Using Large Language Models for Efficient Cancer Registry Coding in the Real Hospital Setting: A Feasibility Study

Chen-Kai Wang*
Department of Computer Science, National Yang Ming Chiao Tung University
Hsinchu, 300093, Taiwan, ROC
Advanced Technology Laboratory, Chunghwa Telecom Laboratories
Taoyuan, 326402, Taiwan, ROC
Email: dennisckwang@gmail.com

Cheng-Rong Ke*
Intelligent System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology
Kaohsiung, 80778, Taiwan, ROC
Email: F111154134@nkust.edu.tw

Ming-Siang Huang
Intelligent System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology
Kaohsiung, 80778, Taiwan, ROC
Email: elephant52381@gmail.com

Inn-Wen Chong
Division of Chest Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University
Kaohsiung, 80708, Taiwan, ROC
Department of Biological Science and Technology, National Yang Ming Chiao Tung University
Hsinchu, 30010, Taiwan, ROC
Email: chong@cc.kmu.edu.tw

Yi-Hsin Yang
National Institute of Cancer Research, National Health Research Institutes
Tainan, 70456, Taiwan, ROC
Email: yhyang@nhri.edu.tw

* C.-K. Wang and C.-R. Ke contributed equally to this work.

† Corresponding author

© 2024 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

Vincent S. Tseng†
Department of Computer Science, National Yang Ming Chiao Tung University
Hsinchu, 300093, Taiwan, ROC
Email: vtseng@cs.nycu.edu.tw

Hong-Jie Dai†
Intelligent System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology
Kaohsiung, 80778, Taiwan, ROC
Email: hjdai@nkust.edu.tw

The primary challenge in reporting cancer cases lies in the labor-intensive and time-consuming process of manually reviewing numerous reports. Current methods predominantly rely on rule-based approaches or custom-supervised learning models, which predict diagnostic codes based on a single pathology report per patient. Although these methods show promising evaluation results, their biased outcomes in controlled settings may hinder adaption to real-world reporting workflows. In this feasibility study, we focused on lung cancer as a test case and developed an agentic retrieval-augmented generation (RAG) system to evaluate the potential of publicly available large language models (LLMs) for cancer registry coding. Our findings demonstrate that: (1) directly applying publicly available LLMs without fine-tuning is feasible for cancer registry coding; and (2) prompt engineering can significantly enhance the capability of pre-trained LLMs in cancer registry coding. The off-the-shelf LLM, combined with our proposed system architecture and basic prompts, achieved a macro-averaged F-score of 0.637 when evaluated on testing data consisting of patients' medical reports spanning 1.5 years since their first visit. By employing chain of thought (CoT) reasoning and our proposed coding item grouping, the system outperformed the baseline by 0.187 in terms of the macro-averaged F-score. These findings demonstrate the great potential of leveraging LLMs with prompt engineering for cancer registry coding. Our system could offer cancer registrars a promising reference tool to enhance their daily workflow, improving efficiency and accuracy in cancer case reporting.

Keywords: Natural Language Processing; Large Language Models; Electronic Health Record; Cancer registry; Patient Journey.

1. Introduction

Lung cancer stands as the foremost cause of cancer-related deaths among individuals aged 50 years and older, surpassing breast, colorectal, and prostate cancers combined in 2020, as reported by the Global Cancer Observatory, an initiative of the International Agency for Research on Cancer (Ferlay et al., 2020). In the United States, it is projected that 611,720 people will succumb to cancer of all types in 2024, equating to approximately 1,680 deaths per day (Siegel et al., 2024). Similarly, lung cancer has persistently held the top position as Taiwan's leading cause of cancer-specific mortality over the years. The survival rates for patients with lung cancer remain persistently low, often due to late-stage diagnosis that precludes complete surgical resection, thereby reducing long-term survival prospects.

The Taiwan Cancer Registry (TCR), established in 1979 by the Taiwan Society of Cancer Registry, aims to comprehensively measure cancer incidence, morbidity, survival, and mortality among individuals with cancer in Taiwan (Chiang et al., 2015). However, the current method of reporting cancer cases involves labor-intensive and time-consuming manual review of extensive reports, including pathology and radiology reports. Dai et al. (2024) conducted a study at a hospital in southern Taiwan, finding that it takes approximately 30 minutes to process a single case in the

reporting processing. A significant challenge contributing to the time-intensive nature of the process is the large volume and diverse nature of reports associated with each patient. Registrars are required to review and understand a wide array of medical reports, such as pathology reports, radiology reports, and discharge summaries. These reports often cover a span of approximately 1.5 years per patient. One proposed solution to address this challenge involves leveraging artificial intelligence (AI) techniques to automatically parse and extract information from cancer pathology reports. However, these reports are commonly presented in unstructured formats, posing difficulties for machine interpretation due to varying writing styles among different hospitals. Current methodologies predominantly rely on specialized rule-based systems (Coden et al., 2009), machine learning models (Alawad et al., 2020; Dubey et al., 2019; Yoon et al., 2019) or the hybrid of neural symbolic system (Dai, Yang, et al., 2021). Most of these presented works (Alawad et al., 2020; Dubey et al., 2019; Yoon et al., 2019) evaluated their approaches based solely on a single pathology report per patient. This approach may lead to biased results and could struggle to adapt to the real reporting process.

