medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

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

Dinithi Vithanage[1], Chao Deng[2], Lei Wang[1], Mengyang Yin[3], Mohammad Alkhalaf[1], Zhenyu 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, Zhenyu 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) is challenging due to the semantic complexity of textual data. Identifying the most effective machine learning method for EHR classification is useful in real-world clinical settings. Advances in natural language processing (NLP) using large language models (LLMs) offer promising solutions. Therefore, this experimental research aims to test the effects of zero-shot and few-shot learning prompting, with and without parameter-efficient fine-tuning (PEFT) and retrieval-augmented generation (RAG) of LLMs, on the multi-label classification of unstructured EHR data from residential aged care facilities (RACFs) in Australia. The four clinical tasks examined are agitation in dementia, depression in dementia, frailty index, and malnutrition risk factors, using the Llama 3.1-8B. Performance evaluation includes accuracy, macro-averaged precision, recall, and F1 score, supported by non-parametric statistical analyses. Results indicate that both zero-shot and few-shot learning, regardless of the use of PEFT and RAG, demonstrate equivalent performance across the clinical tasks when using the same prompting template. Few-shot learning consistently outperforms zero-shot learning when neither PEFT nor RAG is applied. Notably, PEFT significantly enhances model performance in both zero-shot and few-shot learning; however, RAG improves performance only in few-shot learning. After PEFT, the performance of zero-shot learning is equal to that of few-shot learning across clinical tasks. Additionally, few-shot learning with RAG surpasses zero-shot learning with RAG, while no significant difference exists between few-shot learning with RAG and zero-shot learning with PEFT. These findings offer crucial insights into LLMs for researchers, practitioners, and stakeholders utilizing LLMs in clinical document analysis.

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

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 developed to handle structured health data, while much important clinical information is captured in free-text clinical notes, which introduces complexity for model development and deployment.

Electronic health records in RACFs in Australia are digitized systems designed to collect, store, and display data about clients' demographics, medical diagnoses, assessments, progress notes, charts, and forms [1]. Similar to other healthcare settings [1], besides the structured diagnosis data, much important clinical information in RACFs is captured in unstructured, narrative, free-text nursing progress notes. Because free text is a more expressive and

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

natural way for care staff to record care encounters and communicate among team members, these notes are often updated. They, thus, are the closest to real-time reflection of an older person's health condition. Therefore, effectively extracting information from unstructured clinical notes in EHR is essential to support clinical decision-making, improve aged care quality, and advance translational research.

Multi-label classification assigns multiple labels or categories to a single input instance. Multi-label classification of free-text data is a specialized area in machine learning and natural language processing (NLP). It involves the automated extraction of entities, concepts, events, and their relations from unstructured text [2], a challenging task because text data often has different meanings and interpretations [3] and requires precise and expeditious information extraction [4]. Previous works in multi-label classification of EHR text using NLP have explored various approaches. Rule-based systems, including MERKI, MedLEE, SymText, NegEx, MedEx, MedXN, MetaMap, KnowledgeMap, HITEx, and ContextD, have been effectively utilized for this purpose [5-9]. These systems depend on predefined rules and specialized medical dictionaries to extract and categorize information from unstructured text. Although rule-based systems are known for their simplicity, their performance can decline if the narrative text includes terms not covered by the lexical resources [10].

Recently, algorithms like naïve Bayes, k-nearest neighbour, conditional random field, support vector machine, logistic regression, decision tree, random forest, and artificial neural networks have been employed for classification tasks in NLP [11]. Notably, transformer-based encoder-type language models, particularly various BERT models such as ClinicalBERT, SciBERT, and RoBERTa, have been applied to multi-label classification tasks using unstructured EHR data. For example, BERT models have been fine-tuned to handle the semantic complexity of EHR data, demonstrating effectiveness in assigning multiple labels to clinical notes [13]. Despite the advancement of various BERT models, Clinical NLP remains a labour-intensive process that demands a substantial amount of expertise and human efforts to prepare the training data [1, 12]. This limitation has hindered the practical application of the early NLP technique in information extraction from the unstructured, free-text EHR. Privacy and confidentiality concerns also hinder manual curation efforts and the sharing of annotated medical corpora [13].

