Evaluating machine learning approaches for multi-label classification of unstructured electronic health records with a generative large language model

Dinithi Vithanage[1], Chao Deng[3], Lei Wang[1], Mengyang Yin[3], Mohammad Alkhalaf[1], Zhenyua Zhang [1], Yunshu Zhu[1], Alan Christy Soewargo[4], Ping Yu[1, *]

1 School of Computing and Information Technology, University of Wollongong, Wollongong, Australia
2 School of Medical, Indigenous and Health Sciences, University of Wollongong, Wollongong, Australia
3 Opal Healthcare, Sydney, Australia
4 Extranet Systems Pty Ltd, Australia

ORCiD Id: Dinithi Vithanage https://orcid.org/0000-0001-5851-7158, Chao Deng https://orcid.org/0000-0003-1147-5741, Lei Wang http://orcid.org/0000-0002-0961-0441, Mengyange Yin https://orcid.org/0000-0002-0212-4598, Zhenyua Zhang https://orcid.org/0000-0003-1853-4978, Yunshu Zhu https://orcid.org/0000-0003-2786-0775, Alan Christy Soewargo https://orcid.org/0000-0002-4808-4421, Ping Yu https://orcid.org/0000-0002-7910-9396.

Correspondence: ping@uow.edu.au

Abstract

Multi-label classification of unstructured electronic health records (EHR) poses challenges due to the inherent semantic complexity in textual data. Advances in natural language processing (NLP) using large language models (LLMs) show promise in addressing these issues. Identifying the most effective machine learning method for EHR classification in real-world clinical settings is crucial. Therefore, this experimental research aims to test the effect of zero-shot and few-shot learning prompting strategies, with and without Parameter Efficient Fine-tuning (PEFT) LLMs, on the multi-label classification of the EHR data set. The labels tested are across four clinical classification tasks: agitation in dementia, depression in dementia, frailty index, and malnutrition risk factors. We utilise unstructured EHR data from residential aged care facilities (RACFs), employing the Llama 2-Chat 13B-parameter model as our generative AI-based large language model (LLM). Performance evaluation includes accuracy, precision, recall, and F1 score supported by non-parametric statistical analyses. Results indicate the same level of performance with the same prompting template, either zero-shot or few-shot learning across the four clinical tasks. Few-shot learning outperforms zero-shot learning without PEFT. The study emphasises the significantly enhanced effectiveness of fine-tuning in conjunction with zero-shot and few-shot learning. The performance of zero-shot learning reached the same level as few-shot learning after PEFT. The analysis underscores that LLMs with PEFT for specific clinical tasks maintain their performance across diverse clinical tasks. These findings offer crucial insights into LLMs for researchers, practitioners, and stakeholders utilising LLMs in clinical document analysis.

Keywords: Natural language processing, Large language models, Electronic health records, Machine learning, Multi-label classification

1 Introduction

A substantial amount of medical predictive models have been trained, tested, and published, yet the majority of them have never been deployed into the clinical setting, which is coined "a last mile problem" [1]. This is because most of these predictive models are relied on structured health data, while many important clinical information is captured in free text clinical notes, which introduces complexity for model development and deployment.

Electronic health records in residential aged care facilities in Australia are digitised systems designed to collect, store, and display data about clients' demographics, medical diagnoses, assessments, progress notes, charts, and forms [1]. Similar as other healthcare settings [2], besides the structured diagnosis data, many important clinical information in RACFs are captured in unstructured, narrative, free-text nursing progress notes. Because free text is a more expressive and natural way for care staff to record care encounters and communicate among team members, these notes are often updated and the closest to real-time reflection of an older person's health condition.

Therefore, effectively extracting information from unstructured clinical notes in EHR is important to support clinical decision-making, improve aged care quality, and advance translational research.

Multi-label classification of free-text data is a specialised area in machine learning and NLP. Multi-label classification refers to the task of assigning multiple labels or categories to a single input instance. It involves the automated extraction of entities, concepts, events, and their relations from unstructured text [4], a challenging task because text data often has different meanings and interpretations [3] and requires the use of precise and expeditious information extraction tools [7]. Despite the advancement of various transformer-based encoder-type language models, e.g., various BERT models, Clinical NLP remains a labour-intensive process that demands a substantial amount of expertise and human effort to prepare the training data [2, 4]. This limitation has hindered the effective application of the early NLP technique in information extraction from the unstructured, free-text EHR.

The recent advancements in LLMs, such as GPT variants, T5, OPT, and Llama [5], have demonstrated the ability of these decoder models to generate text that is not only human-like but also surpasses human-level performance in certain tasks [5]. These models, when combined with machine learning techniques like pre-training, fine-tuning, and prompt-based learning [6], offer transformative potential for NLP, enabling the development of automated and adaptable systems that can extract valuable insights from the free-text EHR. This marks a significant step towards the goal of integrating health predictive models into real-world clinical systems.

We are still in the early days of applying generative AI-based LLMs to extract clinical insights from the free-text EHR. While LLMs have shown potential in answering clinical questions [7-9] and extracting clinical data from public health data sets [10], their practical application in specific clinical tasks within real-world clinical settings using clinic data remains limited [4, 8, 11]. It is yet to be determined whether prompt engineering for LLMs can meet the stringent safety standards required for healthcare applications, given their limitations in generating outputs that may contain disinformation, misinformation, bias or hallucinations [12] [8]. The optimal prompting strategies for healthcare information extraction, whether zero-shot or few-shot learning, in various contexts remain unclear. Therefore, this research aims to investigate the differential effect of zero-shot and few-shot learning prompting strategies on multi-label classification across diverse clinical domains. Understanding prompting behaviour is crucial for the safe and effective deployment of LLMs in healthcare settings.

A prompt is an input a user enters to instruct a LLM to autonomously generate sequential output [13]. A LLM uses pattern matching to identify the relationships between the words, phrases, and concepts in the prompt and, connect these with its learned patterns from the previous training and uses natural language generation to respond in a human understandable format. Prompts enable the model to adapt and comprehend specific information in a new domain, leveraging its learned knowledge stored within the pre-trained models like Llama 2, thereby expanding the model's applicability and effectiveness. Prompt learning reduces the need to introducing new parameters or extensive retraining of the model using labelled data for various tasks, thus improves efficiency and reduces computational resources required for machine learning.

There are different formats of prompt-based learning. In this study, we test zero-shot and few-shot learning.

1.1 Zero-shot learning

Zero-shot learning uses single-prompt instruction to train LLMs for specific NLP tasks, directly applying previously trained models to predict both seen and unseen classes without using any labelled training instances [14]. Zero-shot learning has achieved impressive performance in a variety of NLP tasks, such as summarisation, dialogue generation, and question-answering [8]. Ge et al. use zero-shot learning to extract six data elements from patients' abdominal imaging reports using an API implementation of the OpenAI GPT-3.5 turbo LLM, achieving an overall high accuracy of 88.9%. They find that the level of accuracy of zero-shot learning reduces with more complex use cases. Their findings prove the feasibility of using general-purpose LLMs to extract structured information from clinical data with minimal technical expertise.

1.2 Few-shot learning

Few-shot learning, also coined as in-context learning, refers to the ability of LLMs to perform tasks guided by a small set of representative examples provided in the prompt [15, 16]. These in-context examples not only teach the LLM the mapping from inputs to outputs but also activate the LLM's parametric knowledge [17]. Only requiring a handful of labelled training examples is a clear advantage of few-shot learning, making it data-efficient and accessible to knowledge domain users without expertise in machine learning [15]. Few-shot learning is particularly useful in situations where annotating text data is not convenient or expensive. By providing just a few examples, domain experts can quickly create a generative AI system for a new task. Importantly, few-shot learning does not change the underlying model weights [18]. This allows for efficient adaptation to new tasks without risking the loss of previously learned information. However, the performance of few-shot learning varies and is highly task-dependent [15]. Its accuracy is also sensitive to the choice of prompt templates and in-context examples [17]. Prior research finds that using semantically similar in-context examples to those with prior success can significantly enhance the performance of few-shot learning [19].

1.3 Parameter-efficient fine-tuning

Fine-tuning involves modifying the LLM, or the parameters used to train the LLM, to improve model response to the same prompt [13]. Fine-tuning changes a model's weight, thus, the model's behaviour to perform better at a specific task. Full fine-tuning will fine-tune all layers of the pre-trained model, which can be computationally expensive and may lead to catastrophic forgetting, i.e., the model forgets the knowledge it gained during pre-training. Thus, it may significantly increase the cost of computational resources and computational skill sets. Parameter-efficient fine-tuning only fine-tunes a small number of (extra) parameters while freezing most parameters of the pre-trained LLMs. It thus overcomes the computational resource constraint and catastrophic forgetting observed in the full-scope fine-tuning of LLM.

Low-Rank Adaptation (LoRA) is a PEFT technique designed to improve training efficiency for LLMs. It freezes the weight of per-trained LLMs and inserts low-rank decomposition matrices into the transformer layers. Previous research has demonstrated that LoRA can allow the fine-tuning process to focus on crucial parameters specific to the target task or domain, thus optimising the model's performance without extensive resource requirements or overfitting concerns. By focusing on PEFT, LoRA minimises the dependency on extensive labelled data for model optimisation, which maximises the utility of available data, making the fine-tuning process more effective and feasible in scenarios with limited annotated datasets [20].

Extracting symptoms of various geriatric diseases is important for early diagnosis, personalised treatment, and improving patient outcomes. To date, there is no reporting of effective tools to execute this multi-label classification task accurately and reliably from free-text notes in an EHR system. Therefore, this study focuses on a comparative analysis of the performance of prompt engineering with and without PEFT in multi-label classification. In this study, we include four clinical tasks with careful consideration of the following factors: (1) the information is recorded in the free text nursing progress notes; (2) the information meets aged care information needs; and (3) the research team has curated labelled datasets to allow model training, validation, and testing to evaluate the performance of the machine. We identified four clinical tasks: agitation in dementia, depression in dementia, frailty index and malnutrition risk factors (see Table 1). Each task has various numbers of labels, ranging from 13 to 83.

Table 1: Clinical tasks for multi-label classification.

