Fine-tuning large language models for effective nutrition support in residential aged care: a domain expertise approach

Authors:

Mohammad Alkhalaf [1,2], Jun Shen[1], Hui-Chen (Rita) Chang[3], Chao Deng[4] , Ping Yu[1]

[1] School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia
[2] School of Computer Science, Qassim University, Qassim, 51452, Saudi Arabia
[3] School of Nursing and Midwifery, Western Sydney University, Penrith NSW 2751, Australia
[4] School of Medical, Indigenous and Health Sciences, University of Wollongong, Wollongong, NSW 2522, Australia

Abstract

Purpose: Malnutrition is a serious health concern, particularly among the older people living in residential aged care facilities. An automated and efficient method is required to identify the individuals afflicted with malnutrition in this setting. The recent advancements in transformer-based large language models (LLMs) equipped with sophisticated context-aware embeddings, such as RoBERTa, have significantly improved machine learning performance, particularly in predictive modelling. Enhancing the embeddings of these models on domain-specific corpora, such as clinical notes, is essential for elevating their performance in clinical tasks. Therefore, our study introduces a novel approach that trains a foundational RoBERTa model on nursing progress notes to develop a RAC domain-specific LLM. The model is further fine-tuned on nursing progress notes to enhance malnutrition identification and prediction in residential aged care setting.

Methods: We develop our domain-specific model by training the RoBERTa LLM on 500,000 nursing progress notes from residential aged care electronic health records (EHRs). The model's embeddings were used for two downstream tasks: malnutrition note identification and malnutrition prediction. Its performance was compared against baseline RoBERTa and BioClinicalBERT. Furthermore, we truncated long sequence text to fit into RoBERTa's 512-token sequence length limitation, enabling our model to handle sequences up to1536 tokens.

Results: Utilizing 5-fold cross-validation for both tasks, our RAC domain-specific LLM demonstrated significantly better performance over other models. In malnutrition note identification, it achieved a slightly higher F1-score of 0.966 compared to other LLMs. In prediction, it achieved significantly higher F1-score of 0.655. We enhanced our model's predictive capability by integrating the risk factors extracted from each client's notes, creating a combined

data layer of structured risk factors and free-text notes. This integration improved the prediction performance, evidenced by an increased F1-score of 0.687.

Conclusion: Our findings suggest that further fine-tuning a large language model on a domain-specific clinical corpus can improve the foundational model's performance in clinical tasks. This specialized adaptation significantly improves our domain-specific model's performance in tasks such as malnutrition risk identification and malnutrition prediction, making it useful for identifying and predicting malnutrition among older people living in residential aged care or long-term care facilities.

Keywords:
Large language model, domain-specific fine-tuning, RoBERTa, prediction, nursing notes, unstructured EHR, malnutrition

1. Introduction

Malnutrition is a serious health problem with many negative health consequences for older people, such as a weakened immune system and impaired cognition [1]. It may also contribute to vulnerabilities of infections, anemia and other diseases [2-4]. Malnutrition has been identified as a key area for urgent review by the Australian government for residential aged care [5] with identification of poor nutrition and weight loss as important indicators measuring the quality of care in residential aged care facilities (RACF). Healthcare professionals are requested to regularly screen older adults for early detection of malnutrition [6, 7]. To date, the common malnutrition screening tools used at RACFs include Mini Nutritional Assessment (MNA) and Subjective Global Assessment (SGA). However, since using these tools are time-consuming, they were not consistently applied [7]. Predicting and addressing malnutrition can lead to better health outcomes and improved quality of life [8]. Thus, it is crucial to develop new methods to improve the efficiency and effectiveness in malnutrition detection. However, scarcity of reliable datasets detailing the nutritional intake, lack of domain-specific knowledge models, and the novelty of the transformer approach have been primary obstacles to the development of a malnutrition prediction model for older people.

1.1 Electronic health records

Electronic health records (EHRs) have been widely adopted in RACFs in Australia to document clients' diagnosis, health assessment, nursing care plans, personal preferences, activities of daily living and care received [9]. The datasets in these EHRs can be classified as structured data and unstructured data. The structured data include client demographics and diagnosis that are recorded in structured tables. The unstructured data include nursing care plans, assessment records and free-

text clinical notes [10]. Most information about clients in RACFs, including nutritional information, is recorded in unstructured progress notes in EHR. Since EHR data is captured real-time in the care service delivery process, models trained on EHR can be more readily applied to clinical practise [11]. This provides the opportunity for natural language processing (NLP) to extract insights from the unstructured data in EHR for aged care services.

1.2 Natural language processing

Recent advancements in artificial intelligence, more specifically NLP, have opened doors for extracting relevant information and automating clinical diagnoses and predictions using language models on patient EHR [12-14]. One of NLP's recent advancements is the word embedding technique, which is a way of representing text as multi-dimensional vectors. Models such as GloVe [15] and word2vec [16] apply such text representation and have achieved promising results in different fields [17, 18]. However, these models lack context awareness, a core competency in text analysis.

1.3 Large language models

The emergence of the encoder-based large language models (LLM) such as Bert [19] and RoBERTa [20] have brought in positive disruption to the field of NLP. They utilize contextualized embeddings that account for both the prior and subsequent contexts of a token, adjusting its weight vector accordingly. By analyzing relationships between all pairs of words, LLMs introduce context-awareness, addressing the weakness of previous models. LLMs also have the ability to transfer previously acquired knowledge, thus are more efficient and have achieved state-of-the-art (SOTA) performance in many general downstream tasks with minimal to no need to modify architecture [21]. They can be further fine-tuned on specific corpus for specific-domain tasks. This

has enabled them to be successfully fine-tuned for various complex applications. Previous studies have demonstrated that LLMs can be trained on medical corpus to achieve high reliability in medical diagnoses and predictions [22-24]. While LLMs have demonstrated their utility in extracting data from public health data sets, their practical application in specific clinical tasks within real clinical settings, using clinic data, remains limited [25, 26].

1.4 RoBERTa

RoBERTa is a robust encoder-based LLM that is further optimized from its predecessor BERT model for better performance on a variety of NLP tasks [20]. It has achieved SOTA performance after being trained with massive text data with increased parameters, larger batch size and learning rate. RoBERTa utilizes byte-level tokenization instead of word tokenization in BERT. In addition, it randomizes the masking place, which eliminates the chance for the model to memorize the training data. Previous studies found that encoder-based LLMs such as RoBERTa outperform or at least are as effective as decoder-based LLMs, e.g. ChatGPT, in classification task [27, 28]. RoBERTa's architecture is highly suitable for fine-tuning on domain specific data sets. It has smaller model size, requires less computational power and memory, and often provides faster inference times compared to larger models like Llama model [29] or ChatGPT [30]. These make it more feasible for deployment in various health systems and devices [31]. Therefore, we choose RoBERTa as the candidate model for our task of generating knowledge about nutrition care in RACFs over other models.

1.5 Objective

Since, to date, there are no models that have been reported to be specifically fine-tuned for classifying and predicting malnutrition in older people, this study aimed to conduct NLP on free-

text notes in the RAC EHR for two downstream tasks: (1) identifying malnutrition notes and (2) malnutrition prediction. We fine-tuned an encoder-based LLM, RoBERTa checkpoint, to produce a nutrition domain-specific LLM in Australian RAC setting. We evaluated the performance of our model in comparison with other baseline models, including BioClinicalBERT and RoBERTa. In addition, free text nursing notes within EHR often contain extensive and detailed documentation, however RoBERTa has a maximum sequence length limitation of only 512 tokens. To address this challenge, we developed a new method for processing long notes with length over the 512 token limit.

2. Methodology

2.1 Dataset

The dataset was obtained from 40 aged care facilities in the state New South Wales (NSW), Australia. Overall, 4,405 de-identified clients' data was included in this analysis. The data was extracted from 1,616,820 notes of dietitians and nursing care staff recorded between Jan 2019 and October 2020, with an average number of 366 notes for each client. The Human research ethics approval for this study was granted by the Human Research Ethics Committee, the University of Wollongong and the Illawarra Shoalhaven Local Health District (Year 2020).