The recent advancements in LLMs, such as GPT variants, T5, OPT, and Llama [2], have demonstrated the ability of these decoder models to generate text that is not only human-like but also surpasses human-level performance in specific tasks [2]. These models, when combined with machine learning techniques like pre-training, fine-tuning, retrieval augmented generation (RAG) and prompt-based learning [14], 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 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 [4, 15, 16] and extracting clinical data from public health data sets [17], their practical application in specific tasks using clinical data produced in real-world clinical settings remains limited [4, 12, 17, 18]. 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 [4, 19]. The optimal prompting strategies for healthcare information extraction, whether zero-shot or few-shot learning, using PEFT or RAG, in various contexts, remain unclear.

A prompt is an input a user enters to instruct an LLM to automatically generate sequential output [20]. An LLM uses pattern matching to identify the relationships between the words, phrases, and concepts in the prompt and connect these with the patterns learned from the previous training. It then 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 3, thereby expanding the model's applicability and effectiveness. Prompt learning reduces the need to introduce new parameters or extensive retraining of the model using labelled data for various tasks, thus improving efficiency and reducing computational resources required for machine learning.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Extracting symptoms of various geriatric diseases is essential for symptom assessment, early disease diagnosis, personalised treatment, and improving patient outcomes. To date, there is no reporting of practical tools to execute this multi-label classification task accurately and reliably from free-text notes in an EHR system. There is also a lack of prior research on the difference in performance between zero-shot and few-shot learning for the same clinical classification task and the effect of PEFT or RAG 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. Understanding machine learning methods, e.g., prompting behaviour, is also crucial for the safe and effective deployment of LLMs in healthcare settings.

Therefore, we compared zero-shot and few-shot learning, with and without PEFT or RAG, on multi-label clinical classification tasks. In this study, we included 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 machine learning performance. 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, 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,	83

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

	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 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	36
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

We designed experiments to test the research hypotheses (see Table 2).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Table 2: Research hypotheses in the study.

Hypothesis	Null Hypothesis	Alternative Hypothesis
1	Zero-shot and few-shot learning with similar prompting templates have the different level of performance when applied to multi-label classification for various clinical tasks.	Zero-shot and few-shot learning with similar prompting templates have the same level of performance when applied to multi-label classification for various clinical tasks.
2	Few-shot learning does not perform better than zero-shot learning for the multi-label classification of the same clinical task.	Few-shot learning performs better than zero-shot learning for the multi-label classification of the same clinical task.
3	Parameter-efficient fine-tuning does not improve the performance of either zero-shot or few-shot learning.	Parameter-efficient fine-tuning improves the performance of both zero-shot and few-shot learning.
4	Zero-shot learning does not reach the same level of performance as few-shot learning for the same clinical task after PEFT.	Zero-shot learning reaches the same level of performance as few-shot learning for the same clinical task after PEFT.
5	Fine-tuning for one clinical task does not impact model performance across other clinical tasks.	Fine-tuning for one clinical task impacts model performance across other clinical tasks.
6	Retrieval augmented generation does not improve the performance of both zero-shot and few-shot learning.	Retrieval augmented generation improves the performance of both zero-shot and few-shot learning.
7	Zero-shot learning does not reach the same level of performance as few-shot learning for the same clinical task with RAG.	Zero-shot learning reaches the same level of performance as few-shot learning for the same clinical task with RAG.
8	Few-shot learning with RAG and zero-shot learning with PEFT do not have the same performance levels in multi-label classification.	Few-shot learning with RAG and zero-shot learning with PEFT has the same performance levels in multi-label classification.