Clinical Tasks	Label Names	Number of Labels
Agitation in dementia	Disruptive vocalisation, verbally aggressive behaviour, arguing, complaining, cursing, threat, using abusive language, using accusatory language, using foul language, using hostile language, using obscene language, using profane language,	83

	verbally nonaggressive behaviour, ceaseless talking, constant repetition of word, constant unwarranted requests for attention, constant unwarranted requests for help, constant unwarranted requests for reassurance, echolalia, groaning, grunting, howling, making bizarre noise, rambling, repetitive questioning, roaring, screaming, shouting, speaking in excessively loud voice, emotional distress, anger, frustration, irritability, mood swing, negativism, outburst, physically aggressive behaviour, biting, destroying property, fighting, grabbing, hitting, hurting self, hurting someone, kicking, pushing, resisting, scratching, shoving, slamming, spitting on people, staring, striking people, tearing, throwing object, physically nonaggressive behaviour, constant manipulation of object, fidgeting, gesturing, hand wringing, inappropriate dressing, inappropriate handling object, inappropriate undressing, pacing, pointing finger, repetitive physical mannerism, restlessness, rocking, rummaging, searching, wandering, bruxism, resisting, punching, absconding, calling out, physical agitation, facial grimacing, moving furniture, hoard items, intrusive of others privacy, gets up and down from constantly, urinating on the floor	
Depression in dementia	Diminished ability of thinking, feeling discouraged, feeling empty, feeling hopeless, feeling of excessive guilt, feeling sad, feeling worthless loss of energy, loss of interest, loss of pleasure, suicidal ideation, suicide, suicide attempt, tearfulness	13
Frailty index	Activity limitation, anaemia and haematinic deficiency, arthritis, atrial fibrillation, cerebrovascular disease, chronic kidney disease, diabetes, dizziness, dyspnoea, falls, foot problems, fragility fracture, hearing impairment, heart failure, heart valve disease, housebound, hypertension, hypotension/syncope, ischaemic heart disease, memory and cognitive impairment, mobility and transfer problems, osteoporosis, Parkinsonism	36

	and tremor, peptic ulcer, peripheral vascular disease, polypharmacy, requirement for care, respiratory disease, skin ulcer, sleep disturbance, social vulnerability, thyroid disease, urinary incontinence, urinary system disease, visual impairment, weight loss and anorexia	
Risk factors for malnutrition	Anxiety, bowel blockage, cancer, chest infection, chronic wound, confusion, constipation, delirium, dementia, depression, diabetes, diarrhoea, difficulty swallow, dysphagia, eating disorder, food preference, frailty, gastritis, heart disease, HIV, hospital admission, isolation, kidney disease, liver disease, malabsorption medication, nausea, Parkinson, pneumonia, poor appetite, poor intake, poor oral health, pressure ulcer, sepsis, stroke, suboptimal intake, surgery, vomiting	37

There is a lack of prior research on the difference in performance between zero-shot and few-shot learning for the same clinical classification task and on the effect of PEFT on the tasks. As healthcare demands high safety standards for machine learning, it is imperative to conduct experimental comparisons of the performances of various machine learning methods. In this research, we focus on the comparison of zero-shot and few-shot learning, with and without PEFT, on multi-label clinical classification tasks. We design experiments to test the research hypotheses (see Table 2).

Table 2: Research hypotheses in the study.

Hypothesis 1	Zero-short and few-shot learning with similar prompting templates have different levels of performance when applied to multi-label classification for different clinical tasks
Hypothesis 2	Few-shot learning performs better than zero-shot learning for the multi-label classification of the same clinical task without PEFT.
Hypothesis 3	Parameter-efficient fine-tuning can improve both zero-shot and few-shot learning performance.
Hypothesis 4	Zero-shot learning reaches the same level of performance as few-shot learning for the same clinical task after PEFT.
Hypothesis 5	Fine-tuning for one clinical task impacts model performance across other clinical tasks.

2 Methodology

We conduct the experiment in seven stages: generative AI-based large language model selection, data set selection, data preprocessing, designing prompting templates for zero-shot and few-short learning in each clinical task, machine learning methods execution, model performance evaluation and statistical analysis.

2.1 Ethics approval

The Human Research Ethics Committee of the University of Wollongong approved the study (Ethics Number 2019/159).

2.2 Generative AI-based LLM selection

We select the Llama 2-Chat 13B-parameter model as the generative AI-based LLM. The selection considers the following factors: (1) the optimal model in terms of open source and favourable review at the time of the experiment; (2) practical considerations regarding the availability of GPU resources; (3) feasibility for local server deployment, convenience and control over usage; (4) compliance with health data privacy regulations in Australia; (4) the presence of diverse variants spawned through fine-tuning, including Alpaca, Baizem, Koala, and Vicuna [5, 21]. We obtain the Llama 2-Chat 13B-parameter model from the Hugging Face repository (https://huggingface.co/meta-llama/Llama-2-7b-chat-hf).

2.3 Data set selection

De-identified demographic data and free-text nursing progress notes are collected for the same population of older people living in 40 RACFs in New South Wales, Australia, for the period of 2019 to 2021. Residential aged care facilities are the equivalent of long-term care facilities in the USA. The dataset encompasses over 890,000 records of 3,528 de-identified individuals. The structured demographic information includes masked sequence number for client de-identification, age, and gender. The unstructured nursing notes include nursing assessment and progress reporting. They document clients' daily activities, care staff's clinical observations, assessments of client's care needs (including risk factors), and carer interventions.

2.4 Data preprocessing

Text preprocessing involves the removal of URLs and non-textual characters, such as extra delimiters and empty spaces in the dataset. We make a choice not to exclude stop words because many of them, like "a," "be," "very," "should," etc., held semantic relevance to the content [22].

2.5 Designing prompting templates for zero-shot and few-short learning in each clinical task

First, we select prompt-based training via zero-shot and few-shot learning. We adopt the template developed by Abdallaha et al. [23] to construct our prompt (see Figure 1).

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:

{ instruction }

Input:

{input}

Response: { output }

Figure 1: Prompt template adapted from Abdallaha et al. [23]

The final prompts use in our experiment are listed in Table 3. The example results generated from the final prompts are showcased in Supplementary Table 1.

Table 3: Prompts used in the study.

Prompt Learning Technique	Domain	Prompt
Zero-shot	Agitation in dementia	As a nursing expert, you are tasked with reviewing a nursing progress note for a resident with dementia residing in a Residential Aged Care environment. The note may contain one or more symptoms indicative of agitation in dementia, including but not limited to resisting, wandering, speaking in an excessively loud voice, pacing, restlessness, pushing,

		shouting, complaining, frustration, using profane language, screaming, gesturing, threatening, grabbing, absconding, using abusive language, arguing, punching, spitting, expressing anger, groaning, tearing, hitting, and more. Please follow these steps. 1. Identify any symptoms of agitation. 2. If symptoms of agitation are evident, please list them.
Few-shot	Agitation in dementia	As a nursing expert, you are tasked with reviewing a nursing progress note for a resident with dementia residing in a Residential Aged Care environment. The note may contain one or more symptoms indicative of agitation in dementia, including but not limited to resisting, wandering, speaking in an excessively loud voice, pacing, restlessness, pushing, shouting, complaining, frustration, using profane language, screaming, gesturing, threatening, grabbing, absconding, using abusive language, arguing, punching, spitting, expressing anger, groaning, tearing, hitting, and more. Please follow these steps. 1. Identify any symptoms of agitation. 2. If symptoms of agitation are evident, please list them. Example 1: You have identified that the resident exhibits agitation symptoms, including physical agitation/aggression and verbal behaviours, including verbal disruption, calling out, and screaming, as documented in the note below. “Jenny, who suffers from vascular dementia, displays confusion, disorientation, and notable physical agitation/aggression and verbal behaviours. She exhibits verbal disruption, calling out and screaming, which disturbs others. Jenny frequently requires staff assistance for reassurance, comfort, and distraction using strategies such as music therapy, playing cards, or engaging in simple puzzles. Additionally, when highly distressed, staff contact her daughter to speak with her for comforting purposes. Although she can follow simple instructions, at times, she experiences considerable distress.” Example 2: You have identified that the resident exhibits agitation symptoms, including wandering, frightening others, refusing care, and arguing, as documented in the note below. "Elli faces challenges with poor balance and is at risk of falls due to impulsivity and reduced balance. Staff supervise Elli's transfers and mobility using a 4-wheelie walker. As Elli's dementia progresses, she tends to wander into other residents' rooms, mistakenly believing they occupy her bed. This behaviour frightens other residents, resulting in arguments. Staff frequently intervene, redirect Elli, and provide extra reassurance and diversion throughout the day. As her dementia advances, Elli lacks insight into her care needs. She adamantly refuses staff assistance changing her continence aids and attending to her hygiene. Due to this progression, staff guide her through mealtimes, set the table, provide cutlery, and supervise and encourage her during meals and drinks.”

Zero-shot	Frailty index	As a nursing expert, you are tasked with reviewing a nursing progress note for a resident residing in a Residential Aged Care environment. The note may contain one or more frailty index, including but not limited to activity limitation, anaemia and haematinic deficiency, arthritis, atrial fibrillation, cerebrovascular disease, chronic kidney disease, diabetes, dizziness, dyspnoea, falls, foot problems, fragility fracture, hearing impairment, heart failure, heart valve disease, housebound, hypertension, hypotension/syncope, ischaemic heart disease, memory and cognitive impairment, mobility and transfer problems, osteoporosis, Parkinsonism and tremor, peptic ulcer, peripheral vascular disease, polypharmacy, the requirement for care, respiratory disease, skin ulcer, sleep disturbance, social vulnerability, thyroid disease, urinary incontinence, urinary system disease, visual impairment, weight loss and anorexia and more. Please follow these steps. 1. Identify any frailty index. 2. If the frailty index is evident, please list them along with the corresponding evidence in the note.
Few-shot	Frailty index	As a nursing expert, you are tasked with reviewing a nursing progress note for a resident residing in a Residential Aged Care environment. The note may contain one or more frailty index, including but not limited to activity limitation, anaemia and haematinic deficiency, arthritis, atrial fibrillation, cerebrovascular disease, chronic kidney disease, diabetes, dizziness, dyspnoea, falls, foot problems, fragility fracture, hearing impairment, heart failure, heart valve disease, housebound, hypertension, hypotension/syncope, ischaemic heart disease, memory and cognitive impairment, mobility and transfer problems, osteoporosis, Parkinsonism and tremor, peptic ulcer, peripheral vascular disease, polypharmacy, the requirement for care, respiratory disease, skin ulcer, sleep disturbance, social vulnerability, thyroid disease, urinary incontinence, urinary system disease, visual impairment, weight loss and anorexia and more. Please follow these steps. 1. Identify any frailty index. 2. If the frailty index is evident, please list them along with the corresponding evidence in the note. Example 1: You have identified that the resident exhibits a frailty index, including mobility and transfer problems, based on the evidence that the resident mobilises with a wheelie walker, as documented in the note below. "Ethan utilises a wheelie walker for mobility, and staff members provide supervision during his use of a chair lift. All of Ethan's meals take place in the dining room, with staff members responsible for pouring his drinks. Despite these assistance needs, Skeet maintains a healthy appetite." Example 2: You have identified that the resident exhibits a frailty index, including a skin ulcer, based on the evidence that the resident's wound shows an unchanged state with a small amount of exudate in the same cavity, as documented in the note below.