Table 1:The proportion of malnourished clients in the studied population (n = 4,405)

	Well-nourished (n=3,204)	Malnourished (n = 1,201)
Age	Mean (SD) 85.2 (8.9)	Mean (SD) 85.1 (8.9)
Female	2,071 (74%)	726 (26 %)
Male	1,133 (70%)	475 (30 %)

2.2 Data cleaning

All notes were cleansed of noise, including removing white spaces, special symbols and unwanted characters that do not contribute to the meaning of the text.

2.3 Overview of the methodology

Figure 1 depicts our NLP pipeline. It consists of three pathways: Path 1, fine-tuning a domain specific LLM; Path 2, finetuning a malnutrition note identification model; Path 3, finetuning malnutrition prediction model.

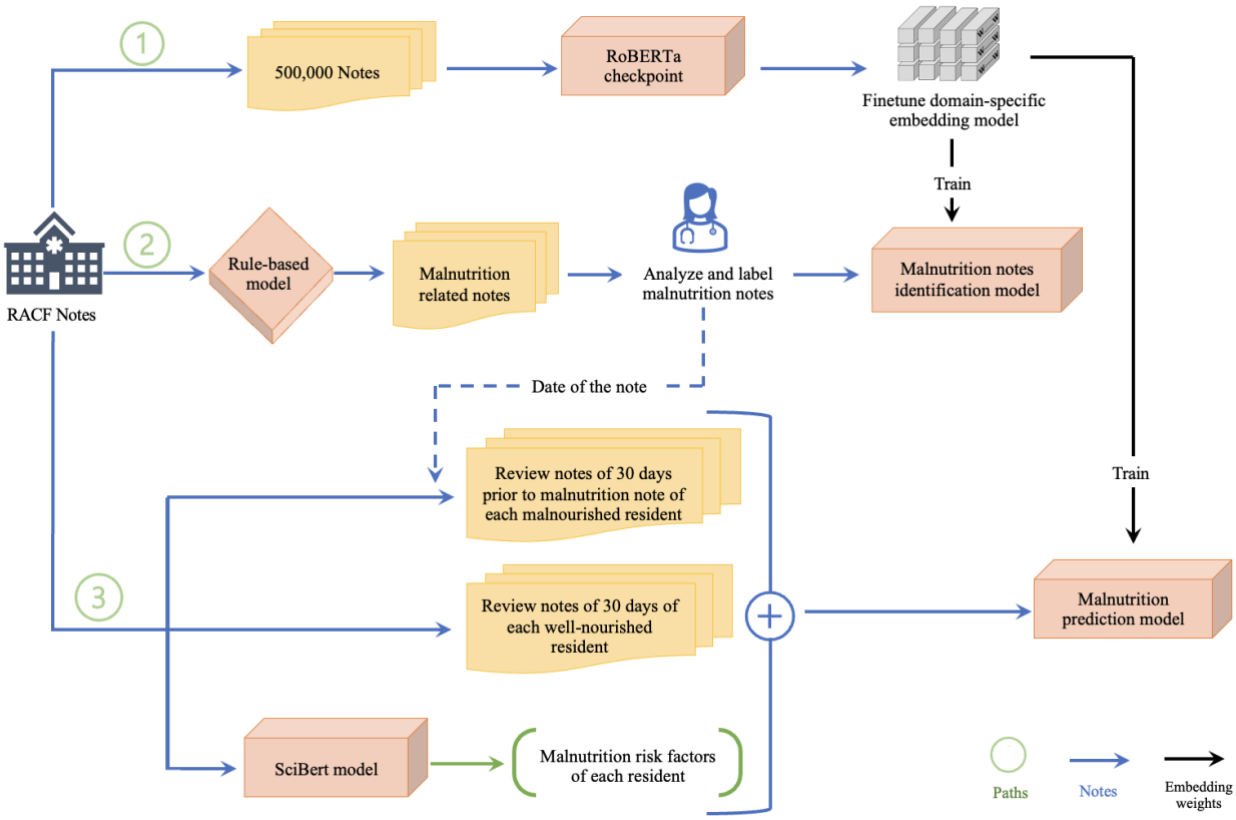


Fig. 1: An overview of the model development pathway.

Path 1: Fine-tuning a RAC domain-specific model
Path 2: Fine-tuning a malnutrition note identification model
Path 3: Fine-tuning a malnutrition prediction model

2.4 Path 1: Fine-tuning domain-specific embedding model

2.4.1 Dataset construction

We randomly selected 500,000 free-text nursing notes with an average token length of 64. The training dataset included 21,969,925 words, which we considered adequate for domain-specific fine-tuning as more notes do not necessarily lead to better results [32]. We then extracted the raw text to a single text file and processed it into chunks of 512 tokens. This chunking procedure resulted in a training set containing 62,273 text chunks (rows).

2.4.2 Model fine-tuning

The weights of the baseline RoBERTa model checkpoint ("roberta-base") downloaded from the Huggingface transformer library [33] contain substantial information regarding the English language corpus. This substantially reduces the fine-tuning time adapting RoBERTa to our specific task than training a model entirely from scratch. However, the corpus of nursing progress notes contains many RAC domain-specific terms, abbreviations and unconventional expressions that do not present in general English, which could affect the performance of the baseline RoBERTa model. Therefore, we chose to train a RAC domain-specific model initialized from RoBERTa on our nursing note dataset to improve the model's ability to understand the words and phrases used in the RAC nursing corpus. The task for the model is to predict words randomly masked out of an input chunk. The knowledge of the resulting model can be transferred and further trained with an additional output layer to create models for various downstream tasks.

Tokenization was conducted on the nursing text corpus using a pre-trained byte-level tokenizer to fit with the RoBERTa model. We randomly split the dataset into 80% training and 20% validation sets. Then, we set the masking probability to 15% of the words in each input

sequence, like the original RoBERTa training. We used whole word masking instead of token masking for better results [34] (Supplementary Table S1). In addition, we randomize the masking with each batch to avoid over-memorization. After that, we trained the model with the following hyperparameters: learning rate of 1e-4, batch size of 32, and weight decay of 0.01. The model was trained until validation loss started to converge (Supplementary Figure S1). The embeddings of this model were then utilized for the two downstream tasks: malnutrition note identification and malnutrition prediction.

2.5 Path 2: Downstream task 1: Fine-tuning a malnutrition note identification model

2.5.1 Dataset construction

Transformer models need a considerable amount of labelled data to boost their performance. However, to the best of our knowledge, there is not any publicly available, malnutrition-labelled data; therefore, we engaged three nursing domain experts to build a malnutrition-specific labelled dataset. To accomplish this, we developed process to identify and label records with malnutrition [35]. We first constructed a rule-based model to identify all malnutrition notes in the dataset. Using these rules, we extracted 2,474 notes belonging to 1,283 clients. Manual analysis and screening of all extracted notes identified 196 notes that did not fit the malnutrition definition, but either reported planned weight loss or invalid weight recording due to scale or typing errors. At the end, our manually labelled ground truth training dataset contained 2,278 notes reporting malnutrition (labelled: 1) and 15,000 notes with normal nutrition status (labelled: 0).

2.5.2 Model finetuning

We further fine-tuned our model with the embeddings from the RAC domain-specific LLM that we built in Path 1 to identify notes related to malnutrition (see Figure 1). We divided the dataset

into 85% training and validation datasets, and 15% for hold-out testing set. The hyperparameters included learning rate of 3e-5, batch size of 16, weight decay of 0.01 and 50% of dropout rate. We used binary cross-entropy loss with positive weights and the mean pooling output of the last hidden state.

2.6 Path 3: Downstream task 2: Fine-tuning a malnutrition prediction model

2.6.1 Dataset construction

The dataset for this task consisted of the original weekly nursing review notes and the malnutrition risk factors extracted from these notes. Since malnutrition is a health condition that develops over time, to capture each client's health changes over time, we approached this task as a time series data analysis by extracting weekly review notes of each malnourished client recorded in the 30 days prior to the onset of malnutrition. We organized each client's notes chronologically, with the earliest note appearing first in the sequence and the most recent note appearing last. We followed the same procedures to organize data for clients without malnutrition.