Recently, large language models (LLMs) have emerged as an effective method for extracting information from medical reports (Thirunavukarasu et al., 2023). Due to their large number of parameters and extensive pre-trained on diverse text corpora, LLMs have demonstrated impressive performance across numerous natural language processing (NLP) tasks, including zero-shot and few-shot scenarios (Brown et al., 2020; Nori et al., 2023). Although LLMs have achieved remarkable success in various applications, they still face significant limitations, particularly in domain-specific or knowledge-intensive tasks. These limitations include difficulties with processing long context lengths (Wang et al., 2024) and the potential for generating “hallucinations” when dealing with queries outside their training data or requiring up-to-date information (Zhang et al., 2023). On the other hand, retrieval augmented generation (RAG) is an innovative method for tailoring LLMs to tasks in specific domains (Lewis et al., 2020). The core idea behind RAG is to leverage a vast collection of documents to enhance the capabilities of generative models, thereby improving efficiency in handling complex tasks that require integrated knowledge (Zakka et al., 2024). Unlike traditional LLMs, RAG functions like a search engine by retrieving relevant text data from external knowledge bases through semantic similarity calculations in response to queries. By referencing external knowledge and segmenting large documents into smaller chunks, RAG effectively reduces the problem of generating factually incorrect content and improves the handling of long context data (Kandpal et al., 2023).

In an effort to streamline the data curation process over the various reports of a patient journey while upholding high standards of accuracy, we explore the feasibility of employing LLMs alongside agentic RAG to autonomously extract cancer registry coding items pertaining to lung cancer from various types of clinical reports detailing a patient’s medical journey. This methodology mirrors the responsibilities of a cancer registrar in a real setting, involving the analysis of unstructured reports to identify pertinent data elements essential for cancer registry purposes and their conversion into standardized codes.

Our contributions can be summarized as follows:

(1) We develop an agentic RAG system to facilitate the cancer registry coding process in a real hospital setting. Specifically, we assess the feasibility of directly applying openly available

LLM models without any fine-tuning, utilizing sophisticated crafted prompts through the prompt engineering process.

(2) We empirically show that off-the-shelf LLMs can achieve promising performance on certain cancer registry coding tasks based on the proposed system architecture and the compiled prompts. For example, Mistral-7B (Jiang et al., 2023) can achieve a macro-averaged F-score (F) of 0.637 when evaluated on the test data used in the previous study (Dai et al., 2024).

(3) The LLM, employing strategies such as chain-of-thought (CoT) (Wei et al., 2022) and the proposed coding item grouping, performs better by a large margin than those without these features. When evaluated on the test data, the enhanced strategy outperforms the baseline model without CoT by 0.187 in terms of macro-averaged F-score.

(4) The proposed system can provide a reference text to facilitate the interpretation of the generated outcomes. We conducted an analysis of the presented errors with a detailed discussion for future direction. Through the analysis, we believe that by further validating the generated output with the original reports to reduce the potential hallucinations observed in the presented study, the system could offer cancer registrars a promising reference tool to enhance their daily workflow.

2. Methods

To facilitate the coding process over the large and diverse reports associated with each cancer patient, we propose adapting the agentic RAG system. This system incorporates openly available LLM models along with sophisticatedly designed prompts through the prompt engineering process. In this section, we will first outline the dataset used and the target coding items. Then, we will provide an extensive overview of the proposed agentic RAG system. Subsequently, we will detail the design process and methods for our prompts. Finally, we will describe the evaluation metrics employed to assess the performance of our proposed system.

2.1. *Datasets*

In collaboration with a hospital in southern Taiwan, we collected cancer registry records of lung cancer patients linked with corresponding medical reports in our previous work (Dai et al., 2024). In the compiled dataset, we removed records unrelated to lung cancer based on primary site information, along with patients who had fewer than two reports or only one type of report. This resulted in a final dataset comprising 30 coding item records for 1,629 patients. The dataset was further divided into training and testing sets, comprising 1,287 and 342 patients, respectively. Each patient is associated with an average of 14.6 medical records. Despite Mandarin Chinese being Taiwan's official language, all medical reports were documented mainly in English or a mixture of Chinese and English. The dataset was used for the evaluation of the proposed agentic RAG system for automatic cancer registry coding. For this pilot study, we selected eight coding items to develop our LLM-based cancer registry coding assistant system. These items include pathological TNM classifications (TNM), histology types (H), behavior types (B), primary site (PS), laterality (L), and grades (G).