		" Wound location: Left foot. Date: 04/01/2020. Evaluation: The wound has remained unchanged since 02/01/2020, with the same cavity showing small exudate. The resident reports no pain except when the dressing is attended to. Scheduled for a specialist review next week."
Zero-shot	Depression in dementia	As a nursing expert, you are tasked with reviewing a nursing progress note for a resident with dementia residing in a Residential Aged Care environment. The note may contain one or more symptoms of depression, including but not limited to diminished ability to think, downcast gaze, dysphoria, feeling discouraged, feeling empty, feeling hopeless, feeling of excessive guilt, feeling sad, feeling worthless, loss of energy, loss of interest, loss of libido, loss of pleasure, loss of self-esteem, suicide, suicide attempt, tearfulness and more. Please follow these steps. 1. Identify any symptoms of depression. 2. If symptoms of depression are evident, please list them.
Few-shot	Depression in dementia	As a nursing expert, you are tasked with reviewing a nursing progress note for a resident with dementia residing in a Residential Aged Care environment. The note may contain one or more symptoms of depression, including but not limited to diminished ability to think, downcast gaze, dysphoria, feeling discouraged, feeling empty, feeling hopeless, feeling of excessive guilt, feeling sad, feeling worthless, loss of energy, loss of interest, loss of libido, loss of pleasure, loss of self-esteem, suicide, suicide attempt, tearfulness and more. Please follow these steps. 1. Identify any symptoms of depression. 2. If symptoms of depression are evident, please list them. Example 1: You have identified that the resident exhibits depression symptoms, including apathy and refusal of hygiene care, as documented in the note below. "Due to Peter's depression, he is apathetic and refuses hygiene care. Staff need to reapproach Peter multiple times per day and spend extra time with him, providing encouragement and reassurance and explaining the importance of attending to hygiene care. He enjoys mass and music." Example 2: You have identified that the resident exhibits a depression symptom, including a lack of motivation, as documented in the note below. "Due to John's depression, John does not want to mobilise using mobility aids. He does not want to use a 4-wheelie walker or walking stick. Staff need to help get him up from bed and help sit him in a wheelchair for meals or outings after persuasion and encouragement."
Zero-shot	Malnutrition risk factors	As a nursing expert, you are tasked with reviewing a nursing progress note for a resident residing in a Residential Aged Care environment. The note may contain one or more malnutrition risk factors, including but not limited to

		confusion, nausea, surgery, dementia, hospital admission, medication, stroke, poor intake, constipation, anxiety, depression, dysphagia, delirium, vomiting, gastroesophageal reflux disease, gastritis, bowel blockage, malabsorption, diarrhoea, chest infection, chronic obstructive pulmonary disease, upper respiratory tract infection, pneumonia, heart disease, cancer, sepsis, HIV, urinary tract infection, kidney disease, diabetes, liver disease, Parkinson disease, eating disorders, chronic wound, bedsores, poor oral health, difficulty chewing, unfit denture and more. Please follow these steps. 1. Identify any risk factors of malnutrition. 2. If risk factors of malnutrition are evident, please list them.
Few-shot	Malnutrition risk factors	As a nursing expert, you are tasked with reviewing a nursing progress note for a resident residing in a Residential Aged Care environment. The note may contain one or more malnutrition risk factors, including but not limited to confusion, nausea, surgery, dementia, hospital admission, medication, stroke, poor intake, constipation, anxiety, depression, dysphagia, delirium, vomiting, gastroesophageal reflux disease, gastritis, bowel blockage, malabsorption, diarrhoea, chest infection, chronic obstructive pulmonary disease, upper respiratory tract infection, pneumonia, heart disease, cancer, sepsis, HIV, urinary tract infection, kidney disease, diabetes, liver disease, Parkinson disease, eating disorders, chronic wound, bedsores, poor oral health, difficulty chewing, unfit denture and more. Please follow these steps. 1. Identify any risk factors of malnutrition. 2. If risk factors of malnutrition are evident, please list them. Example 1: You have identified that the resident exhibits a malnutrition risk factor, including confusion, as documented in the note below. " John requires comprehensive assistance with his hygiene and toileting needs due to his confusion. He experiences incontinence of both urine and feces, necessitating the use of pads around the clock. A nurse is responsible for assisting him with toileting, changing his pads, cleansing his groin area, and applying barrier cream to mitigate the risk of skin issues or breakdown." Example 2: You have identified that the resident exhibits a malnutrition risk factor, including constipation, as documented in the note below. " Peter requires the assistance of a nurse for his hygiene and toileting needs, primarily due to his unsteady gait. He utilises pads due to incontinence issues. Additionally, he is prone to constipation and receives aperients as necessary. The current intervention measures in place have proven to be effective."

2.6 Machine learning methods execution

We select prompt-based learning on Llama 2 and Llama 2 with PEFT to test the LLM's ability to adapt, generalise, and optimise performance in clinical multi-domain classification tasks.

2.6.1 Experimental setup

Prompt-based learning with zero-shot and few-shot learning on Llama 2

The experiment is conducted on two NVIDIA GeForce GTX -1080 Ti, each equipped with 11GB of memory. Our software environment is Ubuntu 18.04, the programming language is Python 3.10.0, and the deep learning framework is Pytorch 2.0.0. We employ Llama 2 without PEFT, utilising zero-shot and few-shot learning prompts as outlined in Table 3. To prevent model contamination, we approach each clinical task (see Table 1) in two distinct steps. Initially, we employ the Llama 2 model, which is directly downloaded from the Hugging Face repository, for zero-shot learning. Afterwards, we downloaded a second copy of the same model from the same repository to conduct few-shot learning. We have performed multiple iterations of zero-shot learning and few-shot learning with Llama 2 in order to test the research hypotheses outlined in Table 2, using a test dataset of 100 nursing notes in each clinical task. The maximum token limit employed in the Lama 2 is 4096, as none of the test notes available exceeds this token count. The model iteratively processes each note within this defined token limit during testing.

Parameter Efficient Fine-tuning with LoRA on Llama 2

We use the PEFT method to fine-tune the Llama 2 model. The experiment is conducted on four NVIDIA GeForce GTX -1080 Ti, each equipped with 11GB of memory, and employing the hyperparameter setting shown in Table 4. Our software environment is Ubuntu 18.04, the programming language is Python 3.10.0, and the deep learning framework is Pytorch 2.0.0. The instruction data points are employed in the PEFT process.

Table 4: Hyperparameters used in PEFT

Settings	Parameters
Batch size	128
Micro batch size	4
LoRA rank	8
LoRA alpaca	16
LoRA dropout	0.05
Learning rate	3e-4
Training steps	300
Optimizer	AdamW
Trainable parameters (%)	<0.01%

The maximum token limit is set as 4096, the maximum number token that can be taken by the Llama 2 model, which is large enough to encompass the available token for each nursing note. During the fine-tuning process, the model iteratively processes each note within the defined token limit. We randomly divide the labelled data presented in Table 5 in each clinical task into 90% training and 10% validation data sets, respectively. First, we apply PEFT to the training data, i.e., free-text nursing notes for each clinical task (see Table 5). We ensure that no overlapping free text notes in the labelled dataset are used for different labelled clinical tasks in the PEFT processes.

Table 5: Number of labelled data and file size for each clinical task

Clinical Tasks	Training + Validation Data	File Size	Output model
Agitation in dementia	3000 nursing notes	5.89 MB	Agitation in dementia with specialised PEFT of Lama 2
Depression in dementia	700 nursing notes	280 KB	Depression in dementia with specialised PEFT of Lama 2
Frailty index	949 nursing notes	154 KB	Frailty index with specialised PEFT of Lama 2
Malnutrition risk factors	2850 nursing notes	972 KB	Malnutrition risk factors with specialised PEFT of Lama 2

Subsequently, separate test datasets of 100 nursing notes are employed for model performance evaluation for each clinical task in order to test the research hypotheses outlined in Table 2. This dedicated test dataset is explicitly employed for assessing the prompt-based learning method of Llama 2 without PEFT, facilitating a comprehensive comparison and analysis. We use the prompts delineated in Table 3 to evaluate the test data. First, we conducted zero-shot learning with PEFT Llama 2 across the four test datasets for each clinical task. This is followed by few-shot learning with PEFT Llama 2 across the same four test datasets for each clinical task. To ensure that the few-shot learning does not benefit from the residual effect of the previous zero-shot learning in the test process, we download the original Llama 2 model from the Hugging Face repository for training in the fine-tuning process.

2.7 Model performance evaluation

We calculated accuracy, precision, recall and F1score to assess each model's performance for the four clinical tasks. The annotated ground truth is curated by the large research team. Each annotation is independently corroborated by at least two, and sometimes three, domain experts. We compared the machine learning output to the annotated ground truth, employing exact and semantic matching criteria. In our approach, an extracted entity or phrase that overlaps with the text and shares the exact meaning of the entity type or phrase as the annotated ground truth is considered a true positive response [2]. For example, suppose the original text in the annotated ground truth is "shouting" as an agitation in dementia, whereas the model extracted output is "shouting"; the model output is judged as true positive because it exactly matches the annotated ground truth. Since Llama 2 is a generative model capable of producing text that conveys semantic meaning, we also apply the semantic similarity matching to assess true positive results. For example, suppose the original text in the annotated ground truth is "reject meals" as an agitation in dementia, whereas the model extracted output is "refuse meals"; the model output is judged as true positive because its meaning matches the annotated ground truth. The words/phrases that are output from the model but not in the annotated ground truth are considered false positives, and the words/phrases that are in the annotated ground truth but not in the Llama 2 model output are considered false negatives [2]. The example results generated from the evaluation criteria are showcased in Supplementary Table 2.