In addition to text-based notes, we also extracted malnutrition risk factors for each client from the notes using the SciBert model for named entity recognition with UMLS linker [36]. In our previous study, we identified 46 malnutrition risk factors in each client's notes such as poor appetite, suboptimal oral intake and dysphagia [35]. Moreover, we applied a negation technique to distinguish whether a factor mentioned in a note was confirmed or negated [37]. For instance, if a note has the following sentence "no sign of cancer", the algorithm will correctly identify this as a negation and not a confirmed factor. Finally, we combined each client's notes and the risk factors into one file. We added the notes as raw text data and the risk factors as one-hot encoding

tensor, with '0' indicating the absence of the factor and '1' indicating its presence. After that, we used notes and factors as the dataset for the malnutrition prediction model.

2.6.2 Model finetuning

Our model was initialized from the RAC domain-specific model fine-tuned in Path 1. The training dataset consisted of 862 aggregated notes (rows) of malnourished clients and 2,298 aggregated notes (rows) of well-nourished clients. We split the data into 85% for training and validation, and 15% for hold-out testing. Hyperparameters included a learning rate of 3e-5, batch size of 16, weight decay of 0.01 and 50% dropout rate. We used binary cross-entropy loss with positive weights. We concatenated the output with the structured data (malnutrition risk factors). We added to the concatenated data a fully connected layer, with a Sigmoid activation function applied to obtain the final output (Supplementary Figure S2).

2.6.3 Addressing the 512 maximum length challenge

In this downstream task, as opposed to Task 1, the notes were longer due to the inclusion of information gathered over a four-week period. Therefore, we encountered long notes spanning a four-week duration, with an average 644 token length (95% confidence interval:627.17 - 663.31). Therefore, the length of certain records exceeded the maximum sequence length accepted by RoBERTa and BERT, which is 512 tokens.

For these long records, we truncated and padded the text sequence into equal 512 token parts. Each part starts with a start sequence token and concludes with an end sequence token. Padding token was added if the last part has less than 512 tokens. Attention masks were also manually added as (1), informing the model to pay attention to the token, or (0), suggesting the model to ignore the token.

In the model forward function, the last hidden state embeddings of each token, generated by the model, are selectively emphasized through an attention mask. Then the sum of the masked embeddings is calculated. Only tokens with an attention mask value of 1 are considered; tokens with an attention mask of 0, which indicated padding token, are ignored. In addition, the model keeps track of the number of tokens. To capture the main ideas of a note, for short notes, embeddings are averaged across all tokens (Figure 2A). Conversely, as longer notes are divided into several parts with equal length of 512 tokens, embedding tokens are aggregated across all parts (Figure 2B).

All tasks were evaluated using precision, recall, F1-score, specificity, the area under the precision-recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC). To better assess the model's robustness, generalizability and avoid finetuning instability [38], we performed 5-fold cross-validation in each downstream task. We kept the number of epochs in each fold to a low number (four) to avoid the possibility of overfitting the data. In each fold, the model with the least validation loss was utilized for testing on the test dataset. We calculated the cross-validation performance by taking the average of the k performance estimates of all measures (F1-score, recall, etc.) obtained from the testing sets using the arithmetic mean. Additionally, we calculated the confidence interval for each measure.

This study was implemented using Python 3.10.11, PyTorch 2.0.1, transformers 4.29.2 (Huggingface) and Scikit-learn 1.2.2. Our models were trained on the NVIDIA Tesla T4-16GB graphics processing unit (GPU).

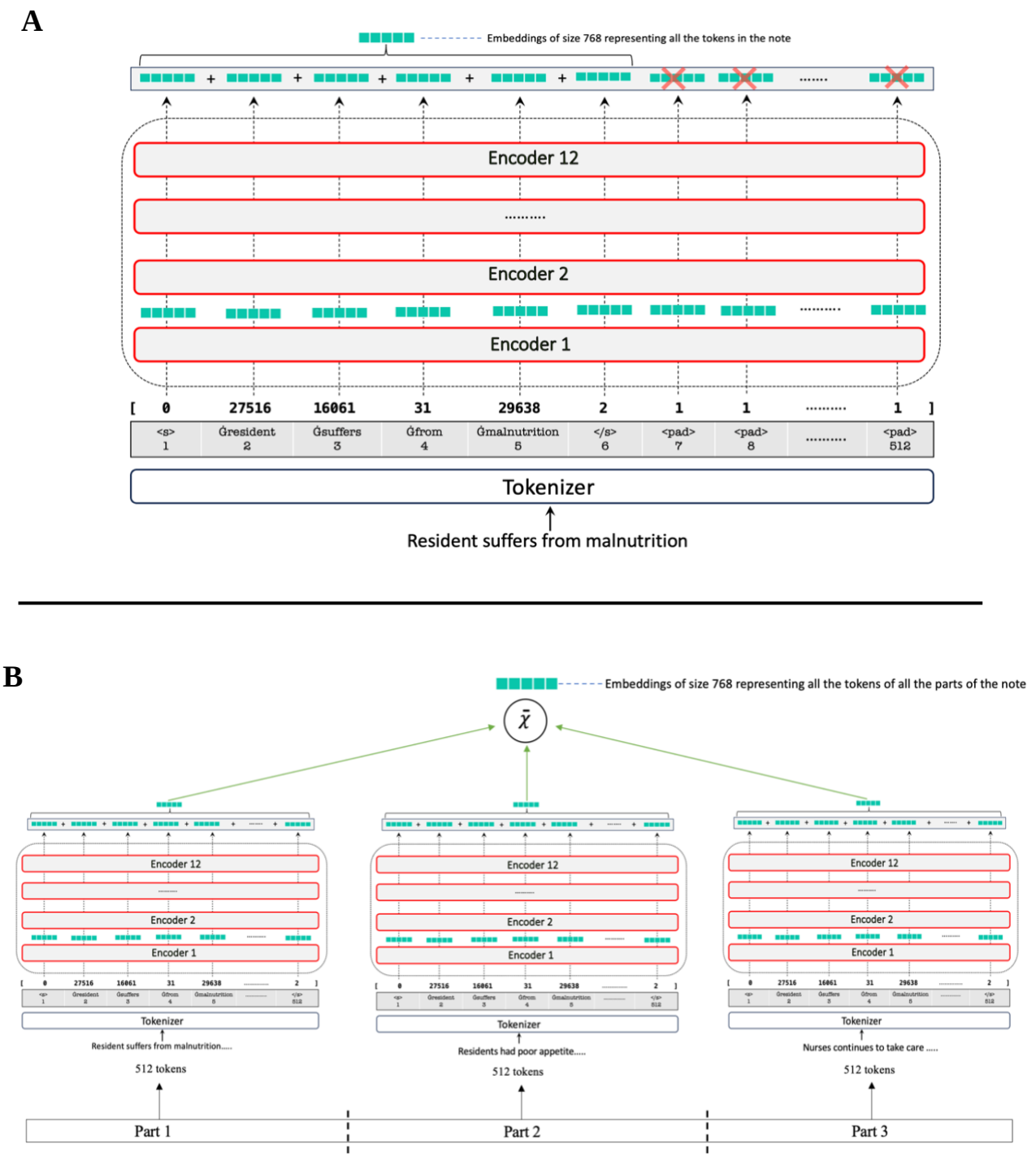


Fig. 2: Methods developed for processing long notes with more than 512 tokens. (**A**) Example of a nursing note with sequence length less than 512 tokens; (**B**) Example of a nursing note with sequence length of 1536 tokens. This note is truncated into 3 parts each with a sequence of 512 tokens.

3. Result

We compared the results of our domain specific LLM to the baseline models, RoBERTa, and BioClinicalBERT. In the first downstream task (malnutrition note identification), LLMs had very similar results, with RAC domain specific LLM producing a slightly better F1-score of 0.966, followed by RoBERTa, which achieved a comparable F1-score of 0.964. Then, BioClinicalBERT had an F1-score of 0.960 (Table 2).