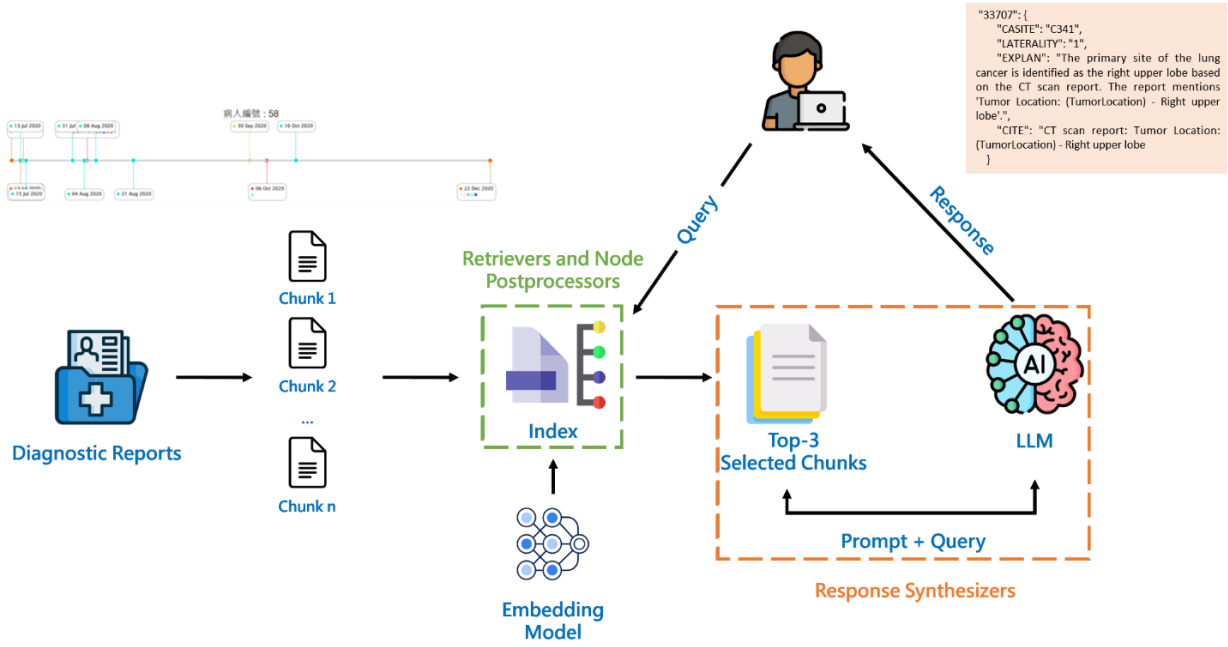

Fig. 1. Workflow of the proposed agentic RAG system.

2.2. *Proposed Agentic RAG System*

We applied RAG to process free-text medical reports collected over approximately 1.5 years to generate recommended cancer registry coding outcomes. Our system employs advanced embedding models to index and retrieve text chunks related to the specific coding prompt from medical reports using retrievers. These chunks are then filtered with post-processors to enhance accuracy before generating standardized cancer registry codes with an LLM. The system functions in two main stages: Top-*k* embedding-based retrieval and LLM-based code generation.

In the first stage, the same embedding model used for indexing the chunks of medical records is used to embed the given prompts for retrieving the most pertinent text chunks from medical reports for each patient. These chunks are refined using keyword-based post-processors to ensure they are among the top three most relevant for the coding task. In the second stage, the refined text chunks are combined with the prompt and query. The LLM then processes this integrated information, learning from patterns in provided examples, analyzing the input, and generating accurate and standardized cancer registry codes.

Figure 1 illustrates the system workflow of the proposed agentic RAG method. The detail workflow operates as follows: Initially, various medical reports for each patient are collected and formatted into JSON lines. These documents are then segmented into smaller, manageable text chunks. Each chunk undergoes processing through an embedding model, transforming it into a vector representation. When a query prompt for a coding task is received, it is similarly converted into a vector representation to facilitate a search within the vector database. This search identifies the most relevant text chunks, which are then combined with the query prompt to create a refined request. This refined request is sent to the LLM for processing, which subsequently provides a comprehensive response. For the underlying LLMs employed by the agents of the proposed RAG

system, we experimented with Mistral-7B and LLaMA3-8B (Touvron et al., 2023). Both Mistral and LlaMA3 are renowned for their balance between computational efficiency and performance across diverse NLP tasks. Therefore, for our implementation of the proposed RAG system, we selected Mistral-7B as the base model and compared its performance with LlaMA3-8B.

2.3. *Prompt Engineering for Cancer Registry Coding*

A prompt is a text-based and task-specific instruction given to a language model to guide its output without altering its parameters. The language model processes the prompt and generates a response based on the provided instructions and context (Marvin et al., 2023). Typically, a prompt may include instructions, input data, context, and an output indicator. According to the information provided, prompts can be categorized into four levels (Heston & Khun, 2023). Level four, known as CoT, breaks down the instruction into step-by-step solutions, offering language models a more structured way to handle the prompt for improved accuracy. Prompt engineering has emerged as a crucial technique for crafting effective prompts. It is an iterative process aimed at refining defined prompts to enhance the capabilities of pre-trained LLMs. In this subsection, we describe the crafted level four prompts through an iterative prompt engineering process.

First, we set the goal to design the initial prompts for the eight coding tasks. We precisely specified the definitions of the coding task along with the desired output formats. In our initial implementation, we used the long-form coding manual of TCR (revision of the 2018v.6) to include detailed explanations and coding guidelines for each coding item. The first and second rows of Table 1 show examples of the PS coding item.