2 Methodology

We conducted the experiment in seven stages: generative AI-based large language model selection, data set selection, data preprocessing, designing prompt 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 selected the Llama 3.1-8B-parameter model as the generative AI-based LLM. The selection considered 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 [2]. We obtained the Llama 3.1 8B-parameter model from the Hugging Face repository (https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct).

2.3 Data set selection

De-identified demographic data and free-text nursing progress notes were collected for the same population of older people living in 40 RACFs in New South Wales, Australia, from 2019 to 2021. The structured demographic information included masked sequence numbers for client de-identification, age, and gender. The unstructured nursing notes included nursing assessment and progress reporting. They documented clients' daily activities, care staff's clinical observations, assessments of client's care needs (including risk factors), and carer interventions.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

2.4 Data preprocessing

Text preprocessing involved the removal of URLs and non-textual characters, such as extra delimiters and empty spaces in the dataset. We made a choice not to exclude stop words because many of them, like "a," "be," "very," "should," etc., held semantic relevance to the content [21].

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

First, we selected prompt-based training via zero-shot and few-shot learning.

2.5.1 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 [22]. Zero-shot learning has achieved impressive performance in a variety of NLP tasks, such as summarization, dialogue generation, and question-answering [4]. 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% [23]. 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.

2.5.2 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 [24, 25]. These in-context examples not only teach the LLM the mapping from inputs to outputs but also activate the LLM's parametric knowledge. 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 [24]. Few-shot learning is particularly useful in situations where annotating text data is not convenient or expensive. Domain experts can quickly create a generative AI system for a new task by only providing a few examples. Importantly, few-shot learning does not change the underlying model weights [26]. This allows for efficient adaptation to new tasks without risking the loss of previously learned knowledge. However, the performance of few-shot learning varies and is highly task-dependent [24]. Its accuracy is also sensitive to the choice of prompt templates and in-context examples. Prior research finds that using semantically similar in-context examples to those with prior success can significantly enhance the performance of few-shot learning [27].

We adopted the zero-shot and few-shot learning template developed by Abdallaha et al. [28] to construct our prompt (see Figure 1).

```
Below is an instruction that describes a task. Write a response that
appropriately completes the request.

###Instruction:
{instruction}

###Input:
{input}

### Response:
{output}
```

Figure 1: Prompt template adapted from Abdallaha et al. [28].

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

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Table 3: Prompts used in this 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 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, 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. 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 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, 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,

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

		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. 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 behaviors, 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,

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

		mobility and transfer problems, osteoporosis, Parkinsonism 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. 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, requirement for care, respiratory disease, skin ulcer, sleep disturbance, social vulnerability, thyroid disease, urinary incontinence, urinary system disease, visual impairment, weight loss and anorexia. 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 of thinking, feeling discouraged, feeling empty, feeling hopeless, feeling of

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

		excessive guilt, feeling sad, feeling worthless loss of energy, loss of interest, loss of pleasure, suicidal ideation, suicide, suicide attempt, tearfulness. 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 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. 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 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. Please follow these steps. 1. Identify any risk factors of malnutrition. 2. If risk factors of malnutrition are evident, please list them.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

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 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 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. 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 selected Llama 3.18B with PEFT, prompt-based learning (without PEFT or RAG), and RAG to test the LLM's ability to adapt, generalize, and optimize performance in clinical multi-domain classification tasks.

2.6.1 Experiment setup

Parameter efficient fine-tuning with LoRA on Llama 3

Parameter-efficient fine-tuning involves modifying the LLM, or the parameters used to train the LLM, to improve model response to the same prompt [20]. Fine-tuning changes a model's weight, thus, the model's behavior 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

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

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 optimizing the model's performance without extensive resource requirements or overfitting concerns. By focusing on PEFT, LoRA minimizes the dependency on extensive labelled data for model optimization, which maximizes the utility of available data, making the fine-tuning process more effective and feasible in scenarios with limited annotated datasets [29].