Table 2: Performance of the five machine learning models on the task of identifying malnutrition notes

Model	Precision (95% CI)	Recall (95% CI)	F1-Score* (95% CI)	Specificity (95% CI)	AUPRC* (95% CI)	AUROC* (95% CI)
BioClinicalBERT	0.932 (0.91 – 0.95)	0.988 (0.98 – 0.99)	0.960 (0.95 – 0.97)	**0.990 (0.99 – 0.99)**	0.966 (0.95 – 0.98)	**1.0 (1.0 – 1.0)**
roberta-base	0.934 (0.92 – 0.95)	**0.994 (0.99 – 1.0)**	0.964 (0.96 – 0.97)	**0.990 (0.99 – 0.99)**	0.958 (0.94 – 0.97)	**1.0 (1.0 – 1.0)**
RAC domain-specific LLM	**0.942 (0.94 – 0.95)**	**0.994 (0.99 – 1.0)**	**0.966 (0.96 – 0.97)**	**0.990 (0.99 – 0.99)**	**0.978 (0.97 – 0.99)**	**1.0 (1.0 – 1.0)**

* F1-score computed using 0.5 threshold

* AUPRC and AUCROC computed across various threshold values

In the second downstream task of malnutrition prediction, LLMs demonstrated superior performance to the older techniques. RAC domain specific LLM with the risk factor layer was the top-performing model with an F1-score of 0.687. It was followed by our domain specific LLM without the risk factors, which had an F1-score of 0.655. Next, RoBERTa had an F1-score of 0.614. Then, BioClinicalBERT achieved an F1-score of 0.582 (Table 3). Supplementary Figures (S3 – S9) show the area under the curve plots for each model.

Table 3: Results of the malnutrition prediction model

Model	Precision (95% CI)	Recall (95% CI)	F1-Score* (95% CI)	Specificity (95% CI)	AUPRC* (95% CI)	AUROC* (95% CI)
BioClinicalBERT	0.554 (0.51 – 0.59)	0.617 (0.54 – 0.70)	0.582 (0.54 – 0.62)	0.787 (0.74 – 0.84)	0.613 (0.56 – 0.67)	0.771 (0.74 – 0.80)
roberta-base	0.579 (0.50 – 0.66)	0.662 (0.59 – 0.74)	0.614 (0.58 – 0.65)	**0.789** **(0.71 – 0.87)**	0.677 (0.66 – 0.70)	0.803 (0.79 – 0.82)
RAC domain-specific LLM	0.592 (0.50 – 0.68)	0.751 (0.64 – 0.86)	0.655 (0.62 – 0.69)	0.766 (0.63 – 0.90)	0.734 (0.67 – 0.80)	0.843 (0.82 – 0.87)
RAC domain-specific LLM + risk factors	**0.615** **(0.53 – 0.70)**	**0.790** **(0.71 – 0.87)**	**0.687** **(0.65 – 0.72)**	0.780 (0.67 – 0.89)	**0.735** **(0.68 – 0.79)**	**0.858** **(0.83 – 0.88)**

* F1-score computed using 0.5 threshold

* AUPRC and AUCROC computed across various threshold values

4. Discussion

The aim of this study was to develop an encoder-based RAC domain specific LLM to accurately identify clients with malnutrition in EHR and develop a model capable of predicting malnutrition in older people one month before its onset. We first utilized our RAC domain-specific LLM, initialized from the well-known RoBERTa model and further fine-tuned on nursing progress notes. Afterwards, we employed the embedding weights generated from the proposed model for two subsequent downstream tasks. We compared the performance of different parameters on three models, BioClinicalBERT, roberta-base and our domain-specific LLM. The results demonstrated the advantage of utilizing domain-specific embeddings. It is worthy to note that this is the first study to utilize LLMs on free text nursing notes to predict malnutrition in older people, although it has been a health risk that has long plagued the care staff members and has casted a negative impact on care quality [1]. This study has also developed a method for processing long notes (>

512 tokens), which is crucial and particularly relevant in health contexts where it is typical to encounter long and detailed documentation.

For the first downstream task, there have been very few attempts to identify malnutrition notes in EHR in the literature. One attempt to classify malnutrition notes applies conditional random fields technique to nursing notes [39]. However, the study reported that the model performed poorly on malnutrition and had a low F1-score of 0.39. The authors of the study stated that classifying malnutrition notes was very challenging which led to the low accuracy of their model. Another attempt is our previous work to classify the malnutrition notes using a rule-based model which achieved a high-level performance; however, the development of the rule-base method was time-consuming and labor-intensive [35]. To address this limitation, we adopted a specific domain LLM in this study. The process of fine-tuning and evaluating the model was much more efficient than the rule-based method.

In accordance with the previous reports [22, 40, 41], LLMs significantly outperformed other comparative models with our RAC domain-specific LLM achieving an F1-score of 0.966. However, all LLMs models yielded comparable performance in this task. We argue that this can be attributed to the notable difference in notes between the positively labelled and negatively labelled instances in the dataset compared to those in the second task.

For the second downstream task of malnutrition prediction, to our knowledge, this is the first study in predicting malnutrition for older people in RACFs applying a transfer learning approach to EHR. Once again, our RAC domain-specific LLM notably outperformed other LLMs and had the highest F1-score of 0.655. This illustrates the importance of fine-tuning a foundational LLM on a domain-specific corpus. In addition, combining a layer of structured data with the output of the note-based model increased the performance of the model, as evidence by an increased F1-

score (0.687). Despite that, the model did not achieve a high F1-score like in the first task, which is arguably because malnutrition is a health risk that is influenced by various factors, many of which are prevalent in all clients and are not specific for clients with malnutrition. Therefore, accurately predicting malnutrition is still a complex and challenging mission [6, 42].

Our findings in this study will be practically and clinically important for the malnutrition management of older people in RACFs. We would also stress that our LLM is not a replacement for already existing reliable screening tools. Instead, it could be incorporated into the nursing care process to help nurses and clinicians efficiently identify clients at risk of malnutrition by automatic screening of their EHR data. This will enable them to implement the tailored malnutrition prevention and intervention actions accordingly. In addition, our study demonstrates the feasibility of using a robust, well-known LLM such as RoBERTa can facilitate researchers to produce the optimal model for various downstream tasks.

The area of LLMs is evolving at a rapid pace; recently, the decoding models such as GPT-3.5 and GPT-4, though with widely doubted hypes, have revolutionized the whole field of NLP. However, there remains a paucity of usable models for health care systems [26]. We intend to compare the prediction ability of our encoder-based LLM with one of the more recent and advanced decoder-based LLMs such as Llama 2 [29]. We will also further gather nursing notes to improve the performance of the model, particularly for the second task.

4.1 Limitation

This study has two notable limitations. Firstly, LLMs typically require substantial training data to achieve high levels of accuracy [15]. In the case of our malnutrition prediction task, the dataset size is relatively modest, potentially impacting the model's predictive performance. The availability of a larger dataset could lead to improved accuracy and more robust predictions.

Secondly, although our models are trained on data sourced from a diverse array of 40 RACFs, it's crucial to acknowledge that these facilities are all part of a single organization. Consequently, the applicability of our models might be constrained when used in RACFs with differing strategies and guidelines for electronic data collection. The lack of diversity in institutional practices could potentially hinder the models' generalizability to a broader range of settings.

5. Conclusion

To address the critical malnutrition issue in older people, we proposed an encoder-based RAC domain-specific LLM that is fine-tuned from the foundation LLM, RoBERTa model, on RAC domain-specific nursing text. The resulted embeddings were successfully utilized for two downstream tasks: malnutrition note identification and malnutrition prediction. Our findings demonstrate that fine-tuning a foundation LLM with domain-specific corpus can improve the performance of the foundation models. In addition, combining risk factors as a structured data with a text model enhance model performance. This study also developed a method to truncate long text into parts that fit into the 512-token limit of RoBERTa model.

Author contribution

Design of work: Mohammad Alkhalaf and Ping Yu. Data analysis and interpretation: Mohammad Alkhalaf and Ping Yu. Drafting of the manuscript: Mohammad Alkhalaf and Ping Yu. Critical review and revision: all authors.

Acknowledgments

The first author, Mohammad Alkhalaf, is supported by a full PhD scholarship from Qassim University, Saudi Arabia. The authors are grateful for the aged care organization that shared the

de-identified electronic health records, which provided the opportunity to conduct this significant research project.

Statements and Declarations

The authors declare that they have no conflicts of interests that could have appeared to influence the work reported in this paper.