2.8 Statistical analysis

As the accuracy, precision, recall and F1 score, serving as measurement indicators, are continuous variables and do not adhere to the assumption of normality of variance required for parametric tests, we utilise the non-parametric Kruskal-Wallis test for comparing results across three or more independent groups and the Mann-Whitney U test for comparing two independent groups to test the hypotheses, as suggested by the previous research [24, 25]. A significant difference is decided if the p-value is smaller than 0.05.

3 Results

3.1 Results of testing Hypothesis 1: Zero-short and few-shot learning with similar prompting templates have different levels of performance when applied to multi-label classification for different clinical tasks.

To evaluate Hypothesis 1, we undertake the following comparisons among the four clinical tasks: (1) the performance of zero-shot learning without PEFT; (2) the performance of zero-shot learning with PEFT; (3) the performance of few-shot learning without PEFT; and (4) the performance of few-shot learning with PEFT.

3.1.1 Comparing the performance of zero-shot learning without PEFT for four clinical tasks.

Figure 2 compares the evaluation results of zero-shot learning among the four clinical classification tasks without PEFT. There is no statistically significant difference in accuracy, precision, recall, and F1 score between these classification tasks when utilising zero-shot learning without PEFT ($p > 0.05$). However, there is a trend that the classification tasks related to agitation in dementia and malnutrition risk factors perform better than those related to frailty index and depression in dementia.

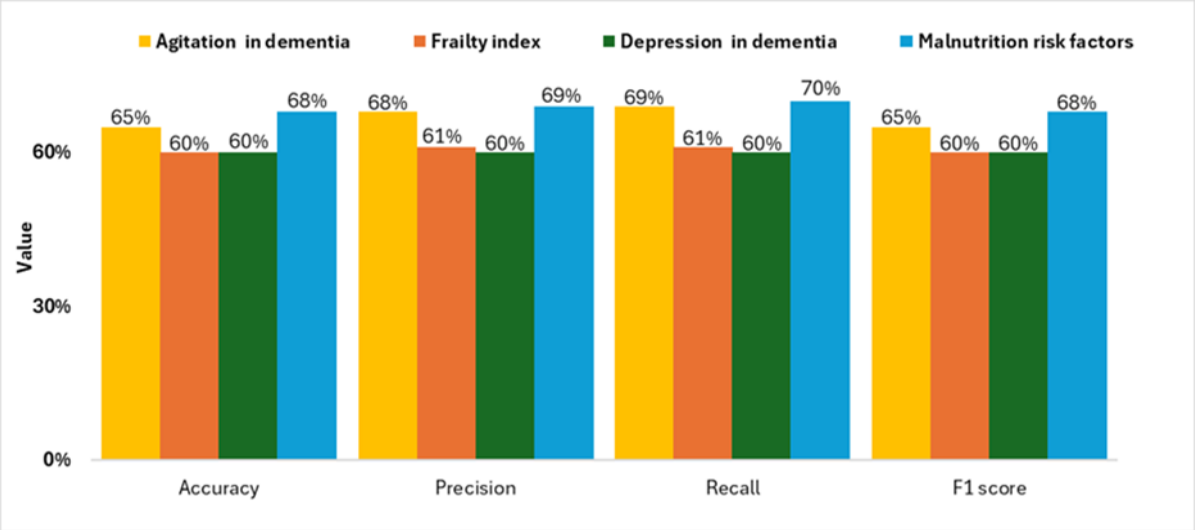

Figure 2: Comparative evaluation of zero-shot learning for clinical multi-label classification tasks without PEFT.

3.1.2 Comparing the performance of zero-shot learning with PEFT for four clinical tasks.

Figure 3 compares the evaluation results of zero-shot learning for the four clinical classification tasks with PEFT. Once again, no statistically significant difference is found in accuracy, precision, recall, and F1 score between these classification tasks when utilising zero-shot learning with PEFT ($p > 0.05$). However, there is a trend that the classification tasks related to agitation in dementia and malnutrition risk factors perform better than those related to frailty index and depression in dementia.

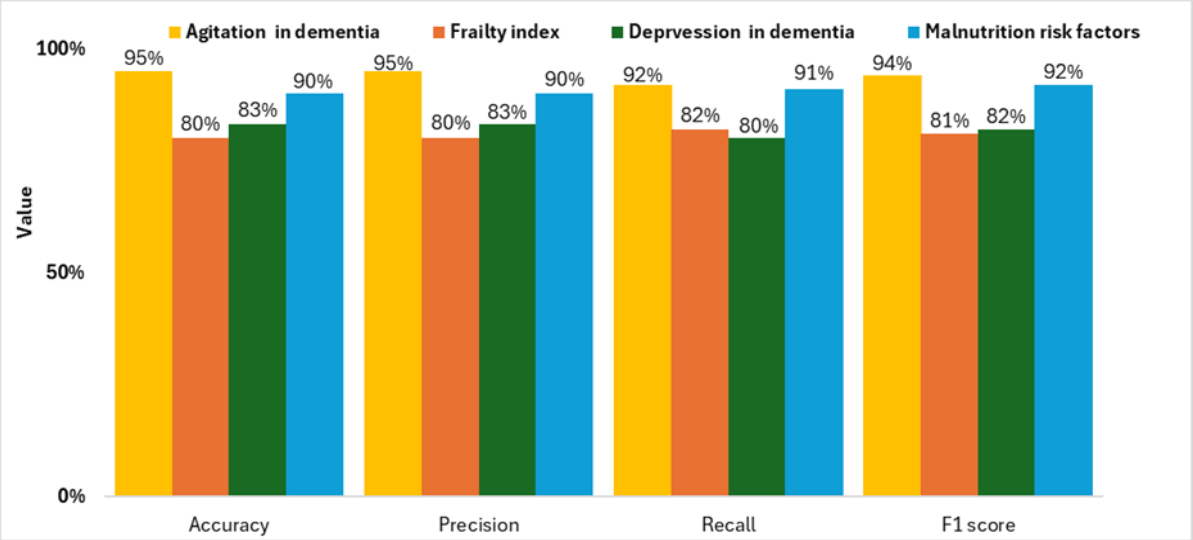

Figure 3: Comparative evaluation of zero-shot learning for the four clinical multi-label classification tasks with PEFT.

3.1.3 Comparing the performance of few-shot learning without PEFT for the four clinical tasks.

No statistically significant difference is found in accuracy, precision, recall, and F1 score between these tasks when utilising few-shot learning without PEFT (Figure 4, $p > 0.05$). However, the same trend as above was found, i.e., the classification tasks related to agitation in dementia and malnutrition risk factors perform better than those related to frailty index and depression in dementia.

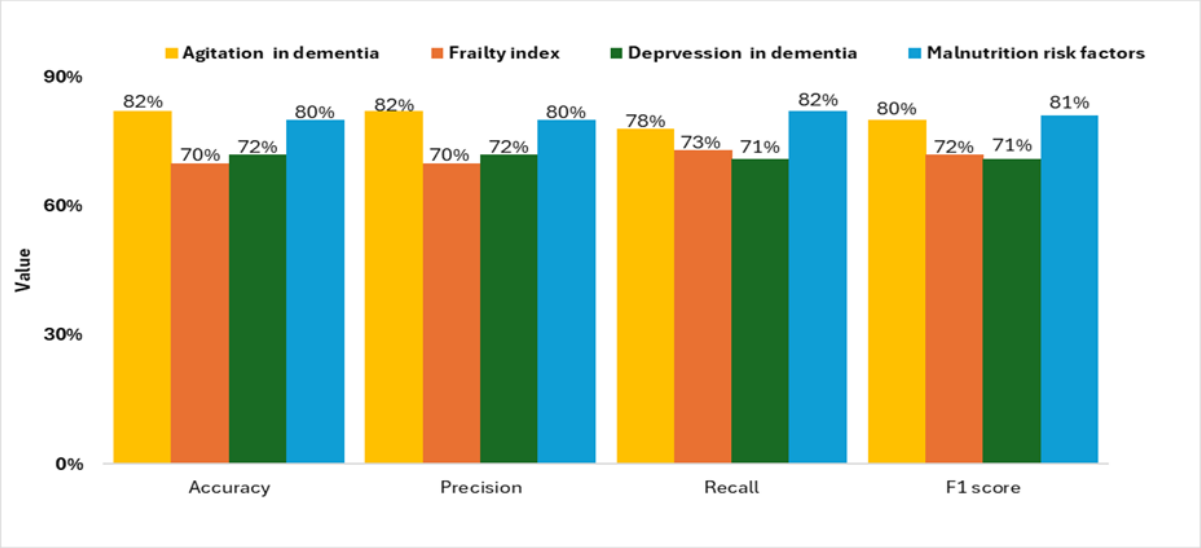

Figure 4: Comparative evaluation of few-shot learning for clinical multi-label classification tasks implemented without PEFT.

3.1.4 Comparing the performance of few-shot learning with PEFT for the four clinical tasks.

Again, no statistically significant difference is found in accuracy, precision, recall, and F1 score among the four clinical tasks (Figure 5, $p > 0.05$); however, the same trend as above is observed; i.e., the classification tasks related to agitation in dementia and malnutrition risk factors perform better than those related to frailty index and depression in dementia.