We used the PEFT method to fine-tune the Llama model. The experiment was conducted on four NVIDIA RTX - A5000, each equipped with 24GB of memory. The use of multiple GPUs not only accelerated the training process but also ensured that the model could be fine-tuned within a reasonable timeframe. Our software environment was Ubuntu 18.04, the programming language was Python 3.10.0, and the deep learning framework was Pytorch 2.0.0. Instruction data points were employed during the PEFT process, and the hyperparameter settings are listed in Table 4. In these hyperparameter settings, batch size refers to the number of training examples processed in a single iteration of the model's training process. However, due to the memory limitations of the available hardware, we employed a micro-batch size strategy, which splits the full batch into smaller chunks and processes them sequentially. This allowed us to handle the memory load efficiently while still achieving the target batch size of 128 with the Llama model [30, 31]. To further optimize memory usage, we applied 8-bit quantization using the BitsandBytes library, compressing the model weights into a lower precision format that fits within our GPU memory [32]. This same quantization strategy was utilized for other experiments as well, including (1) prompt-based learning without PEFT or RAG and (2) RAG, which are discussed in subsequent sections. The combination of micro-batch size and quantization was essential for effectively training the model within the hardware constraints, ensuring the smooth and efficient execution of our experiments.

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%

Llama 3 model's maximum token limit is 8192, which was large enough to encompass the available tokens for each nursing note. During the fine-tuning process, the model iteratively processed each note within the defined token limit. The labelled dataset used in our study was originally annotated following the procedures detailed in Zhu et al. (2024) by the large research group [33]. The annotated data set for agitation in dementia was developed for the study of Zhu et al. (2023) [34], the malnutrition data set was developed for the study of Alkhalaf et al (2023) [35]. Applying the same manual annotation method to the same data set, we further developed annotation datasets for depression in dementia and frailty index. We randomly divided this labelled data (ensuring no overlapping free-text notes within the labelled dataset) detailed in Table 5 into 80% training, 10% validation, and 10% testing sets for each clinical task. This process was repeated three times to mitigate potential bias from different data splits, and cross-validation was conducted to achieve reliable results [36]. The dedicated test datasets were explicitly used to assess the other machine learning methods including prompt-based learning (without PEFT or RAG), and RAG, allowing for a comprehensive comparison and analysis of each clinical task to test the research hypotheses outlined in Table 2.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

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 specialized PEFT of Lama 2
Depression in dementia	700 nursing notes	280 KB	Depression in dementia with specialized PEFT of Lama 2
Frailty index	949 nursing notes	154 KB	Frailty index with specialized PEFT of Lama 2
Malnutrition risk factors	2850 nursing notes	972 KB	Malnutrition risk factors with specialized PEFT of Lama 2

We used the prompts delineated in Table 3 to evaluate the test data. First, we conducted zero-shot learning with or without PEFT or RAG on Llama 3 across the test datasets for each clinical task. This was followed by few-shot learning with or without PEFT or RAG on Llama 3 across the same test datasets for each clinical task. To ensure that few-shot learning does not benefit from any residual effects of the previous zero-shot learning during testing, we began each time with the downloaded Llama 3 model from the Hugging Face repository for each training method and clinical task. Furthermore, to address variability in results caused by model randomness, multiple evaluation sessions were conducted over three-weeks, with results being collected at three distinct time points. We applied the same approach of repeated evaluation sessions for consistency in two other experiments—prompt training without PEFT or RAG, and RAG—allowing for a more reliable and comparable evaluation across all methods.