Statement on funding

No funding was received for conducting this study.

Ethical approval

The Human research ethics approval for this study was granted by the Human Research Ethics Committee, the University of Wollongong and the Illawarra Shoalhaven Local Health District (Year 2020).

References

[1] E. Dent, O. R. L. Wright, J. Woo, and E. O. Hoogendijk, "Malnutrition in older adults," (in English), *The Lancet*, vol. 401, no. 10380, pp. 951-966, 2023 Mar 18 2023-11-29 2023, doi: https://doi.org/10.1016/S0140-6736(22)02612-5.

[2] M. I. T. D. Correia and D. L. Waitzberg, "The impact of malnutrition on morbidity, mortality, length of hospital stay and costs evaluated through a multivariate model analysis," *Clinical Nutrition*, vol. 22, no. 3, pp. 235-239, 2003, doi: 10.1016/S0261-5614(02)00215-7.

[3] J. Edington *et al.*, "Prevalence of malnutrition on admission to four hospitals in England," *Clinical Nutrition*, vol. 19, no. 3, pp. 191-195, 2000, doi: 10.1054/clnu.1999.0121.

[4] R. J. Stratton, C. L. King, M. A. Stroud, A. A. Jackson, and M. Elia, "'Malnutrition Universal Screening Tool' predicts mortality and length of hospital stay in acutely ill elderly," *British Journal of Nutrition*, vol. 95, no. 2, pp. 325-330, 2006, doi: 10.1079/bjn20051622.

[5] *Aged care - Australian Institute of Health and Welfare*. 2019, vailable: URL.

[6] E. Dent, E. O. Hoogendijk, R. Visvanathan, and O. R. L. Wright, "Malnutrition Screening and Assessment in Hospitalised Older People: A Review," *Journal of Nutrition, Health and Aging*, vol. 23, no. 5, pp. 431-441, 2019, doi: 10.1007/s12603-019-1176-z.

[7] J. Kellett, G. Kyle, C. Itsiopoulos, and M. Naunton, "Nutrition screening practices amongst australian Residential Aged Care Facilities," *Journal of Nutrition, Health and Aging*, vol. 20, no. 10, pp. 1040-1044, 2016, doi: 10.1007/s12603-015-0693-7.

[8] E. Amarantos, A. Martinez, and J. Dwyer, "Nutrition and quality of life in older adults," *The Journals of Gerontology: Series A* vol. 56, no. suppl_2, pp. 54-64, 2001, doi: 10.1093/gerona/56.suppl_2.54.

[9] P. Yu and S. Qian, "Developing a theoretical model and questionnaire survey instrument to measure the success of electronic health records in residential aged care," *PLOS ONE*, vol. 13, no. 1, p. e0190749, 2018, doi: 10.1371/journal.pone.0190749.

[10] N. Wang, P. Yu, and D. Hailey, "The quality of paper-based versus electronic nursing care plan in Australian aged care homes: A documentation audit study," *International Journal of Medical Informatics*, vol. 84, no. 8, pp. 561-569, 2015, doi: 10.1016/j.ijmedinf.2015.04.004.

[11] X. Xiong *et al.*, "Knowledge-driven online multimodal automated phenotyping system," (in eng), *medRxiv*, Oct 2 2023, doi: 10.1101/2023.09.29.23296239.

[12] S. Graham *et al.*, "Artificial Intelligence for Mental Health and Mental Illnesses: an Overview," *Current Psychiatry Reports*, vol. 21, no. 11, 2019, doi: 10.1007/s11920-019-1094-0.

[13] E. Iqbal *et al.*, "ADEPt, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health records," *PLoS ONE*, vol. 12, no. 11, pp. 1-16, 2017, doi: 10.1371/journal.pone.0187121.

[14] X. Zhou, Y. Wang, S. Sohn, T. M. Therneau, H. Liu, and D. S. Knopman, "Automatic extraction and assessment of lifestyle exposures for Alzheimer's disease using natural language processing," *Int J Med Inform*, vol. 130, no. August, pp. 103943-103943, 2019, doi: 10.1016/j.ijmedinf.2019.08.003.

[15] J. Pennington, R. Socher, and C. D. Manning, "GloVe : Global Vectors for Word Representation," in *InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Doha, 2014, pp. 1532-1543, doi: 10.3115/v1/D14-1162.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv*, pp. 1-12, 2013, doi: 10.48550/arXiv.1301.3781.

[17] C. Tao, M. Filannino, and Ö. Uzuner, "Prescription extraction using CRFs and word embeddings," *Journal of Biomedical Informatics*, vol. 72, pp. 60-66, 2017, doi: 10.1016/j.jbi.2017.07.002.

[18] H. Wu, B. Zhong, B. Medjdoub, X. Xing, and L. Jiao, "An Ontological Metro Accident Case Retrieval Using CBR and NLP," *Applied Sciences*, vol. 10, no. 15, pp. 5298-5298, 2020, doi: 10.3390/app10155298.

[19] J. Devlin, C. Ming-Wei, K. Lee, and K. Toutanova, "BERT- Pre-training of Deep Bidirectional Transformers for Language Understanding," presented at the 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), Minneapolis, Minnesota, 2019vailable: URL.

[20] O. M. Liu Y, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V., "RoBERTa-A Robustly Optimized BERT Pretraining Approach," *arXiv*, 2019, doi: https://doi.org/10.48550/arXiv.1907.11692.

[21] P. Gupta, P. Malhotra, J. Narwariya, L. Vig, and G. Shroff, "Transfer Learning for Clinical Time Series Analysis Using Deep Neural Networks," *J Healthc Inform Res*, vol. 4, no. 2, pp. 112-137, Jun 2020, doi: 10.1007/s41666-019-00062-3.

[22] E. Alsentzer *et al.*, "Publicly Available Clinical BERT Embeddings," *arXiv*, 2019, doi: 10.18653/v1/W19-1909.

[23] T. L. Chen *et al.*, "Domain specific word embeddings for natural language processing in radiology," *J. Biomed. Inform.*, vol. 113, p. 103665, Jan 2021, doi: 10.1016/j.jbi.2020.103665.

[24] C. C. Chiang *et al.*, "A large language model-based generative natural language processing framework finetuned on clinical notes accurately extracts headache frequency from electronic health records," (in eng), *medRxiv*, Oct 3 2023, doi: 10.1101/2023.10.02.23296403.

[25] J. Ge, M. Li, M. B. Delk, and J. C. Lai, "A comparison of large language model versus manual chart review for extraction of data elements from the electronic health record," (in eng), *medRxiv*, Sep 4 2023, doi: 10.1101/2023.08.31.23294924.

[26] M. Wornow *et al.*, "The shaky foundations of large language models and foundation models for electronic health records," (in eng), *NPJ Digit Med*, vol. 6, no. 1, p. 135, Jul 29 2023, doi: 10.1038/s41746-023-00879-8.

[27] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT," p. arXiv:2302.10198doi: 10.48550/arXiv.2302.10198.

[28] M. M. Amin, E. Cambria, and B. W. Schuller, "Will Affective Computing Emerge From Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT," *IEEE Intelligent Systems*, vol. 38, no. 2, pp. 15-23doi: 10.1109/MIS.2023.3254179.

[29] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," p. arXiv:2307.09288doi: 10.48550/arXiv.2307.09288.

[30] H. Naveed *et al.*, "A Comprehensive Overview of Large Language Models," p. arXiv:2307.06435doi: 10.48550/arXiv.2307.06435.

[31] T. Savage, J. Wang, and L. Shieh, "A large language model screening tool to target patients for best practice alerts: Development and validation," *JMIR Med Inform*, vol. 11, p. e49886, 2023/11/27 2023, doi: 10.2196/49886.

[32] F. Li, Y. Jin, W. Liu, B. P. S. Rawat, P. Cai, and H. Yu, "Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study," *JMIR Medical Informatics*, vol. 7, no. 3, pp. e14830-e14830, 2019, doi: 10.2196/14830.

[33] T. Wolfe, L. Debut, V. Sanh, J. Chaumond, and A. Rush, "Transformers: State-of-the-Art Natural Language Processing," in *EMNLP (Systems Demonstrations)*, 2020, pp. 38-45, doi: 10.18653/v1/2020.emnlp-demos.6.