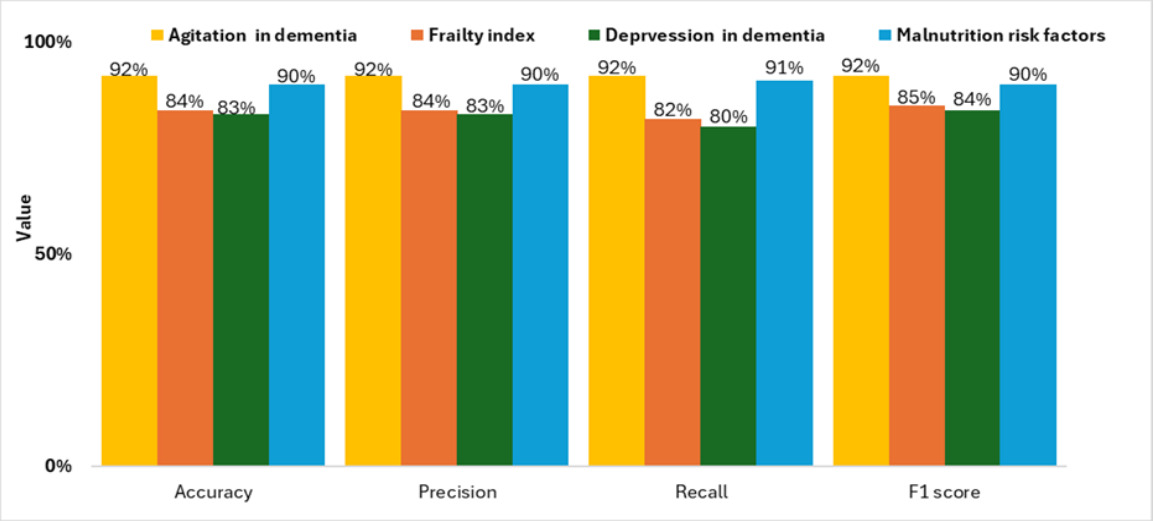

Figure 5: Comparative evaluation of few-shot learning for clinical multi-label classification tasks with PEFT.

3.2 Results of testing Hypothesis 2: Few-shot learning performs better than zero-shot learning for the multi-label classification of the same clinical tasks without PEFT.

To evaluate Hypothesis 2, we undertake the following comparison: the performance of zero-shot and few-shot learning without PEFT for all four clinical tasks.

Few-shot learning adaptation significantly improves model accuracy, precision, recall and F1 score in the multi-label clinical classification task than zero-shot learning (Figure 6, $p < 0.05$). The level of improvement is as follows: an 18% increase in model accuracy, an 18% increase in precision, a 25% increase in recall, and a 28% increase in F1 score.

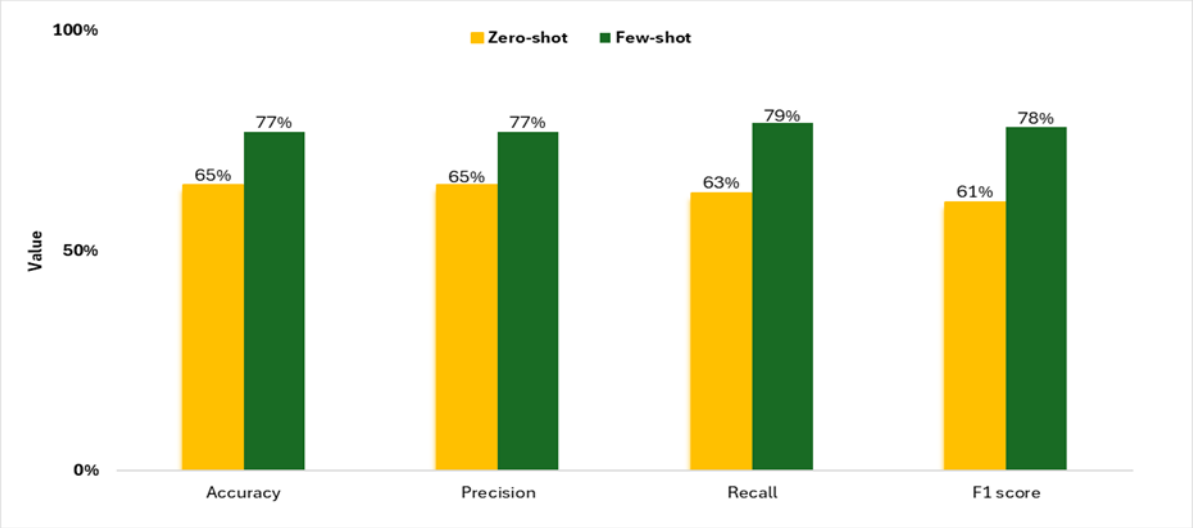

Figure 6. Comparative evaluation of zero-shot learning versus few-shot learning for all four clinical classification tasks without PEFT.

Few-shot learning adaptation significantly improves model accuracy in each multi-label clinical classification task (Figure 7, $p < 0.05$). The level of improvement ranges from 15% in agitation in dementia, 15% in malnutrition risk factors, 17% for frailty index, and the highest of 19% for depression in dementia. Few-shot learning adaptation significantly improves model precision in each multi-label clinical classification task ($p < 0.05$). The level of improvement ranges from 15% in agitation in dementia, 15% in malnutrition risk factors, 17% for frailty index, and the highest of 19% for depression in dementia. Few-shot learning adaptation significantly improves model recall in each multi-label clinical classification task ($p < 0.05$). The level of improvement ranges from 17% in malnutrition risk factors, 18% in agitation in dementia, and the highest of 20% for frailty index and depression in dementia. Finally, few-shot learning adaptation significantly improves model F1 score in each multi-label clinical classification task ($p < 0.05$). The level of improvement ranges from 17% in malnutrition risk factors, 17% in agitation in dementia, 18% for frailty index, and the highest of 20% for depression in dementia.

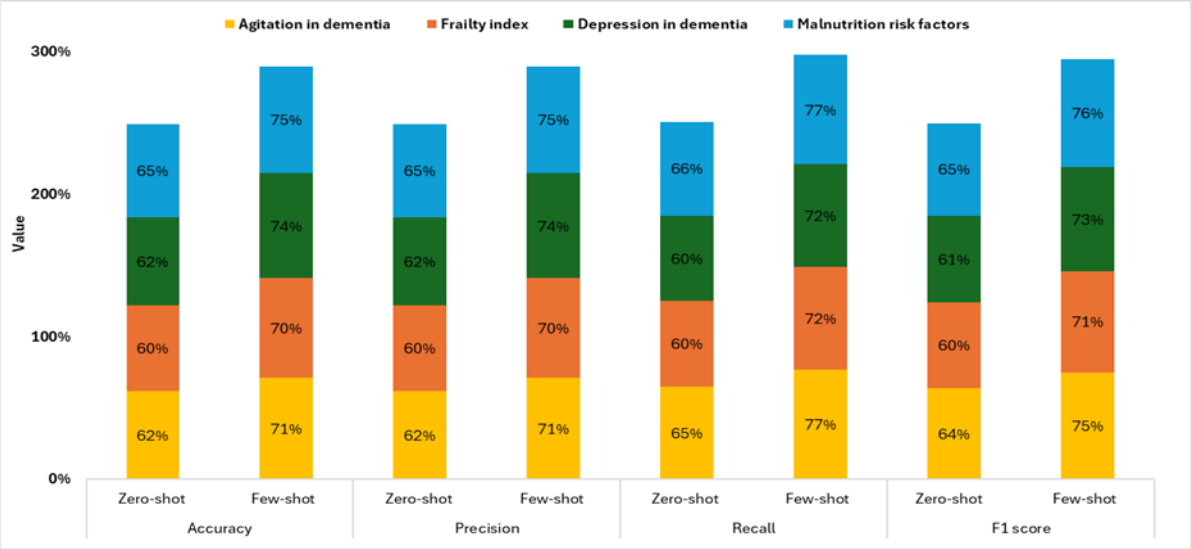

Figure 7. Comparative evaluation of zero-shot learning versus few-shot learning for each clinical classification task without PEFT.

3.3 Results of testing Hypothesis 3: Parameter-efficient finetuning can improve both zero-shot and few-shot learning performance.

To evaluate Hypothesis 3, we undertake the following comparisons: (1) the performance of the model of zero-shot learning without PEFT and with PEFT for all four tasks; (2) the performance of the model of few-shot learning without PEFT and with PEFT for all four tasks.

3.3.1 Comparing the performance of the model of zero-shot learning without PEFT and with PEFT for all four tasks.

Zero-shot learning with PEFT adaptation significantly improves model accuracy, precision, recall and F1 score in the multi-label clinical classification tasks (Figure 8, $p < 0.05$). The level of improvements is as follows: a 37% increase in model accuracy, an 37% increase in precision, a 35% increase in recall, and a 33% increase in F1 score.

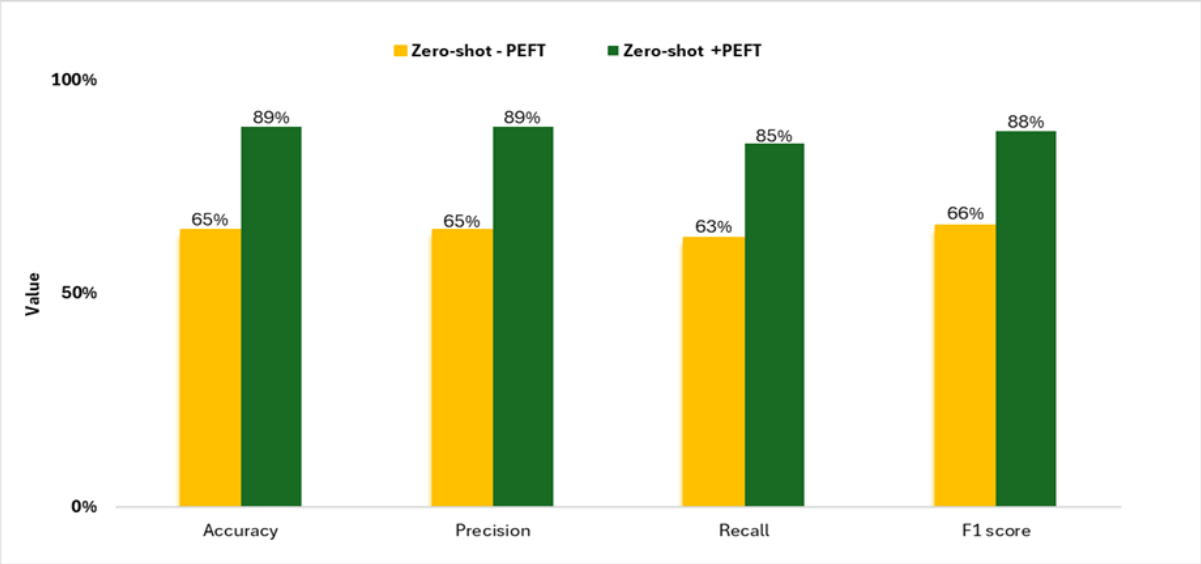

Figure 8: Comparative evaluation of zero-shot learning without PEFT versus with PEFT for all four clinical classification tasks. Note: '-PEFT' denotes without PEFT, '+PEFT' denotes with PEFT.