Prompt-based learning with zero-shot and few-shot learning on Llama 3

The experiment was conducted in an environment similar to the one described in the "Parameter efficient fine-tuning with LoRA on Llama 3" section, with the main difference being the number of GPUs used. Since prompt-based learning with the original Llama 3 requires fewer computational resources due to the absence of fine-tuning [37], we were able to utilize a single GPU with 24GB. The same process was employed for the experiment with RAG, as no fine-tuning was involved in the RAG process.-We employed Llama 3, utilizing zero-shot and few-shot learning prompts as outlined in Table 3. To prevented model contamination, we approached each clinical task (see Table 1) in two distinct steps. Initially, we employed the Llama 3 model that we directly downloaded from the Hugging Face repository, for zero-shot learning. Afterwards, we downloaded a new copy of the same model from the same repository for few-shot learning. The Llama 3 model has a maximum token limit of 8192, and none of the test notes (see the sections titled ‘parameter efficient fine-tuning with LoRA on Llama 3' for details about the test datasets) exceeded this token count during evaluation. Consequently, the model processed each note iteratively within this defined token limit during testing.

Retrieval augmented generation with Llama 3

Retrieval-augmented generation enhances LLMs by enabling access to real-time, relevant information from external knowledge sources, improving response accuracy and relevance [32]. The process includes indexing, where data is converted into embeddings for efficient retrieval; retrieval, where the system identifies the most pertinent documents based on a user query; augmentation, where the retrieved information enriches the query's context; and generation, where the LLM produces a response informed by both the original query and the augmented data. This dynamic approach addresses traditional LLM limitations, ensuring up-to-date and contextually accurate responses [32].

The experiment for Llama 3, with RAG, began with data preparation, utilizing a JSON file containing nursing notes that encompass all the labels for each clinical task (refer to Table 1). Each clinical task, and the number of nursing notes provided in the JSON file, are detailed in Table 6.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Table 6: Data used in the RAG

Clinical Tasks	Number of nursing notes	File Size
Agitation in dementia	83	14KB
Depression in dementia	13	5 KB
Frailty index	36	8 KB
Malnutrition risk factors	37	10 KB

This data was input into an embedding model. The embedding model (https://huggingface.co/sentence-transformers/all-mpnet-base-v2) generated embeddings from the data, which were then stored in a vector store. A query was formulated by combining the prompt with zero-shot or few-shot examples. The query was applied to the dedicated test datasets (see the sections titled 'parameter efficient fine-tuning with LoRA on Llama 3' for details about the test datasets). The retriever used the query to fetch relevant data from the vector store. The retrieval process, managed by LlamaIndex, determined relevance based on the similarity between the query and the stored embeddings, which were then used by the LLM to generate the output [38, 39].

2.7 Model performance evaluation

To assess model performance on the four clinical tasks, we utilized micro-average precision, recall, F1 score, and accuracy, which were more appropriate for imbalanced class distributions [40, 41]. For the automatic evaluation of multi-label text classification, we first calculated the BERTscore to obtain precision, recall, and F1 score following the method by Li et al. [42]. We then used the Scikit-learn library, which offers built-in metrics for multi-label classification, to calculate the micro-average F1 score, precision, and recall [43, 44]. The model outputs, generated in JSON format, were compared against the annotated ground truth data, also stored in JSON format. An extracted entity or phrase was considered correct if it overlapped the text and conveyed the exact or highly similar meaning of the annotated ground truth entity or phrase. The comparison was facilitated by BERTscore and Scikit-learn library tools. For instance, if the ground truth annotation is 'shouting' as a symptom of agitation in dementia, and the model output is also 'shouting,' it is evaluated as an exact match. Alternatively, if the ground truth annotation is 'reject meals' and the model output is 'refuse meals,' it is considered a match based on semantic similarity, as the two phrases share the same meaning. To measure semantic similarity between model predictions and the ground truth, a similarity threshold of 0.5 was used, following prior work in this domain [45, 46].

2.8 Statistical analysis

As the measurement indicators including accuracy, precision, recall and F1 score were not normally distributed, we utilized 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 [28, 47]. A significant difference is decided if the p-value is smaller than 0.05. The results generated from the statistical analysis are showcased in Supplementary Table 2.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

3 Results

3.1 Results of testing Hypothesis 1: Zero-short and few-shot learning with similar prompting templates may (or may not) exhibit the same level of performance when applied to multi-label classification for various clinical tasks.