[34] Y. Dai *et al.*, "Is Whole Word Masking Always Better for Chinese BERT?"- Probing on Chinese Grammatical Error Correction," *arXiv*, 2022, doi: https://doi.org/10.48550/arXiv.2203.00286.

[35] M. Alkhalaf *et al.*, "Malnutrition and its contributing factors for older people living in residential aged care facilities: Insights from natural language processing of aged care records," *Technology and Health Care*, vol. Preprint, pp. 1-12, 2023, doi: 10.3233/THC-230229.

[36] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," p. arXiv:1902.07669doi: 10.48550/arXiv.1902.07669.

[37] *Industrial-Strength Natural Language Processing*. (2015). [Online]. Available: https://spacy.io/

[38] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, "Fine-Tuning Pretrained Language Models- Weight Initializations, Data Orders, and Early Stopping," *arXiv*, 2020, doi: https://doi.org/10.48550/arXiv.2002.06305.

[39] T. Chen, M. Dredze, J. P. Weiner, L. Hernandez, J. Kimura, and H. Kharrazi, "Extraction of Geriatric Syndromes From Electronic Health Record Clinical Notes: Assessment of Statistical Natural Language Processing Methods," *JMIR Med Inform*, vol. 7, no. 1, p. e13039, 2019, doi: 10.2196/13039.

[40] P. Lewis, M. Ott, J. Du, and V. Stoyanov, "Pretrained Language Models for Biomedical and Clinical Tasks-Understanding and Extending the State-of-the-Art," 2020: Association for Computational Linguistics, pp. 146-157, doi: 10.18653/v1/2020.clinicalnlp-1.17.

[41] C. H. Lin *et al.*, "A disease-specific language representation model for cerebrovascular disease research," *Comput Methods Programs Biomed*, vol. 211, p. 106446, Nov 2021, doi: 10.1016/j.cmpb.2021.106446.

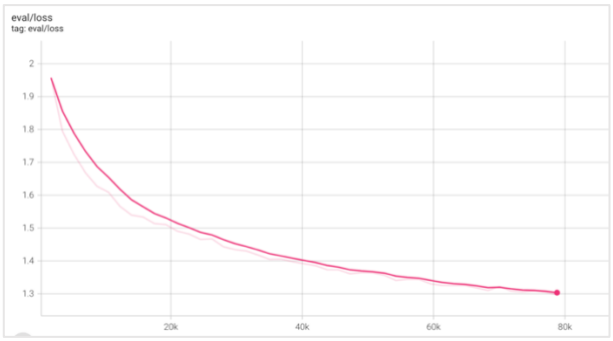
[42] E. Agarwal, M. Miller, A. Yaxley, and E. Isenring, "Malnutrition in the elderly: A narrative review," *Maturitas*, vol. 76, no. 4, pp. 296-302, 2013, doi: 10.1016/j.maturitas.2013.07.013.

6. Supplementary materials

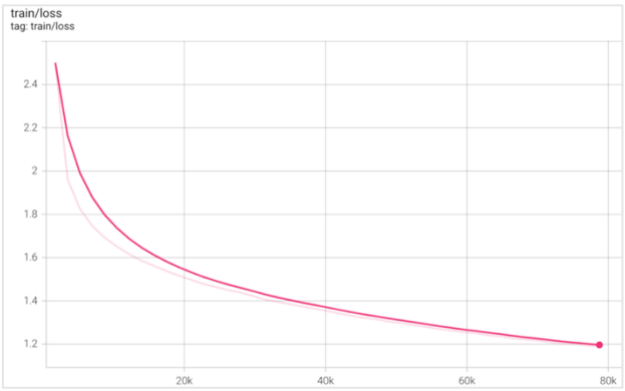
Supplementary Table S1: Example of tokenization and 15% whole word masking

Original	Client had physical impairment and unsteady gait related to Parkinson disease and has high malnutrition risk.
Toknized	'ĠClient', 'Ġhad', 'Ġphysical', 'Ġimpairment', 'Ġand', 'Ġunst', 'ead', 'y', 'Ġg', 'ait', 'Ġrelated', 'Ġto', 'ĠParkinson', 'Ġdisease', 'Ġand', 'Ġhas', 'Ġhigh', 'Ġmalnutrition', 'Ġrisk', '.'
Masking (Whole word masking is in red colour)	Client had physical impairment<mask><mask><mask> gait<mask> to Parkinson disease and has high malnutrition <mask>.

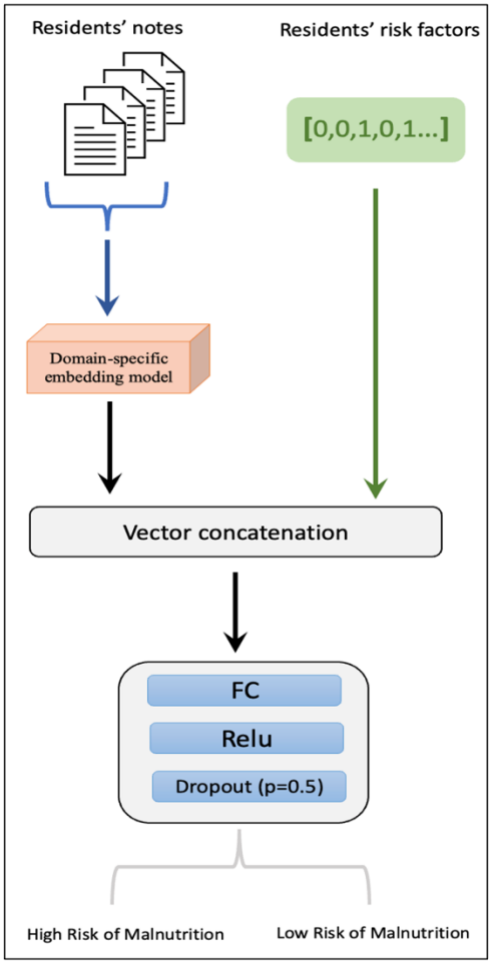
A.



B.



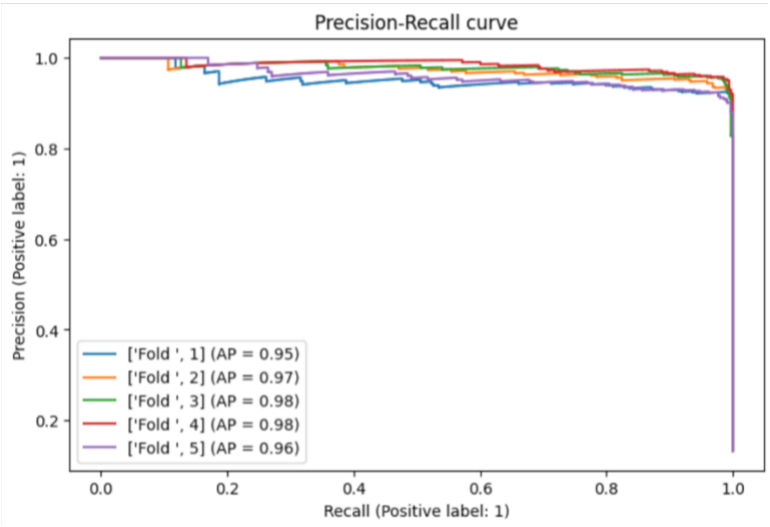
Supplementary Figure S1. The RAC domain-specific LLM training, (A) validation loss, (B) train loss