Furthermore, to achieve a more automated and controllable process, we designed output format prompts to instruct the LLM on how to format its output. As shown in the third row of Table 1, we specified that the LLM should generate its response in JSON format to facilitate the extraction of the conclusions. The output JSON object contains three keys: “explain”, “cite”, and the names of the target coding items. The target coding item name key holds the final coding result suggested by the LLM. If the LLM cannot determine the result based on the given report, the values for this key is instructed to assign “NA”. The “explain” key holds the explanation provided by the LLM for the reason why the coding results are suggested. The “cite” key includes the relevant paragraphs from the documents referenced by the LLM to support the coding results. We developed a simple parser based on regular expressions to convert the decoded text response from an LLM into a JSON structured format. If the response for a report cannot be parsed, “NA” is assigned for that report.

During the refining phase, we evaluated the performance of each prompt on the coding tasks using the training set. Biomedical expert MS Huang (listed as the second author) carefully analyzed the models’ responses to identify any errors or areas where the response fell short. Based on the error analysis and the potential solutions observed, we adjusted the prompt content to get a more precise response. This process was repeated until satisfactory performance was achieved. We then evaluate the developed prompts on the test set for performance comparison.

During the iterative process, we observed that certain coding items are often considered together in the actual cancer registration process. Therefore, we treated these related items as a coding item set and integrated their instructions into a single prompt during the design phase. For example, PS and L are often addressed together. This integration helps streamline the process and ensures that

Table 1. Example of the level 3 structured prompt defined for the "primary site" coding item. The ellipsis indicates the placeholder for the prompt string for other coding items belonging to the same group.

Prompt Component	Example
Coding item set definition	Your task as an assistant is to identify and confirm **the primary site […]** of lung cancer. **The primary site refers to specific regions within the respiratory system.** **[…]**
Coding rules for an individual item	It is essential to use only the information provided in the document at hand, considering its date and the pertinent organs or tissues examined. Choose from the following standard codes: - **Primary site codes: C339: Trachea, C340: Main bronchus, C341: Upper lobe, lung, …** […]
Output format (including **examples**)	Your response must be a valid JSON object containing the following keys: **-'primary site': A string containing the code for the primary site.** **[…]** Ensure your response is limited to the provided options for **primary site […]**. For instance, if the pathologic diagnosis specifies '**Lung; upper lobe; left**', this indicates that **the primary site […]** are located in the '**upper lobe**' of the 'left' lung, according to the provided options your JSON response should be: { "explain": "[Insert your explanation here based on the document]", "cite": "[Insert the relevant passages extracted from the document used for your decision]", **"primary site": "C341"**, **[…]**

related items are coded consistently and accurately. Table 2 shows the pre-defined coding item set. The grouped coding items are instructed within the same prompt.

Another significant improvement during the iterative process to the above basic prompt was the introduction of CoT reasoning for coding items like TNM and G. This method involves decomposing the coding task into intermediate steps and solving each step before arriving at the final answer (Wei et al., 2022). For example, consider the coding item G. Initially, we provided detailed coding rules in the prompt, such as:

Table 2. The pre-defined related item groups.

Coding item group type	Coding item
Grouped	- Pathological TNM classification (TNM) - Primary site and Laterality (PS and L) - Histology and Behavior (H and B)
Isolated	Pathological grades (G)

> *…Exclude any data from metastatic sites or recurrent tumors. If an excisional biopsy was conducted at the primary site and subsequent tumor resection shows no residual tumor, use the pathological grade/differentiation from the excisional biopsy. For patients who underwent neoadjuvant treatment before surgery, record the grade/differentiation based on post-surgical tumor tissue pathology. …*

We revised these rules by breaking down the coding task of G into three steps resulting a level four prompt:

1. Identify relevant reports: First, we requested the LLM to identify pathology reports that include surgical procedures from all available medical reports using a list of predefined common surgical terms.
2. Define reference range: Next, we instructed the model to produce the coding result for G based solely on the pathology reports identified in the first step. Coding definition rules similar to the initial detailed definitions shown above were also applied in this step.
3. Point out other key points: Finally, we instruct the model to improve its accuracy by considering the dates of the reports and the specific organs or tissues examined, followed by applying the exact "coding rules" for G.

2.4. *Evaluation Metrics*

We evaluate the performance of the proposed agentic RAG system using the commonly used metrics for evaluating information extraction results: precision (P), recall (R), and F_1-measure (F). P and R are also known as positive predictive value and sensitivity, respectively. The F-score is the weighted harmonic mean of P and R. The formulae for the three metrics are defined as follows:

$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$

$$F = \frac{2\times P\times R}{P+R}$$

In these formulas, TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively, for each coding item. Specifically, if the model outputs "NA" for a coding item for a patient's entire report set, it is counted as one FN for that patient.

Table 3. Performance comparison of the proposed systems across eight coding items. The highest F-scores for each type are highlighted in bold.