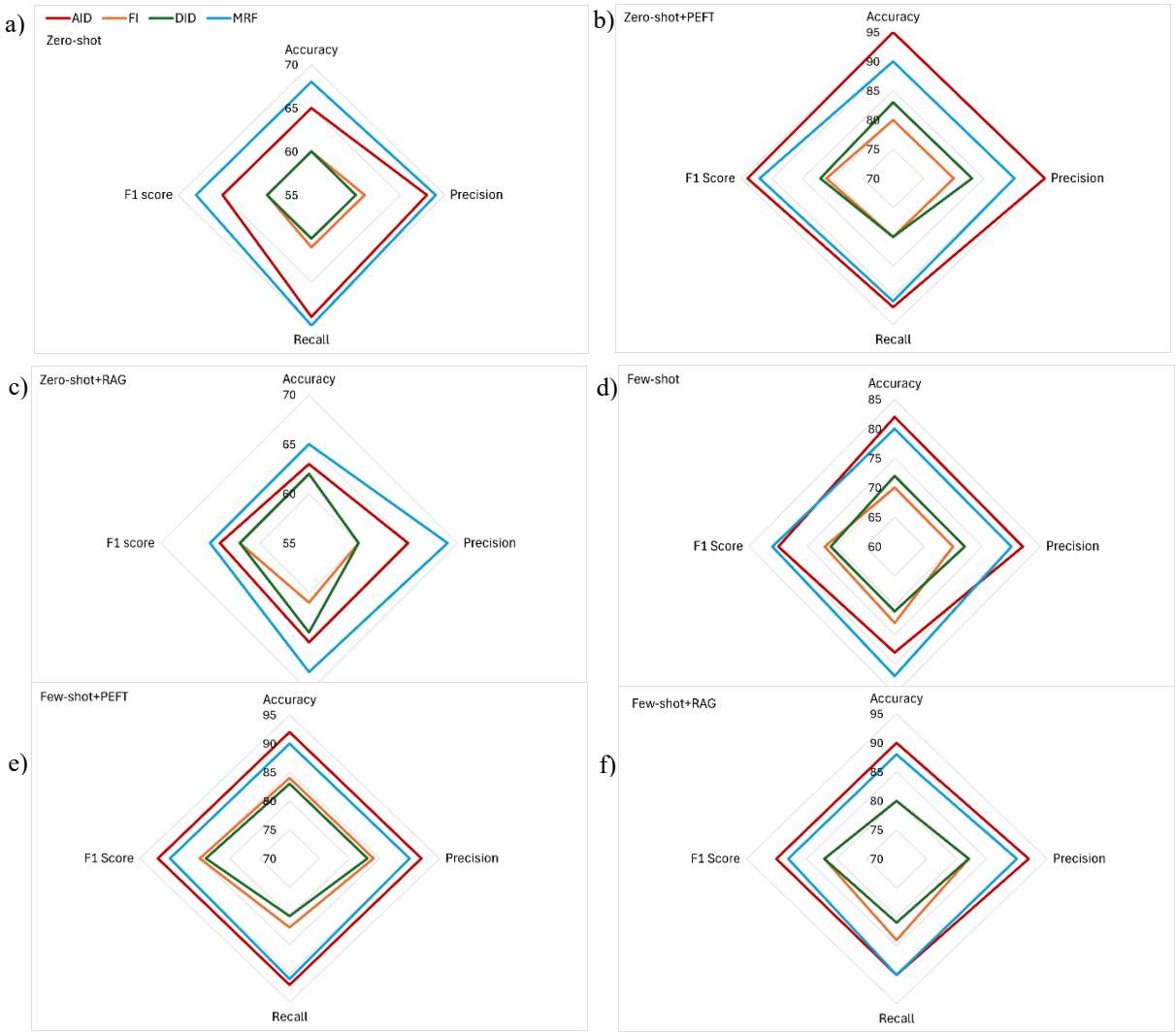
To evaluate Hypothesis 1, we compared: (1) the performance of zero-shot learning (without PEFT or RAG); (2) the performance of zero-shot learning with PEFT; (3) the performance of zero-shot learning with RAG; (3) the performance of few-shot learning (without PEFT or RAG); (5) the performance of few-shot learning with PEFT; and (6) the performance of few-shot learning with RAG across the four clinical tasks.