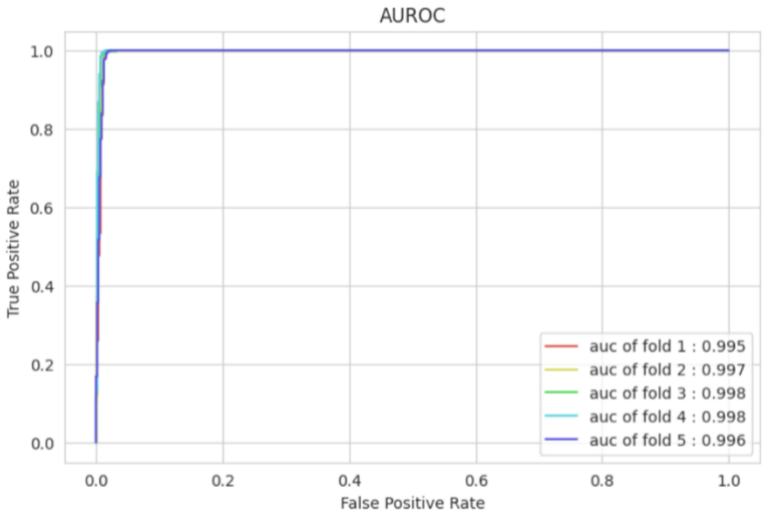
FC: Fully connected layer
P: Probability

Supplementary Figure S2. The Malnutrition prediction model

A

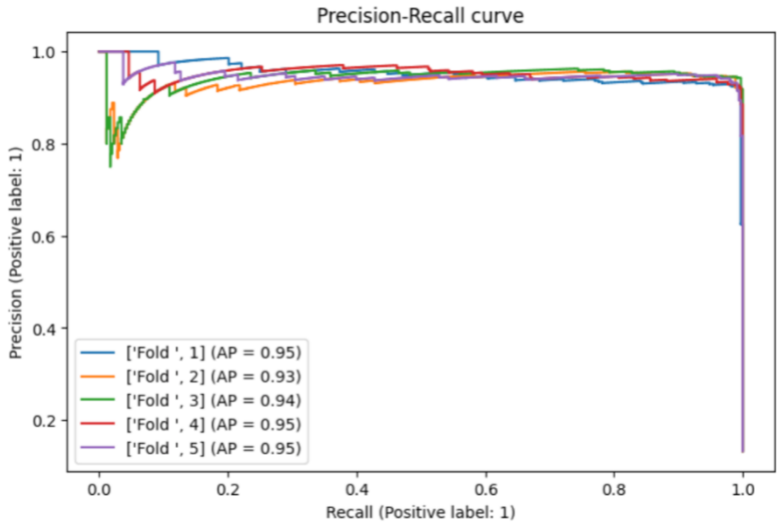


B

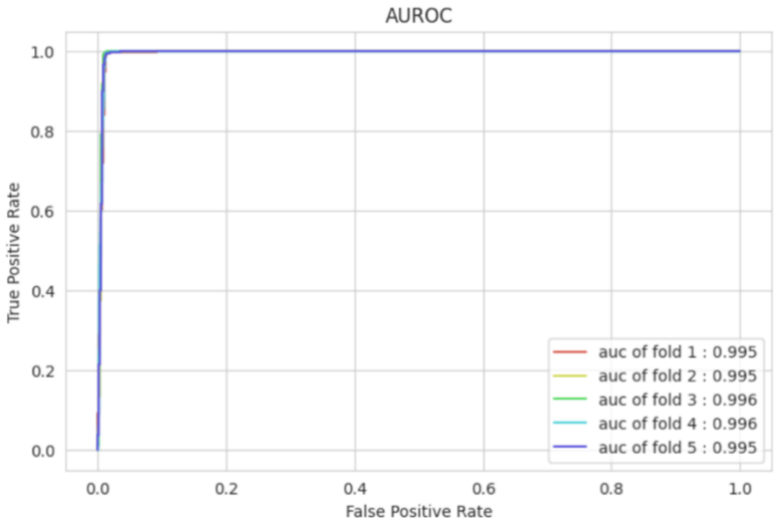


Supplementary Figure S3. AUPRC (A) and AUROC (B) of malnutrition note identification model - BioClinicalBERT

A

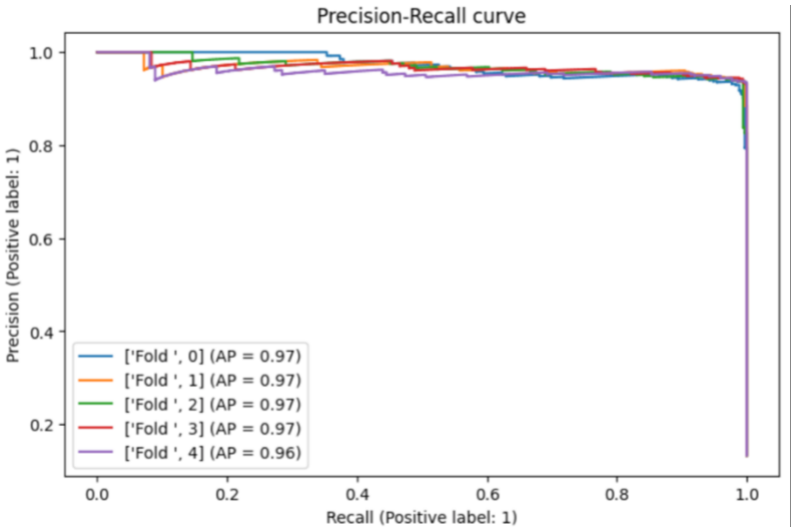


B

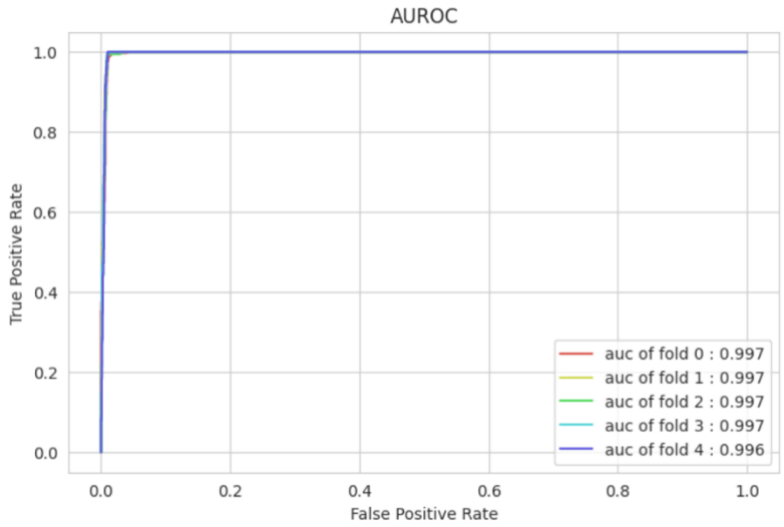


Supplementary Figure S4. AUPRC (A) and AUROC (B) of malnutrition note identification model – RoBERTa base

A

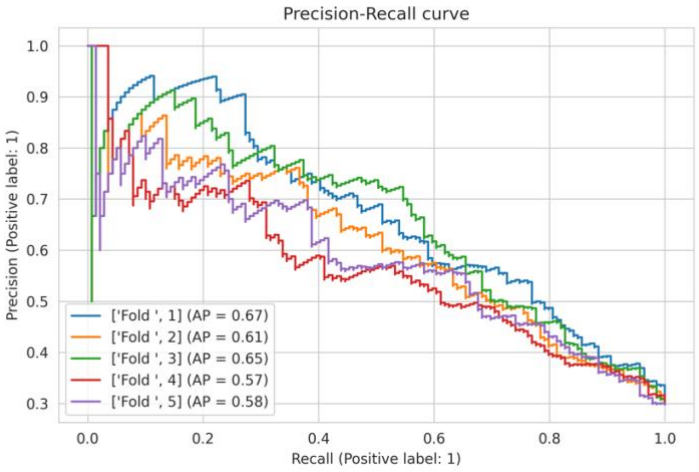


B

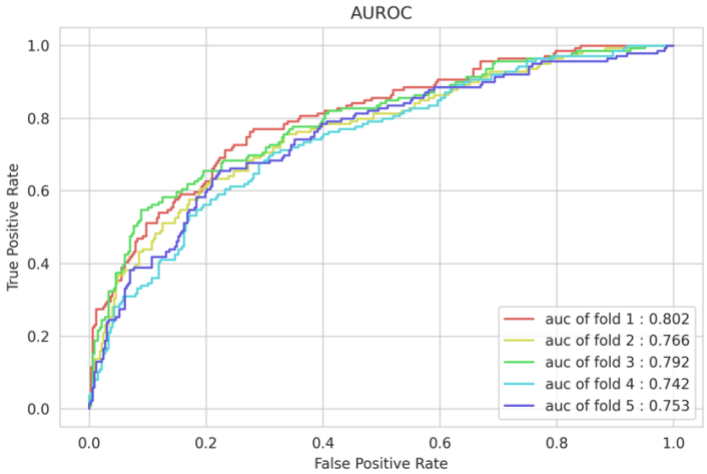


Supplementary Figure S5. AUPRC (A) and AUROC (B) of malnutrition note identification model - RAC domain-specific LLM