	Mistral-7B			LLaMA3-8B			Neural-symbolic	MT-CNN	HAN
Coding Item	P	R	F	P	R	F	F	F	F
T	0.707	0.915	0.798	0.844	0.972	0.904	**0.905**	0.730	0.763
N	0.845	0.955	0.897	0.860	0.976	0.914	**0.928**	0.830	0.904
M	0.433	0.898	0.584	0.400	0.917	0.557	**0.930**	0.799	0.822
PS	0.877	0.914	0.895	0.894	0.987	**0.938**	0.884	0.750	0.710
L	0.911	0.917	0.914	0.926	0.987	**0.956**	0.948	0.910	0.951
H	0.724	0.710	0.717	0.721	0.964	0.825	**0.871**	0.700	0.760
B	0.942	0.855	0.897	1.000	0.977	0.988	0.934	**0.994**	**0.994**
G	0.815	0.975	0.888	0.883	0.970	0.925	0.932	0.797	**0.939**

3. Results

3.1. *Performance Comparison of the Proposed Agentic RAG System*

To illustrate the effectiveness of the proposed system, we compared it with the previously developed neural symbolic hybrid system (Dai et al., 2024). and two baseline models, as shown in Table 3. For the hierarchical attention network (HAN) model (Gao et al., 2018), we followed the binary relevance transformation method (Dai, Su, et al., 2021) to formulate the coding task for each coding item as a multiclass classification task, training the corresponding number of HAN-based classifiers. For the multi-task convolutional neural network (MT-CNN) model (Alawad et al., 2020), BioWordVec (Zhang et al., 2019) was used to represent tokens, and a single model was trained to generate all eight cancer registry items.

For the proposed RAG systems, LLaMA3-8B clearly outperformed Mistral-7B in almost all coding items under the same configuration and prompt design. LLaMA3-8B also outperformed MT-CNN and HAN in five and six coding items, respectively. Notably, LLaMA3-8B also performed comparably to the neural symbolic system developed in our previous work, achieving the best F-scores in coding items such as PS and L. These promising results demonstrate the feasibility of using LLM models without any fine-tuning for cancer registry coding tasks in the real hospital setting.

3.2. *Ablation Study Results on Different Prompt Engineering Techniques*

To further evaluate the effectiveness of the exectuted prompt engineering process for downstream task performance. We execute our ablation study on four cases: (1) full prompt: the complete prompt with all components shown in Tables 1 and 2 and CoT; (2) a level 3-G prompt: a prompt without CoT; (3) a level 3-I prompt: a prompt without CoT and all coding item groups shown in Table 2 are isolated; and (4) a level 2 prompt: a level 3-I prompt without adding the context. In our implementation, the context refers to the part of “Coding rules for an individual item”. The results are shown in Table 4.

The results from the level three prompt demonstrate the potential of LLMs in performing cancer registry coding tasks from medical reports. This finding is particularly inspiring as it highlights the broader potential of leveraging off-the-shelf LLMs for processing medical text without sophisticated

Table 4. The ablation study results on different prompt engineering techniques

Technique	Macro-P	Macro-R	Macro-F
Full prompt (level 4)	0.782	0.892	0.824
w/o CoT (level 3-G)	0.595	0.733	0.637
w/o CoT & Group (level 3-I)	0.571	0.699	0.609
w/o Coding Rules (level 2)	0.000	0.000	0.000

prompt engineering. Specifically, we observed that the performance for the coding items L, B, and PS is satisfactory, with F-scores over 0.85. However, the performance of the proposed RAG system with grouped prompts on the TN and G coding items is less satisfactory, with F-scores lower than 0.6. Additionally, using isolated prompts alone, the proposed system struggles with additional coding items including TNM and G, showing even lower F-scores (F-score <0.6). By comparing the results of the level three prompt with the full prompt, we found that the inclusion of CoT reasoning significantly boosts the macro-averaged F-score from 0.637 to 0.824. This highlights the effectiveness of the employed prompt engineering process. Additionally, the level two prompt failed to extract any coding items, demonstrating the lack of practical cancer registry coding knowledge in the current off-the-shelf LLMs.

4. Discussion

4.1. *Error Analysis*

Benefiting from the development of pre-trained LLMs, the proposed RAG system can rapidly support most of the cancer registry coding item extraction tasks without further fine-tuning steps. However, from our results, we also observe that the system may occasionally produce conclusions contrary to the facts, even when clear clues are present in the reference texts. These “hallucinations” indicate that the system’s performance has room for improvement. In this section, we outline common error profiles derived from the overall design and present corresponding examples along with potential solutions for future work.

Reference Data Flaws: A single patient may have several to dozens of reports at different times and for different examination items during their treatment period. Using all reports can avoid missing critical information but also introduces computational burdens and noise that may interfere with the decision of the coding results. Therefore, in the retrieval phase for evidential chunks, we only retrieve the top three chunks to narrow the inference space. However, this approach has a double-edged sword effect, which may lead to inappropriate reference chunk citations. Such errors arise when the provided chunks do not offer clear and appropriate clues, leading the model to either refrain from responding or generate hallucinations not mentioned in the original text. Based on our analysis of the presented system errors, it is evident that the current implementation sometimes suffers from the dilemma of similar information retrieved from the top-3 reference data. To address this issue, a post-retrieval process mechanism could be introduced to enhance the diversity among candidate chunks. Balancing data coverage would be helpful for this shortcoming.