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

There is no statistically significant difference in accuracy, precision, recall, and F1 score between these classification tasks undertaking this training method (Figure 2-a, $p > 0.05$). 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.

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

Once again, no statistically significant difference is found in accuracy, precision, recall, and F1 score between these classification tasks undertaking this training method (Figure 2-b, $p > 0.05$). 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.



medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Figure 2: Comparative evaluation of model performance for four multi-label classification tasks with the following training methods: (a) zero-short learning, (b) zero-short learning with PEFT, (c) zero-short learning with RAG, (d) few-short learning, (d) few-short learning with PEFT, and (f) few-short learning with RAG.
Note: 'AID' denotes agitation in dementia, 'FI' denotes frailty index, 'DID' denotes depression in dementia, 'MRF' denotes malnutrition risk factors. The same notation applies to the figures that follow.
Note: '+PEFT' denotes with PEFT. The same notation applies to the figures that follow.
Note: '+RAG' denotes with RAG. The same notation applies to the figures that follow.

3.1.3 Comparing the performance of zero-shot learning with RAG for four clinical tasks.

No statistically significant difference is found in accuracy, precision, recall, and F1 score between these tasks undertaking this training method (Figure 2-c, $p > 0.05$). However, the same trend as above is 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.

3.1.4 Comparing the performance of few-shot learning (without PEFT or RAG) for the four clinical tasks.

No statistically significant difference is found in accuracy, precision, recall, and F1 score between these tasks undertaking this training method (Figure 2-d, $p > 0.05$). However, the same trend as above is 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.

3.1.5 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 2-e, $p > 0.05$); 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.

3.1.6 Comparing the performance of few-shot learning with RAG for four clinical tasks.

No statistically significant difference is found in accuracy, precision, recall, and F1 score between these tasks undertaking this training method (Figure 2-f, $p > 0.05$). However, the same trend as above is 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.

3.2 Results of testing Hypothesis 2: Few-shot learning may (or may not) perform better than zero-shot learning for the multi-label classification of the same clinical tasks.

To evaluate Hypothesis 2, we compared the performance of zero-shot and few-shot learning (without PEFT and RAG) for all four clinical tasks. Few-shot learning significantly improves model accuracy, precision, recall and F1 score in all four multi-label clinical classification tasks than zero-shot learning (Figure 3-a, $p < 0.05$). The level of improvement includes an 18% increase in model accuracy, an 18% increase in precision, a 25% increase in recall, and a 28% increase in F1 score.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