A

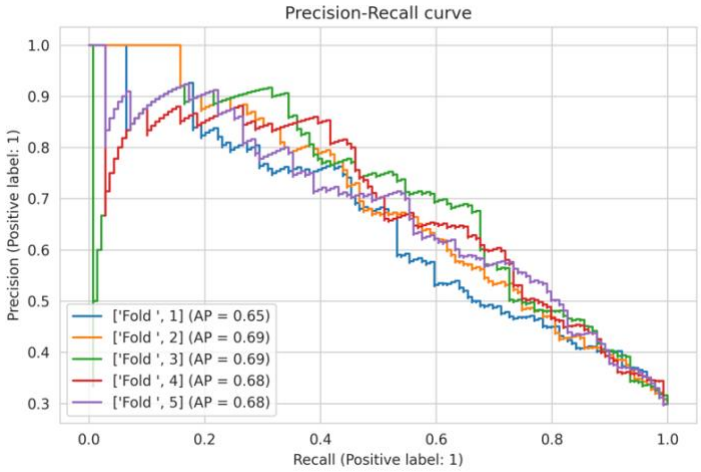


B

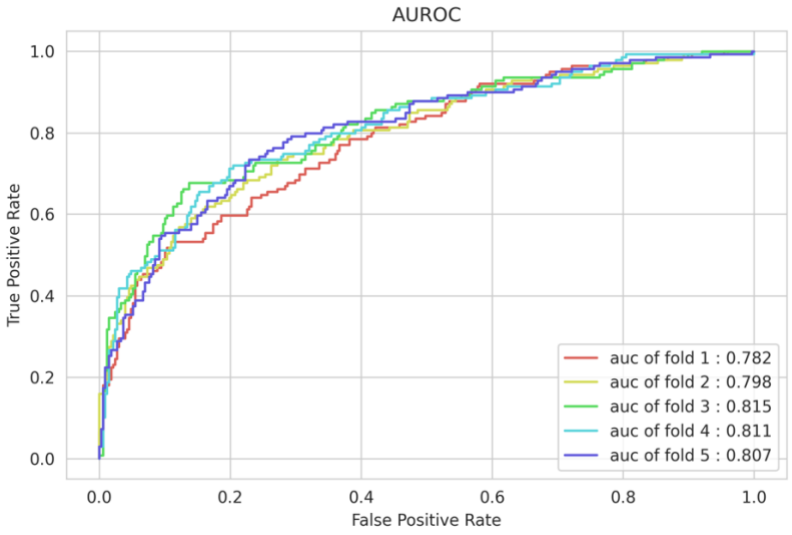


Supplementary Figure S6. AUPRC (A) and AUROC (B) for malnutrition prediction model - BioClinicalBERT

A

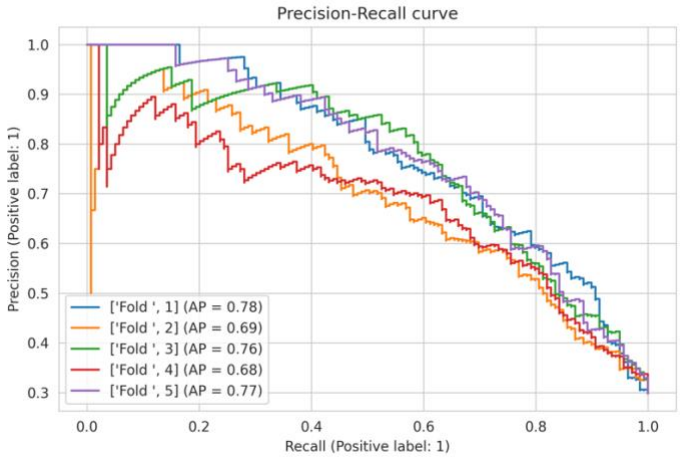


B

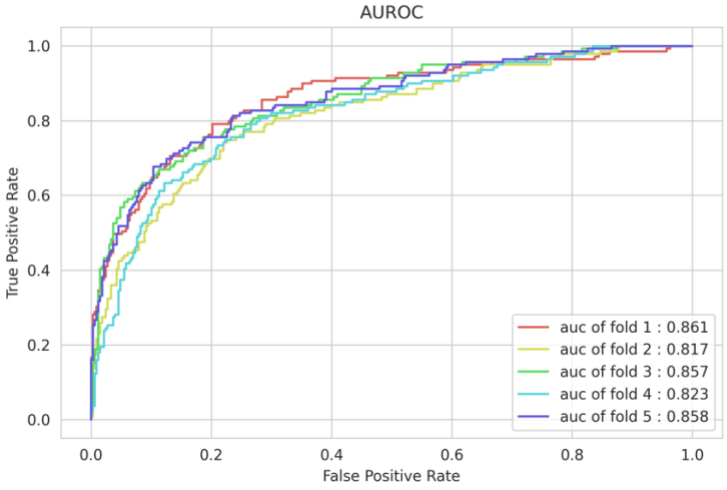


Supplementary Figure S7. AUPRC (A) and AUROC (B) for malnutrition prediction model - RoBERTa base

A

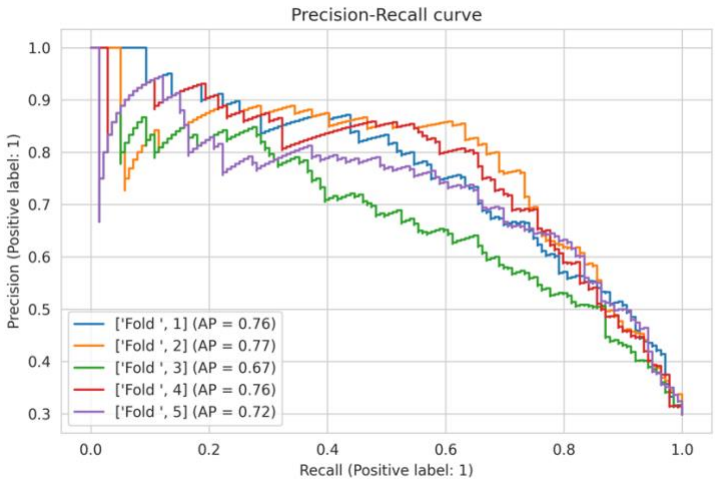


B

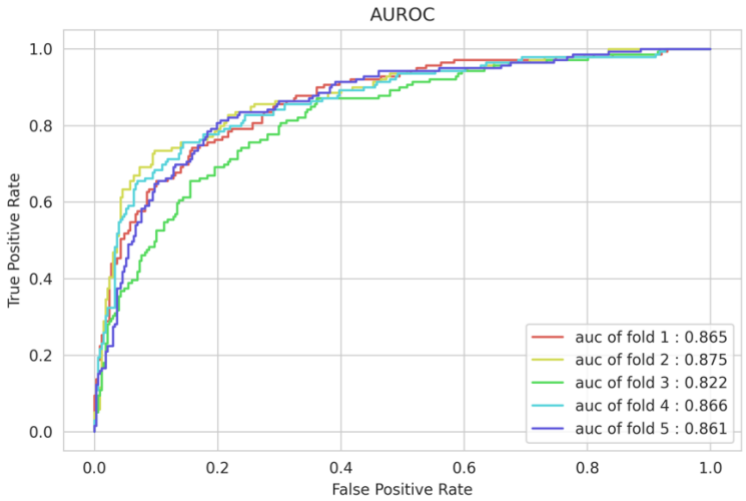


Supplementary Figure S8. AUPRC (A) and AUROC (B) for malnutrition prediction model - RAC domain-specific LLM

A



B



Supplementary Figure S9. AUPRC (A) and AUROC (B) for malnutrition prediction model - RAC domain-specific LLM with risk factor layer