Inconsistency with Facts: There are instances where the model produces outputs deviating from the retrieved facts, even when the medical reports already provide a clear basis for concluding the coding results. Despite the defined prompts guiding and restricting the model's behavior, situations that exceed these controls still occur. This type of hallucination, where the model lacks fidelity to the source facts, has also been noted in recent research on LLMs (Tonmoy et al., 2024). Both the initial one-shot prompting and the current self-consistency CoT (Wang et al.) approaches may not be robust enough to assist the model in recalibrating its responses. Future work could explore techniques like Re-Reading (RE2), which enhances understanding by processing questions twice to better focus on the input (Xu et al., 2024), and Self-Reflective Retrieval-Augmented Generation (Self-RAG), which improves both quality and factual accuracy through retrieval and self-reflection (Asai et al., 2023). These methods could help the model produce more consistently and progressively refined outputs.

Knowledge Boundary Limitations: In the process of diagnosing cancer, different examination methods may yield varying results. Summarizing multiple possibilities and ultimately providing a final answer is challenging for both professionals and support systems. For instance, when identifying cancer histology, conclusions derived from surgical pathology are generally more reliable than those obtained from specimens, gross examinations, or microscopic examinations. We noticed that the current applied LLMs are limited by inherent knowledge gaps and may lack the capability to accurately assess the strength of evidence across reports, leading to a higher likelihood of errors.

4.2. *Prompting Engineering for Cancer Registry Coding*

The extraction of target information from clinical texts using LLMs heavily depends on effective prompt design. Due to the multifunctional capabilities of pre-trained models, prompts can be crafted in various ways. This flexibility is particularly useful when considering the professional nature of the input texts and the need for post-processing the output data.

In this study, the aim was to extract specific cancer registry codes from medical reports. The prompt design included a detailed instruction section, coding definitions, and examples, with the output required in a specific JSON format. This comprehensive approach, although necessary for accuracy, resulted in longer prompts. Different studies adopt varying prompt strategies. For instance, Hyeon Seok's work (Choi et al., 2023), which involved extracting cancer features from breast ultrasound and surgery reports, utilized simpler prompts without strict format requirements, as the outputs underwent manual validation. This streamlined approach achieved an accuracy of 87.7%. On the other hand, Huang et al. (2024) study, similar to ours, used detailed prompts for extracting data from public cancer data repositories, requiring output in a JSON format. Their structured prompt design, supported by thorough data preprocessing, achieved an F_1-score of 88%.

These examples demonstrate that while detailed prompts can enhance accuracy, they must be balanced with the need for efficiency and simplicity. A well-designed prompt, aligned with clean data sources and logical objectives, can significantly improve system performance, showcasing the importance of thoughtful prompt construction in utilizing LLMs effectively.

5. Conclusion

Cancer registry tasks involve referencing numerous clinical imaging and diagnostic reports to abstract patient information according to the AJCC-defined codes. These tasks are typically performed by certified clinical personnel with specialized cancer knowledge. The development of cancer registry support systems has the potential to reduce clinical workload and improve healthcare quality. Unlike traditional machine learning models, LLMs can utilize knowledge-guided prompts to predict field codes, making them valuable tools for supporting clinical tasks. In this study, we utilized the Mistral-7B and LLaMA3-8B pre-trained models and designed prompts for eight cancer registry items, including PS, L, H, B, G, and TNM. We observed that providing context and coding rules in a single prompt led to weaker performance due to insufficient reference report extraction. Incorporating CoT prompts, which provide step-by-step guidance toward the final coding output, significantly improved system performance. Additionally, we found that without specific cancer registry rules, the model's outputs became inconsistent and unreliable.

Overall, our findings indicate that LLMs can achieve promising results in lung cancer registry coding tasks even without the need for fine-tuning. Specifically, LLMs demonstrate impressive performance and efficiently utilize auxiliary data for task completion without specific training examples. This underscores their potential as invaluable tools for automating and optimizing cancer data management processes.

Appendix A. Prompt for Grouped Primary Site and Laterality in the Proposed RAG System

Your task as an assistant is to identify and confirm the primary site and laterality of lung cancer. The primary site refers to specific regions within the respiratory system. The laterality refers to whether the cancer originates from a paired organ and is applicable only to primary tumors. It is essential to use only the information provided in the document at hand, considering its date and the pertinent organs or tissues examined. Choose from the following standard codes for lung cancer sites and laterality:

Primary site codes:
- C339: Trachea
- C340: Main bronchus
- C341: Upper lobe, lung
- C342: Middle lobe, lung
- C343: Lower lobe, lung
- C348: Overlapping lesion of lung
- C349: Lung NOS (Not Otherwise Specified)

Laterality codes:
- 1: Primary origin of the cancer is on the right side.
- 2: Primary origin of the cancer is on the left side.
- 3: Unilateral involvement only, but origin unclear whether from left or right side.
- 4: Bilateral involvement with unclear side of origin, and medical records describe a single primary.

Your response must be a valid JSON object containing the following keys:
-'primary site': A string containing the code for the primary site.
-'laterality': A string containing the code for laterality.