Zero-shot learning with PEFT adaptation significantly improves model accuracy in each multi-label clinical classification task (Figure 9, $p < 0.05$). The level of improvement ranges from 23% in frailty index, 29% in malnutrition risk factors, 30% for depression in dementia, and the highest of 38% for agitation in dementia. Zero-shot learning with PEFT adaptation significantly improves model precision in each multi-label clinical classification task ($p < 0.05$). The level of improvement ranges from 23% in frailty index, 29% in malnutrition risk factors, 30% for depression in dementia, and the highest of 38% for agitation in dementia. Zero-shot learning with PEFT adaptation significantly improves model recall in each multi-label clinical classification task ($p < 0.05$). The level of improvement ranges from 26% for depression in dementia, 29% for frailty index, 32% for malnutrition risk factors, and the highest of 36% for agitation in dementia. Zero-shot learning with PEFT adaptation significantly improves model F1 score in each multi-label clinical classification task ($p < 0.05$). The level of improvement ranges from 28% in depression in dementia, 31% in malnutrition risk factors, 33% for frailty index, and the highest of 36% for agitation in dementia.

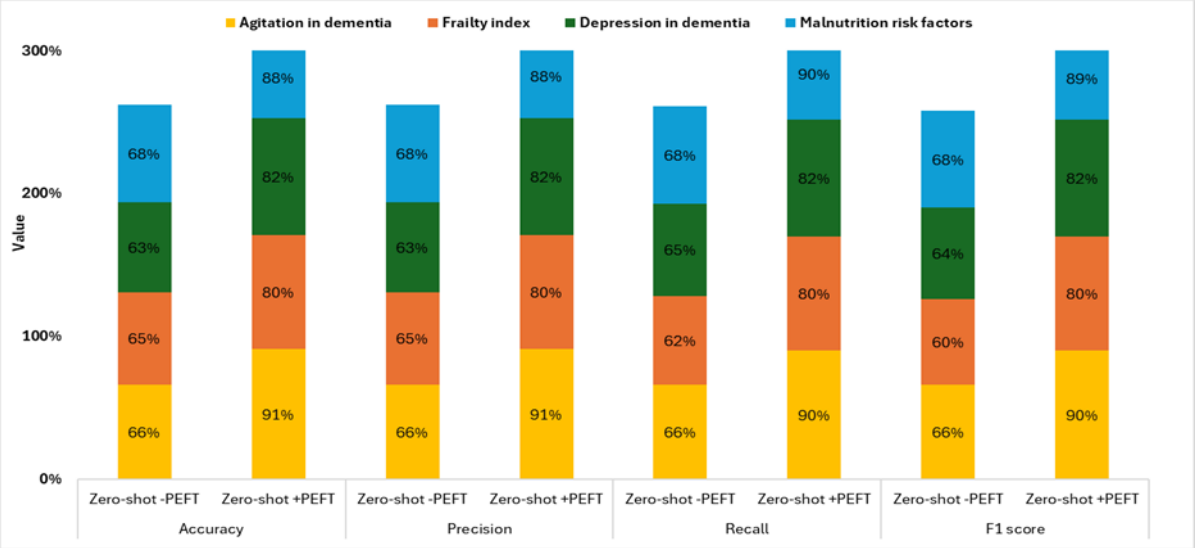

Figure 9: Comparative evaluation of zero-shot learning without PEFT versus zero-shot learning with PEFT for each clinical classification task. Note: '-PEFT' denotes without PEFT, '+PEFT' denotes with PEFT.

3.3.2 Comparing the performance of the model of few-shot learning without PEFT and few-shot learning with PEFT for all four multi-label clinical classification tasks.

Few-shot learning with PEFT adaptation significantly improves model accuracy, precision, recall and F1 score in all four multi-label clinical classification tasks (Figure 10, $p < 0.05$). The level of improvements is as follows: a 15% increase in model accuracy, a 15% increase in precision, a 23% increase in recall, and a 24% increase in F1 score.

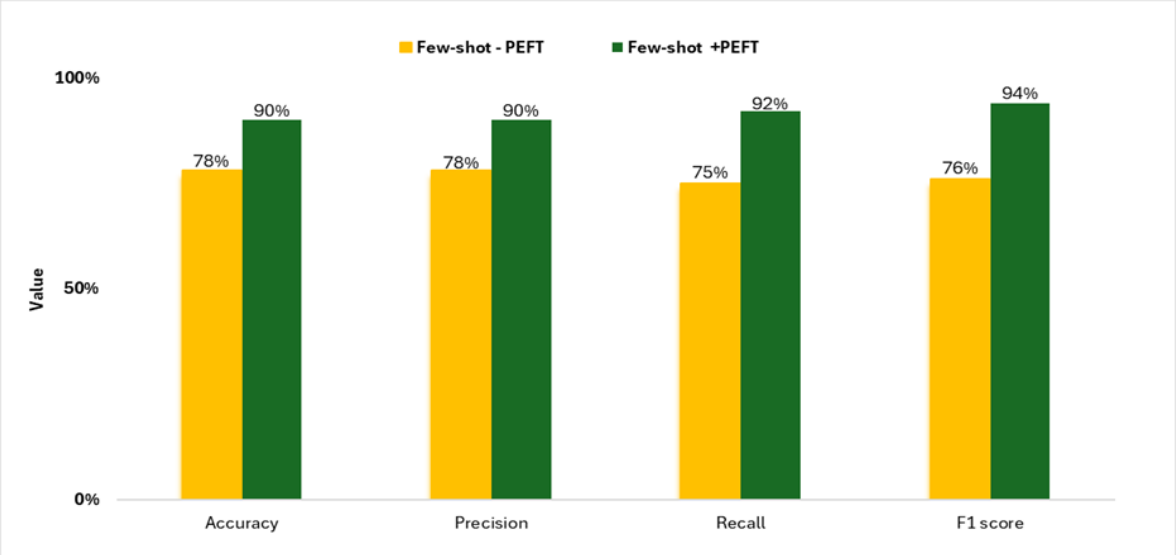

Figure 10: Comparative evaluation of few-shot learning without PEFT versus few-shot learning with PEFT for various clinical classification tasks. Note: '-PEFT' denotes without PEFT, '+PEFT' denotes with PEFT.

Few-shot learning with PEFT adaptation significantly improves model accuracy in each multi-label clinical classification task (Figure 11, $p < 0.05$). The level of improvement ranges from 9% in frailty index, 10% in malnutrition risk factors, the highest of 11% in agitation in dementia, and 11% for depression in dementia. Few-shot learning with PEFT adaptation significantly improves model precision in each multi-label clinical classification task ($p < 0.05$). The level of improvement ranges from 9% in frailty index, 10% in malnutrition risk factors, and the highest of 11% in agitation in dementia, and 11% for depression in dementia. Few-shot learning with PEFT adaptation significantly improves model recall in each multi-label clinical classification task ($p < 0.05$). The level of improvement ranges from 13% in frailty index, 13% in malnutrition risk factors, 15% in agitation in

dementia, and the highest of 17% for depression in dementia. Few-shot learning with PEFT adaptation significantly improves model F1 score in each multi-label clinical classification task ($p < 0.05$). The level of improvement ranges from 13% for malnutrition risk factors, 15% for agitation in dementia, 15% for frailty index, and the highest of 20% for depression in dementia.

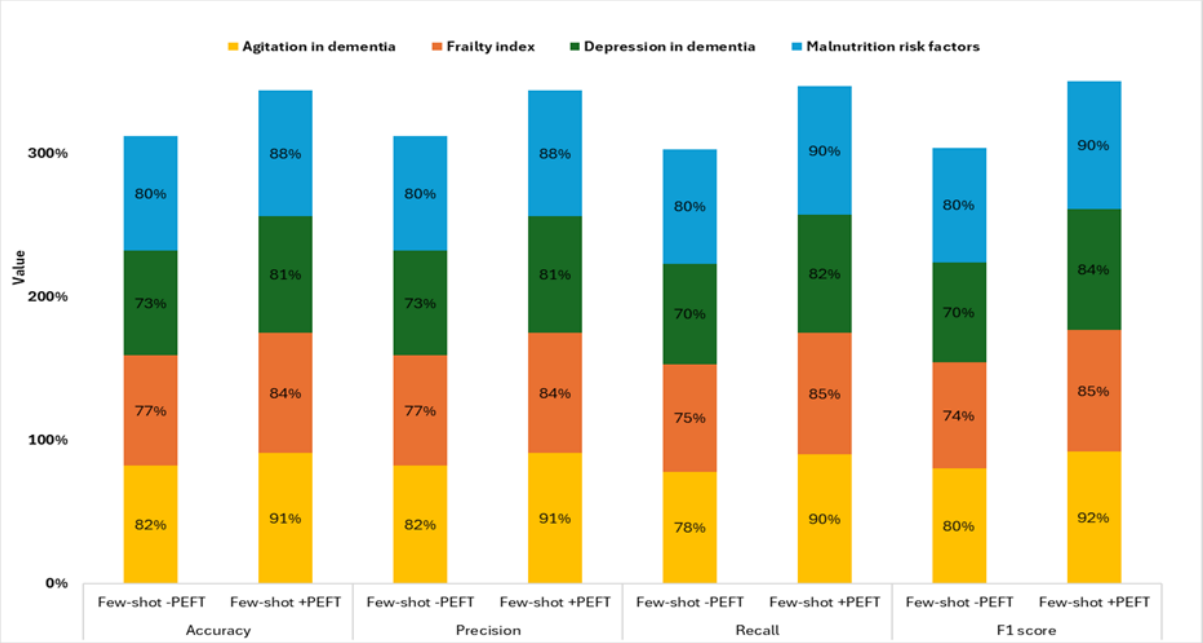

Figure 11: Comparative evaluation of few-shot learning without PEFT versus few-shot learning with PEFT for each clinical classification task. Note: '-PEFT' denotes without PEFT, '+PEFT' denotes with PEFT.

3.4 Results of testing Hypothesis 4: Zero-shot learning reaches the same level of performance as few-shot learning for the same clinical task with PEFT.

To evaluate Hypothesis 4, we undertake the following comparison: the performance of zero-shot and few-shot learning with PEFT for all four tasks.

