electronic Health Record Adoption," Health IT Quick-Stat #50. https://www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption.
2. Burde H. "Health Law the HITECH ACT - An Overview" in the Virtual Mentor, 13(3):172-175 (2011)
3. U.S. Food & Drug. 21st Century Cures Act. Available: https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act (2020)
4. Nelson, Deborah, D. "Copying and Pasting Patient Treatment Notes" in American Medical Association Journal of Ethics. March 2011, Volume 13, Number 3: 144-147 (2011).
5. Moore, W., & Frye, S.A. "Review of HIPAA, Part 1: History, Protected Health Information, and Privacy and Security Rules" in The Journal of Nuclear Medicine Technology, 47, 269 – 272 (2019).
6. Arndt, B.G., et al. "Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations" in The Annals of Family Medicine, 15, 419 – 426 (2017).
7. Gellert, G. A., Ramirez, R., & Webster, S. L. "The rise of the medical scribe industry: implications for the advancement of electronic health records" in Journal of the American Medical Association (JAMA), 313(13), 1315–1316 (2015).
8. Kroth, P.J., et al. "Association of Electronic Health Record Design and Use Factors with Clinician Stress and Burnout" in Journal of the American Medical Association (JAMA) Network Open, 2(8), e199609 (2019).
9. Liu, P.J. "Learning to Write Notes in Electronic Health Records" (2018).
10. Brown, T.B., et al. "Language Models are Few-Shot Learners" in the 34th Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada (2020).
11. OpenAI. GPT-4 Technical Report (2023).
12. Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D.F., & Chao, L.S. "A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions" in ArXiv, (2023).
13. Ye, S., Hwang, et. al. "In-Context Instruction Learning" in ArXiv, (2023).

14. Wei, J., et. al. “Chain of Thought Prompting Elicits Reasoning in Large Language Models” in ArXiv, (2022).
15. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., & Smola, A.J. “Multimodal Chain-of-Thought Reasoning in Language Models” in ArXiv, (2023).
16. Li, J., Tang, T., Zhao, W.X., Nie, J., & Wen, J. “Pre-Trained Language Models for Text Generation: A Survey”. ACM Computing Surveys, 56, 1 – 39, (2022).
17. Sahoo, P., et. al. “A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications” in ArXiv, (2024).
18. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., & Narasimhan, K. “Tree of Thoughts: Deliberate Problem Solving with Large Language Models” in the 37th Conference on Neural Information Processing Systems (NeurIPS) (2023).
19. Long, J. “Large Language Model Guided Tree-of-Thought” in ArXiv, (2023).
20. Besta, M., et. al. “Graph of Thoughts: Solving Elaborate Problems with Large Language Models” in the 38th AAAI Conference on Artificial Intelligence (AAAI-24) (2024).
21. 28. Yao, Y., Li, Z., & Zhao, H. “Beyond Chain-of-Thought, Effective Graph-of-Thought Reasoning in Large Language Models” in ArXiv, (2023).
22. Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, K., & Zhao, J. “Large Language Models are Better Reasoners with Self-Verification” in the Findings of the Association for Computational Linguistics: EMNLP 2023, pages 2550–2575, December 6-10 (2023).
23. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E.H., & Zhou, D. “Self-Consistency Improves Chain of Thought Reasoning in Language Models” (2022).
24. Li'evin, V., Hother, C.E., & Winther, O. “Can large language models reason about medical questions?” in Patterns. 17 July, Vol. 5 (2022).
25. Makohon, I., & Li, Y. “Multi-Label Classification of ICD-10 Coding & Clinical Notes Using MIMIC & CodiEsp” in the 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), 1-4 (2021).
26. Lee, Simon A., and Lindsey, Timothy. “Do Large Language Models understand Medical Codes?” in arXiv, June 6 (2024).
27. Biswas, A., & Talukdar, W. “Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation” in ArXiv, (2024).
28. Yuan, D., et. al. “A Continued Pretrained LLM Approach for Automatic Medical Note Generation” in the North American Chapter of the Association for Computational Linguistics, (2024).
29. Leong, H.Y., et. al. “A GEN AI Framework for Medical Note Generation”, (2024).
30. Wang, H., et al. “Adapting Open-Source Large Language Models for Cost-Effective, Expert-Level Clinical Note Generation with On-Policy Reinforcement Learning” in ArXiv, (2024).
31. Miranda-Escalada, A., et. al. “Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non-English Clinical Cases at CodiEsp Track of CLEF eHealth 2020” in the Conference and Labs of the Evaluation Forum (2020).
32. Stenetorp, P., et. al. “brat: a Web-based Tool for NLP-Assisted Text Annotation” in the Conference of the European Chapter of the Association for Computational Linguistics (2012).
33. Devlin, J., Chang, M., Lee, K., & Toutanova, K. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pages 4171–4186, Minneapolis, Minnesota (2019).
34. Johnson, A.E., et. al. “MIMIC-III, a freely accessible critical care database” in Scientific Data 3, 160035 (2016).