Ensure your response is limited to the provided options for primary site and laterality. For instance, if the tissue in the report is labeled as 'Lung; NOS' and the pathologic diagnosis specifies 'Lung; upper lobe; left', this indicates that the primary site and laterality are located in the 'upper lobe' of the 'left' lung, according to the provided options your JSON response should be:

```
{
"explain": "[Insert your explanation here based on the document]",
"cite": "[Insert the relevant passages extracted from the document used for your decision]",
"primary site": "C341",
"laterality": "2"
}
```

Appendix B. Implementation Details for the Proposed RAG System

For the proposed RAG system, we utilize the LlamaIndex (Liu, 2022) framework. The developed system is deployed on a machine equipped with PyTorch libraries and CUDO12.0 along with an Intel i7-13700 processor, 64GB of RAM, and an NVIDIA GeForce RTX 4090 24GB VRAM (video RAM) graphics card. We employ M3-Embedding (Chen et al., 2024) as our embedding model for encoding a patient's every medical report during the indexing stage. For the retrieval module, we set the number of top K candidate chunks to three. In our configuration settings, we set the temperature to 0 and the seed to 42.

It is worth noting that loading models for inference demands a substantial amount of GPU memory. A general rule of thumb is that every billion parameters require 3 GB of graphics double data rate (GDDR) 6 VRAM for the default precision of parameter values (Lin et al., 2024). Due to the limitations of our machine hardware specifications, we quantize the employed LLMs to fixed-point 4 (FP4) for inference, which recasts these model weights into lower precision data types. This method slightly reduces performance but significantly lowers the memory requirement to a quarter of the original.

Appendix C. Definition of 30 Lung Cancer Coding Items in the Dataset for This Study

Coding Type	Description
AJCC Edition	The version and chapters of the AJCC (American Joint Committee on Cancer) cancer staging manual used to determine the cancer stage of the case.
Behavior Code	The morphological code (M-code) in the pathological diagnosis. The 5th code in the M-code is the behavior code. The first four digits of M-code indicate the specific histological term. The fifth digit is the behavior code, which indicates whether a tumor is malignant, benign, in situ, or uncertain.
Clinical Other Staging Group	The classification standards of the selected "Other Staging Systems" (defined below) chosen for staging cancer cases.
Clinical Stage Descriptor	The prefix or suffix used in conjunction with clinical TNM fields. The prefix/suffix denotes special circumstances that may affect the staging and analysis of the data and is based on the clinical T, N, and M categories prior to treatment.

Date of First Microscopic Confirmation	The earliest date when the case's cancer was confirmed by microscopy.
Date of First Surgical Procedure	The earliest date of surgery for cancer performed at any medical institution.
Date of Initial Diagnosis	The earliest date the cancer was diagnosed by a physician.
Date of Surgical Diagnostic and Staging Procedure	The date of the surgical treatment performed for diagnosis or staging at any medical institution.
Diagnostic Confirmation	The most accurate basis of diagnosis at the reporting hospital or an external hospital for the case.
Grade Clinical	The grading/differentiation of the solid tumor before the first treatment. Grading/differentiation refers to the degree of similarity between the tumor and normal tissues. Well differentiated (Grade I) is most similar to normal tissue; undifferentiated (Grade IV) is most dissimilar from normal tissue.
Grade Pathological	The grading/differentiation of the solid tumor after surgery at the primary site. Grading/differentiation refers to the degree of similarity between the tumor and normal tissues. Well differentiated (Grade I) is most similar to normal tissue; undifferentiated (Grade IV) is most dissimilar from normal tissue.
Histology	The structure of the primary tumor cells under the microscope.
Laterality	The specification of whether the cancer originates from one side of a pair of organs or the body. It is a only applicable to the primary tumor site.
Lymph vessels or Vascular Invasion	The code is recorded based on the pathological report of the primary site to indicate the presence or absence of invasion into lymph vessels or blood vessels.
Nodes Examined	The total number of regional lymph nodes examined by a pathologist.
Nodes Positive	The total number of positive regional lymph nodes examined by a pathologist.
Other Staging System	The selection of alternative staging criteria if the AJCC Cancer Staging System is not utilized.
Pathologic M	The presence of distant metastases of the primary tumor.
Pathologic N	The regional lymph nodes involvement of the tumor. The item is encoded based on all clinical evaluations done prior to definitive surgery, plus all information through completion of definitive surgeries in the first course of treatment in the absence of disease progression or within 4 months of diagnosis, whichever is longer.

Pathologic Stage Descriptor	The prefix or suffix used in conjunction with pathologic TNM fields. The prefix/suffix denotes special circumstances that may affect the staging and analysis of the data and is based on the pathologic T, N, and M categories after completion of surgical treatment.
Pathologic T	The size of the primary tumor and its invasion into adjacent tissues. The item is encoded based on all clinical evaluations done prior to definitive surgery, plus all information through completion of definitive surgeries in the first course of treatment in the absence of disease progression or within 4 months of diagnosis, whichever is longer.
Perineural Invasion	The presence of neural invasion as noted in the pathological report of the primary site in the medical records.
Primary Site	The primary site of the cancer.
Scope of Regional Lymph Node Surgery	The extent of regional lymph nodes removed, sectioned, or aspirated during the primary site surgery or another separate surgery at the reporting hospital.
SSF 2 SSF 5 SSF 6 SSF 7	Cancer site-specific factors (SSF) related to prognosis and treatment decisions. SSF2: Visceral pleural Invasion (VPI)/elastic layer value set. SSF5: Sampling or dissection of mediastinal lymph nodes (N2 Nodes) value set. SSF6: EGFR (epidermal growth factor receptor) gene mutation value set. SSF7: ALK (Anaplastic lymphoma kinase) gene translocation value set.
Surgical Margins	The final status of the surgical margins after the primary tumor is removed.
Surgical Margins Date	The closest distance of tumor cells to the surgical margins in the pathological report after the primary tumor is removed.