Although no statistically significant difference is found in accuracy, precision, recall, and F1 score between the zero-shot and few-shot learning with PEFT in the multi-label clinical classification tasks (Figure 12, $p > 0.05$), there is a trend that few-shot learning performs above zero-shot learning.

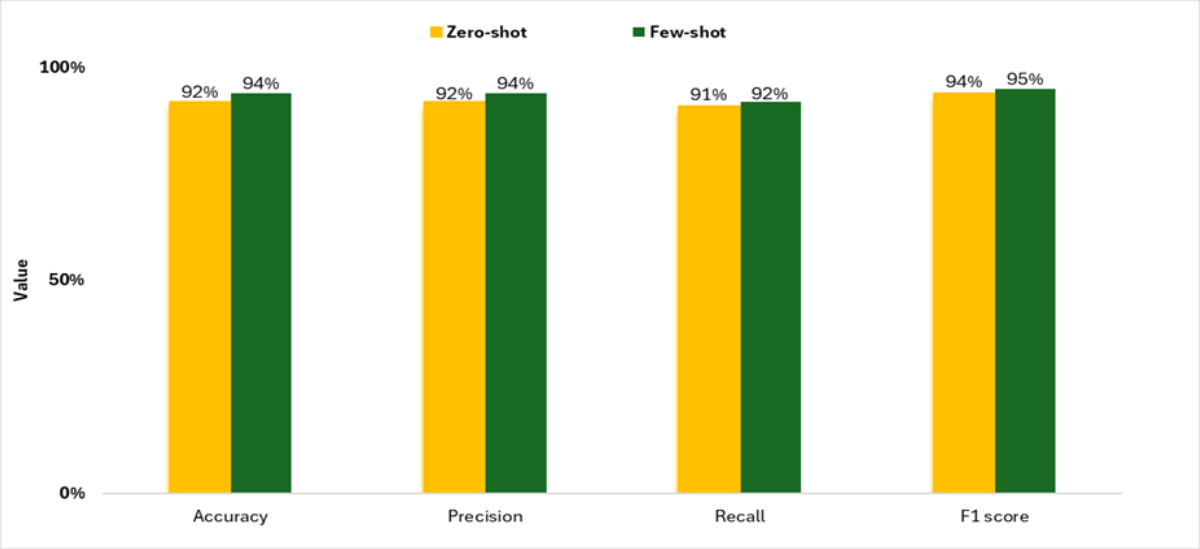

Figure 12: Comparative evaluation of zero-shot learning versus few-shot learning for various clinical classification tasks with PEFT. Note: '-PEFT' denotes without PEFT, '+PEFT' denotes with PEFT.

Overall, there is no statistically significant difference between zero-shot and few-shot learning for each clinical task (see Figure 13). The slight variation in values is task-specific, and there is no overall pattern.

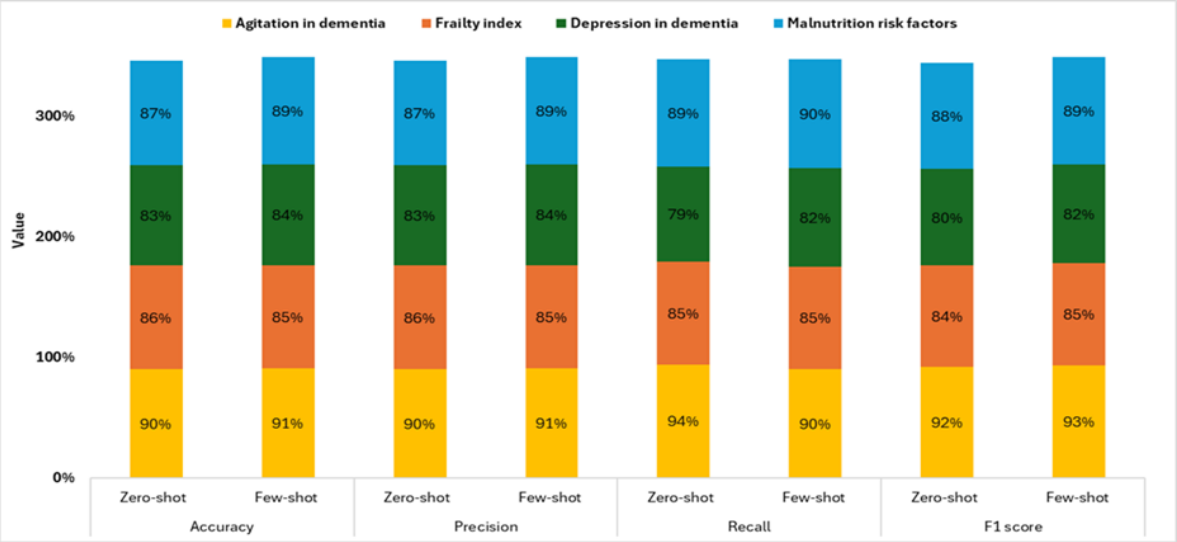

Figure 13: Comparative evaluation of zero-shot learning versus few-shot learning for each clinical classification task with PEFT. Note: '-PEFT' denotes without PEFT, '+PEFT' denotes with PEFT.

3.5 Results of testing Hypothesis 5: Fine-tuning for one clinical task impacts model performance across other clinical tasks.

To evaluate hypothesis 5, we undertake the following comparisons: (1) the performance of a clinical task-specialised PEFT model with zero-shot learning and its impact across other clinical tasks with zero-shot learning; (2) the performance of a clinical task-specialised PEFT model with few-shot learning and its impact on across other clinical tasks with few-shot learning.

3.5.1 Comparing the performance of a clinical task-specialised PEFT model with zero-shot learning and its impact on other clinical tasks trained via zero-shot learning.

No significant difference is found in accuracy, precision, recall, and F1 score for the other group of clinical tasks between the zero-shot learning with or without PEFT in one clinical task (see Table 6, $p > 0.05$).

Table 6: Comparing the impact of zero-shot learning on one clinical task with or without PEFT on other clinical tasks trained via zero-shot learning.

Clinical Task for PEFT	Training Method	Other Clinical Tasks	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Agitation in dementia	-PEFT	Frailty index, depression in dementia, malnutrition risk factors	60	60	60	60
	+PEFT		62	62	63	60
Frailty index	-PEFT	Agitation in dementia, depression in dementia, malnutrition risk factors	64	64	67	66
	+PEFT		62	62	64	60
Depression in dementia	-PEFT	Agitation in dementia, frailty index,	66	66	61	66
	+PEFT		66	66	64	66

		malnutrition risk factors				
Malnutrition risk factors	-PEFT	Agitation in dementia, depression in dementia, frailty index	63	63	65	66
	+PEFT		62	62	62	62

Note: '-PEFT' denotes without PEFT, '+PEFT' denotes with PEFT.

No significant difference is found in accuracy, precision, recall, and F1 score for each individual clinical tasks between the zero-shot learning with or without PEFT (Table 7, $p > 0.05$).

Table 7: Comparing the impact of zero-shot learning on one clinical task with or without PEFT on each other clinical task trained via zero-shot learning.

Clinical Task for PEFT	Training Method	Other Clinical Tasks	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Agitation in dementia	-PEFT	Frailty index	62	62	66	64
		Depression in dementia	67	67	62	66
		Malnutrition risk factors	66	66	67	62
	+PEFT	Frailty index	65	65	68	63
		Depression in dementia	64	64	66	62
		Malnutrition risk factors	68	68	64	64
Frailty index	-PEFT	Agitation in dementia	66	66	64	63
		Depression in dementia	63	63	61	61
		Malnutrition risk factors	64	64	61	65
	+PEFT	Agitation in dementia	65	65	67	64
		Depression in dementia	61	61	68	60
		Malnutrition risk factors	63	63	60	62
Depression in dementia	-PEFT	Agitation in dementia	63	63	64	61
		Frailty index	64	64	61	61
		Malnutrition risk factors	64	64	65	65
	+PEFT	Agitation in dementia	65	65	63	62
		Frailty index	61	61	60	60
		Malnutrition risk factors	62	62	66	60
Malnutrition risk factors	-PEFT	Agitation in dementia	65	65	68	64
		Frailty index	66	66	60	66
		Depression in dementia	65	65	64	63
	+PEFT	Agitation in dementia	67	67	64	66

		Frailty index	65	65	62	65
		Depression in dementia	66	66	61	65

Note: '-PEFT' denotes without PEFT, '+PEFT' denotes with PEFT.

3.5.2 Comparing a clinical task-specialised PEFT model performance with few-shot learning and its impact on other clinical tasks trained via few-shot learning.

No significant difference is found in accuracy, precision, recall, and F1 score for the group of other three clinical tasks between the few-shot learning with or without PEFT in one clinical task (see Table 8, $p > 0.05$).

Table 8: Comparing the impact of few-shot learning on one clinical task with or without PEFT on a group of other clinical tasks trained via few-shot learning.

Clinical Task for PEFT	Training Method	Other Clinical Tasks	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Agitation in dementia	- PEFT	Frailty index, depression in dementia, malnutrition risk factors	80	80	80	81
	+ PEFT		82	82	80	80
Frailty index	- PEFT	Agitation in dementia, depression in dementia, malnutrition risk factors	76	76	78	77
	+ PEFT		78	78	79	79
Depression in dementia	- PEFT	Agitation in dementia, frailty index, malnutrition risk factors	80	80	78	78
	+ PEFT		81	81	82	80
Malnutrition risk factors	- PEFT	Agitation in dementia, depression in dementia, frailty index	84	84	85	82
	+ PEFT		86	86	86	81

Note: '-PEFT' denotes without PEFT, '+PEFT' denotes with PEFT.

No significant difference is found in accuracy, precision, recall, and F1 score for the other individual clinical tasks between the few-shot learning with or without PEFT on one clinical task (Table 9, $p > 0.05$).

Table 9: Comparing the impact of few-shot learning on one clinical task with or without PEFT on each other clinical task trained via few-shot learning.