References

Alawad, M., Gao, S., Qiu, J. X., Yoon, H. J., Blair Christian, J., Penberthy, L., Mumphrey, B., Wu, X.-C., Coyle, L., & Tourassi, G. (2020). Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. *Journal of the American Medical Informatics Association*, *27*(1), 89-98.

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Chiang, C.-J., You, S.-L., Chen, C.-J., Yang, Y.-W., Lo, W.-C., & Lai, M.-S. (2015). Quality assessment and improvement of nationwide cancer registration system in Taiwan: a review. *Japanese journal of clinical oncology*, *45*(3), 291-296.

Choi, H. S., Song, J. Y., Shin, K. H., Chang, J. H., & Jang, B.-S. (2023). Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiation Oncology Journal*, *41*(3), 209.

Coden, A., Savova, G., Sominsky, I., Tanenblatt, M., Masanz, J., Schuler, K., Cooper, J., Guan, W., & De Groen, P. C. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of biomedical informatics*, *42*(5), 937-949.

Dai, H.-J., Chen, C.-C., Mir, T. H., Wang, T.-Y., Wang, C.-K., Chang, Y.-C., Yu, S.-J., Shen, Y.-W., Huang, C.-J., & Tsai, C.-H. (2024). Integrating predictive coding and a user-centric interface for enhanced auditing and quality in cancer registry data. *Computational and Structural Biotechnology Journal*, *24*, 322-333.

Dai, H.-J., Su, C.-H., Lee, Y.-Q., Zhang, Y.-C., Wang, C.-K., Kuo, C.-J., & Wu, C.-S. (2021). Deep learning-based natural language processing for screening psychiatric patients. *Frontiers in psychiatry*, *11*, 533949.

Dai, H.-J., Yang, Y.-H., Wang, T.-H., Lin, Y.-J., Lu, P.-J., Wu, C.-Y., Chang, Y.-C., Lee, Y.-Q., Zhang, Y.-C., & Hsu, Y.-C. (2021). Cancer registry coding via hybrid neural symbolic systems in the cross-hospital setting. *IEEE Access*, *9*, 112081-112096.

Dubey, A. K., Hinkle, J., Christian, J. B., & Tourassi, G. (2019). Extraction of tumor site from cancer pathology reports using deep filters. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics,

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., & Bray, F. (2020). Global cancer observatory: cancer today. International Agency for Research on Cancer. *Lyon, France*.

Gao, S., Young, M. T., Qiu, J. X., Yoon, H.-J., Christian, J. B., Fearn, P. A., Tourassi, G. D., & Ramanthan, A. (2018). Hierarchical attention networks for information extraction from cancer pathology reports. *Journal of the American Medical Informatics Association*, *25*(3), 321-330.

Heston, T. F., & Khun, C. (2023). Prompt engineering in medical education. *International Medical Education*, *2*(3), 198-205.

Huang, J., Yang, D. M., Rong, R., Nezafati, K., Treager, C., Chi, Z., Wang, S., Cheng, X., Guo, Y., & Klesse, L. J. (2024). A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digital Medicine*, *7*(1), 106.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., & Saulnier, L. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. International Conference on Machine Learning,

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459-9474.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., & Han, S. (2024). AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. *Proceedings of Machine Learning and Systems*, *6*, 87-100.

Liu, J. (2022). *LlamaIndex*. https://github.com/jerryjliu/llama_index

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. International conference on data intelligence and cognitive informatics,

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. *CA: a cancer journal for clinicians*, *74*(1).

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, *29*(8), 1930-1940.

Tonmoy, S., Zaman, S., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., & Azhar, F. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, X., Salmani, M., Omidi, P., Ren, X., Rezagholizadeh, M., & Eshaghi, A. (2024). Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2402.02244*.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. Self-consistency improves chain of thought reasoning in language models. arXiv 2022. *arXiv preprint arXiv:2203.11171*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Xu, X., Tao, C., Shen, T., Xu, C., Xu, H., Long, G., & Lou, J.-g. (2024). Re-Reading Improves Reasoning in Large Language Models.

Yoon, H.-J., Gounley, J., Gao, S., Alawad, M., Ramanathan, A., & Tourassi, G. (2019). Model-based hyperparameter optimization of convolutional neural networks for information extraction from cancer pathology reports on HPC. 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI),

Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., Fong, R., Phillips, C., Alexander, K., & Ashley, E. (2024). Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, *1*(2), AIoa2300068.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, *6*(1), 52.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., & Chen, Y. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.