Clinical Task for PEFT	Training Method	Other Clinical Tasks	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
	- PEFT	Frailty index	78	78	77	81
		Depression in dementia	76	76	77	79
		Malnutrition risk factors	82	82	84	80
	+PEFT	Frailty index	80	80	78	84

Agitation in dementia		Depression in dementia	79	79	74	78
		Malnutrition risk factors	81	81	86	80
Frailty index	- PEFT	Agitation in dementia	84	84	82	86
		Depression in dementia	81	81	80	79
		Malnutrition risk factors	84	84	84	81
	+PEFT	Agitation in dementia	80	80	84	82
		Depression in dementia	80	80	78	78
		Malnutrition risk factors	82	82	83	80
Depression in dementia	- PEFT	Agitation in dementia	80	80	82	81
		Frailty index	84	84	78	80
		Malnutrition risk factors	80	80	85	80
	+PEFT	Agitation in dementia	81	81	83	80
		Frailty index	79	79	76	77
		Malnutrition risk factors	82	82	87	82
Malnutrition risk factors	- PEFT	Agitation in dementia	83	83	80	82
		Frailty index	79	79	74	76
		Depression in dementia	80	80	80	76
	+PEFT	Agitation in dementia	80	80	84	79
		Frailty index	77	77	76	75
		Depression in dementia	81	81	77	78

Note: '-PEFT' denotes without PEFT, '+PEFT' denotes with PEFT.

4 Discussion

This study explores the impact of zero-shot and few-shot prompt learning strategies, both with and without PEFT, on multi-label classification across four clinical tasks. These include agitation in dementia, depression in dementia, frailty index and malnutrition risk factors. To achieve this, five research hypotheses have been formulated, and experimental designs have been implemented to rigorously test these hypotheses.

Three of the five proposed hypotheses are confirmed and two rejected. Our findings do not support Hypothesis 1, which proposed that zero-short and few-shot learning with similar prompting templates have different levels of performance when applied to multi-label classification for the different clinical task. Our study consistently did not find statistically significant difference among the four clinical tasks. We also observe the pattern that two clinical tasks, agitation in dementia and malnutrition risk factor classification tasks, achieved a slightly higher performance in accuracy, precision, recall and F1 score than the other two clinical tasks, frailty index and depression in dementia classification tasks in all tests. This slight difference may be attributed to Llama 2 possessing more knowledge, with higher number of training data, for the two former clinical tasks than the two later clinical tasks.

Our finding supports Hypothesis 2, which proposes that few-shot learning performs better than zero-shot learning for the multi-label classification of the same clinical task without PEFT, few-shot domain adaptation effectively minimises false positives and false negatives while increasing true positives in classification tasks. These results imply that exposing a LLM to initial information from the target domain can significantly improve the model's performance in classification tasks.

Our finding supports Hypothesis 3, which posits that PEFT can improve both zero-shot and few-shot learning performance. The outcomes underscore the performance improvement with fine-tuning techniques. When implementing PEFT within a specified domain, it showcases the capability to mitigate false positives and false negatives while concurrently increasing true positives in information extraction tasks unique to that domain. This strategy facilitates more effective modifications to the model's parameters within the domain, thereby enhancing the model's overall performance in handling the multi-label classification tasks within that domain.

Our finding supports Hypothesis 4, proposing that zero-shot learning reaches the same level of performance as few-shot learning for the same clinical task after PEFT. Our study consistently indicates no performance difference when we employ the model with PEFT and zero-shot or few-shot learning. Due to the model's prior exposure to fine-tuning with the domain, the provision of additional exposure through few-shot learning does not significantly impact its performance.

Contrary to Hypothesis 5, proposing that fine-tuning for one clinical task impact model performance across other clinical tasks, our findings do not support this assertion. Instead, our study indicates that fine-tuning a specific task does not significantly hinder the model's performance when it is applied to another classification task. The rationale behind this lies in the methodology of PEFT, which focuses on training only a selective subset of the pre-trained model's parameters. This strategy entails identifying and updating only the most relevant parameters for the new task during training. This insight suggests that the model can be effectively tailored to a specific clinical task without compromising its effectiveness in handling diverse tasks. This adaptability underscores the potential of the PEFT approach within the LLM for various clinical tasks. Additionally, another factor for this result may be that the four clinical tasks are overall similar to each other in term of the nature.

This study encompasses three notable limitations. Firstly, our study encompasses four multi-label clinical classification tasks. However, we recognize that these tasks may not represent a diverse spectrum of clinical scenarios. To address this, we intend to broaden our scope by incorporating additional clinical classification tasks into our study. Secondly, the limitation pertains to the study's scope. Although we have examined four multi-label clinical classification tasks, we plan to expand our research by incorporating additional tasks, such as question answering, summarisation, and relation extraction in the future. This expansion would offer a more comprehensive view of the LLM's performance in EHR data. Thirdly, the limitation relates to the selection of evaluation metrics. This research solely utilises accuracy, precision, recall, and F1 score as the primary evaluation metrics. In future studies, we will broaden our evaluation process to encompass metrics like calibration, robustness, fairness, bias, toxicity, and efficiency [26]. Diversifying our evaluation criteria will provide a more comprehensive and nuanced assessment of the model's performance across various dimensions. This will enhance our findings' reliability and applicability in real-world EHR applications.

5 Conclusion

This study compares the performance of zero-shot and few-shot learning on multi-label clinical classification tasks and the impact of PEFT on their performance. Our findings indicate that the same prompting template (either zero-short or few-shot) has the same level of performance when it is applied to different multi-label classification tasks. Few-shot learning outperforms zero-shot learning without PEFT in classification tasks, while zero-shot learning achieve the same performance level as few-shot learning in classification tasks with PEFT. Furthermore, the study reveals the notably enhanced effectiveness of PEFT with both zero-shot and few-shot learning performance. Our analysis demonstrates that fine-tuning LLMs for a particular clinical task does not significantly compromise the model's performance when applied to other clinical tasks. These insights emphasise the adaptability and effectiveness of PEFT within the LLMs for various clinical tasks.

Acknowledgements: This research is a part of a PhD project supported by the University of Wollongong, Australia and the University Grant Commission, Sri Lanka, and the authors would like to acknowledge and thank them.

Author Approval

All authors have reviewed and approved the final version of this manuscript.

Author Contributions

Conception of idea: [Dinithi Vithanage], [Ping Yu]. Literature search and data analysis: [Dinithi Vithanage], [Ping Yu]. Critical review and revision: [Dinithi Vithanage], [Ping Yu], [Lei Wang], [Chao Deng]. Data acquisition: [Mengyang Yin], Data annotation: [Mengyang Yin], [Mohammad Alkhalaf], [Zhenyua Zhang], [Yunshu Zhu], [Alan Christy Soewargo].

Declarations

Availability of Data and Material

Data is not available.

Funding Statement

The authors did not receive support from any organization for the submitted work.

Competing Interest

The authors have no conflicts of interest to declare.

Consent for Publication

All authors have approved the manuscript and agree with its submission to Journal of Healthcare Informatics Research.

Supplementary Material

The Supplementary Material for this article can be found from Supplementary Material.pdf.

Statements and Declarations

This manuscript has not been published elsewhere and is not under consideration by another journal. All authors have approved the manuscript and agree with its submission to the Journal of Healthcare Informatics Research.

References

1. Jiang, L.Y., et al., Health system-scale language models are all-purpose prediction engines. Nature, 2023. **619**(7969): p. 357-362.
2. Bhate, N.J., et al., Zero-shot Learning with Minimum Instruction to Extract Social Determinants and Family History from Clinical Notes using GPT Model. arXiv preprint arXiv:2309.05475, 2023.
3. Dhingra, L.S., et al., Cardiovascular Care Innovation through Data-Driven Discoveries in the Electronic Health Record. Am J Cardiol, 2023. **203**: p. 136-148.

4. Ge, J., et al., A comparison of large language model versus manual chart review for extraction of data elements from the electronic health record. medRxiv, 2023.
5. Ji, B., VicunaNER: Zero/Few-shot Named Entity Recognition using Vicuna. arXiv preprint arXiv:2305.03253, 2023.
6. Yu, H., et al., Open, Closed, or Small Language Models for Text Classification? arXiv preprint arXiv:2308.10092, 2023.
7. Singhal, K., et al., Large language models encode clinical knowledge. Nature, 2023. 620(7972): p. 172-180.
8. Zakka, C., et al., Almanac: Retrieval-augmented language models for clinical medicine. Res Sq, 2023.
9. Lee, P., S. Bubeck, and J. Petro, Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med, 2023. **388**(13): p. 1233-1239.
10. Goel, A., et al. Llms accelerate annotation for medical information extraction. in Machine Learning for Health (ML4H). 2023. PMLR.
11. Wornow, M., et al., The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med, 2023. **6**(1): p. 135.
12. Yu, P., et al. Leveraging generative AI and large language models: A comprehensive roadmap for healthcare integration. Healthcare, 2023. **11**, DOI: 10.3390/healthcare11202776.
13. Shah, M. Prompt engineering vs. fine tuning: Which approach is right for your enterprsie generative AI strategy? 2023; Available from: https://www.prophecy.io/blog/prompt-engineering-vs-fine-tuning-which-approach-is-right-for-your-enterprise-generative-ai-strategy.
14. Liu, P., et al., Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 2023. **55**(9): p. 1-35.
15. Fu, H.Y., et al., Estimating large language model capabilities without labeled test data. arXiv preprint arXiv:2305.14802, 2023.
16. Brown, T., et al., Language models are few-shot learners. Advances in neural information processing systems, 2020. **33**: p. 1877-1901.
17. Lee, Y., et al., Crafting in-context examples according to LMs' parametric knowledge. arXiv preprint arXiv:2311.09579, 2023.
18. Williams, K. Building confidence in LLM outputs: Approaches to increase confidence in content generated by large language models. 2023.
19. Rubin, O., J. Herzig, and J. Berant, Learning to retrieve prompts for in-context learning. arXiv preprint arXiv:2112.08633, 2021.
20. Ding, N., et al., Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 2023. **5**(3): p. 220-235.
21. Nguyen, T.T., C. Wilson, and J. Dalins, Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. arXiv preprint arXiv:2308.14683, 2023.
22. Chai, C.P., Comparison of text preprocessing methods. Natural Language Engineering, 2023. **29**(3): p. 509-553.
23. Abdallah, A., et al., Amurd: annotated multilingual receipts dataset for cross-lingual key information extraction and classification. arXiv preprint arXiv:2309.09800, 2023.
24. Breslow, N., A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika, 1970. **57**(3): p. 579-594.
25. Nachar, N., The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. Tutorials in quantitative Methods for Psychology, 2008. **4**(1): p. 13-20.
26. P. Liang et al., "Holistic evaluation of language models," arXiv preprint arXiv:2211.09110, 2022.