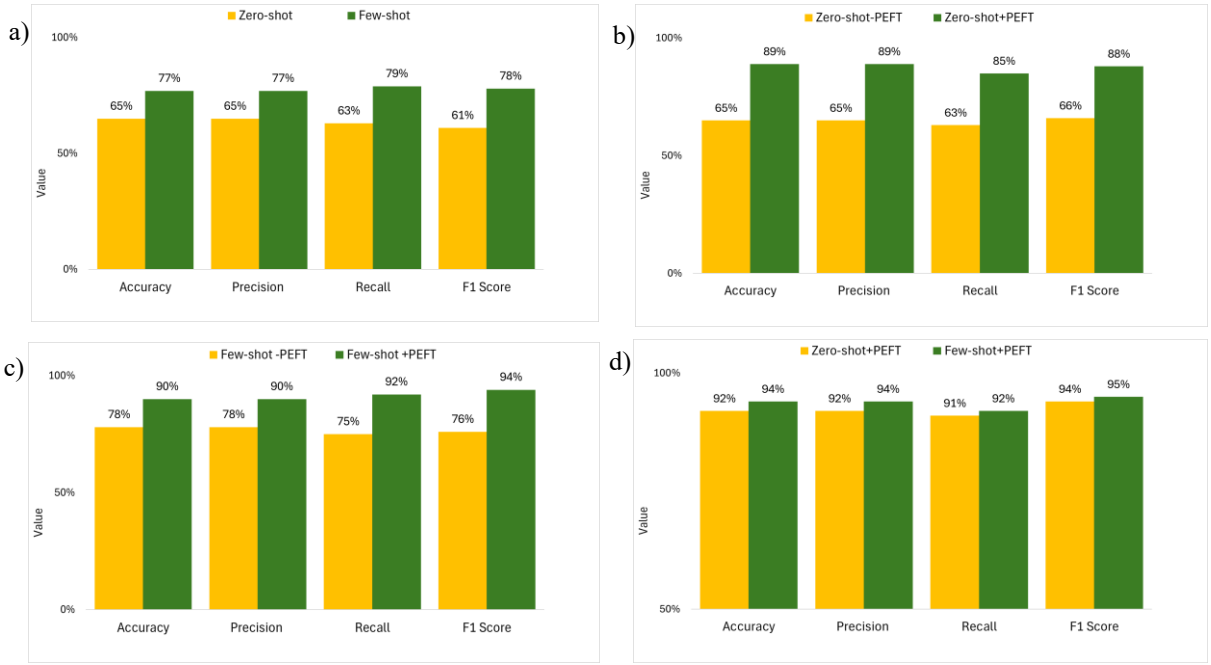


Figure 3: Comparative evaluation of model performance for four multi-label classification tasks with the following training methods: (a) zero-shot learning versus few-shot learning, (b) zero-shot learning versus zero-shot learning with PEFT, (c) few-shot learning versus few-shot learning with PEFT, and (d) zero-shot learning versus few-shot learning with PEFT.
Note: '-PEFT' denotes without PEFT. The same notation applies to the figures that follow.

3.3 Results of testing Hypothesis 3: Parameter-efficient finetuning may (or may not) improve both zero-shot and few-shot learning performance.

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

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

Zero-shot learning with PEFT significantly improves model accuracy, precision, recall and F1 score in all four multi-label clinical classification tasks (Figure 3-b, $p < 0.05$). The level of improvement is as follows: a 37% increase in model accuracy, a 37% increase in precision, a 35% increase in recall, and a 33% increase in F1 score.

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

Few-shot learning with PEFT significantly improves model accuracy, precision, recall and F1 score in all four multi-label clinical classification tasks (Figure 3-c, $p < 0.05$). The level of improvement 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.

3.4 Results of testing Hypothesis 4: Zero-shot learning may (or may not) reach the same level of performance as few-shot learning for the same clinical task with PEFT.

To evaluate Hypothesis 4, we compared 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 all four multi-label clinical classification tasks (Figure 3-d, $p > 0.05$), there is a trend that few-shot learning performs above zero-shot learning.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

3.5 Results of testing Hypothesis 5: Fine-tuning for one clinical task may (or may not) impact model performance across other clinical tasks.

To evaluate Hypothesis 5, we compared the performance of a clinical task-specific PEFT model with zero-shot learning, and its impact across other clinical tasks. For example, we conducted PEFT on the clinical task of agitation in dementia. We then compared the model's impact on multi-label classification for depression in dementia, frailty index and malnutrition risk factors. We then compared the performance of a clinical task-specific PEFT model with few-shot learning and its impact across other clinical tasks, following the same pattern we did for zero-shot learning.

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

No significant difference is found in accuracy, precision, recall, and F1 score for the other group of clinical tasks between two training models, pure zero-shot learning and zero-shot learning with PEFT for a specific clinical task. (see Figure 4-a, Figure 4-b, p >0.05).

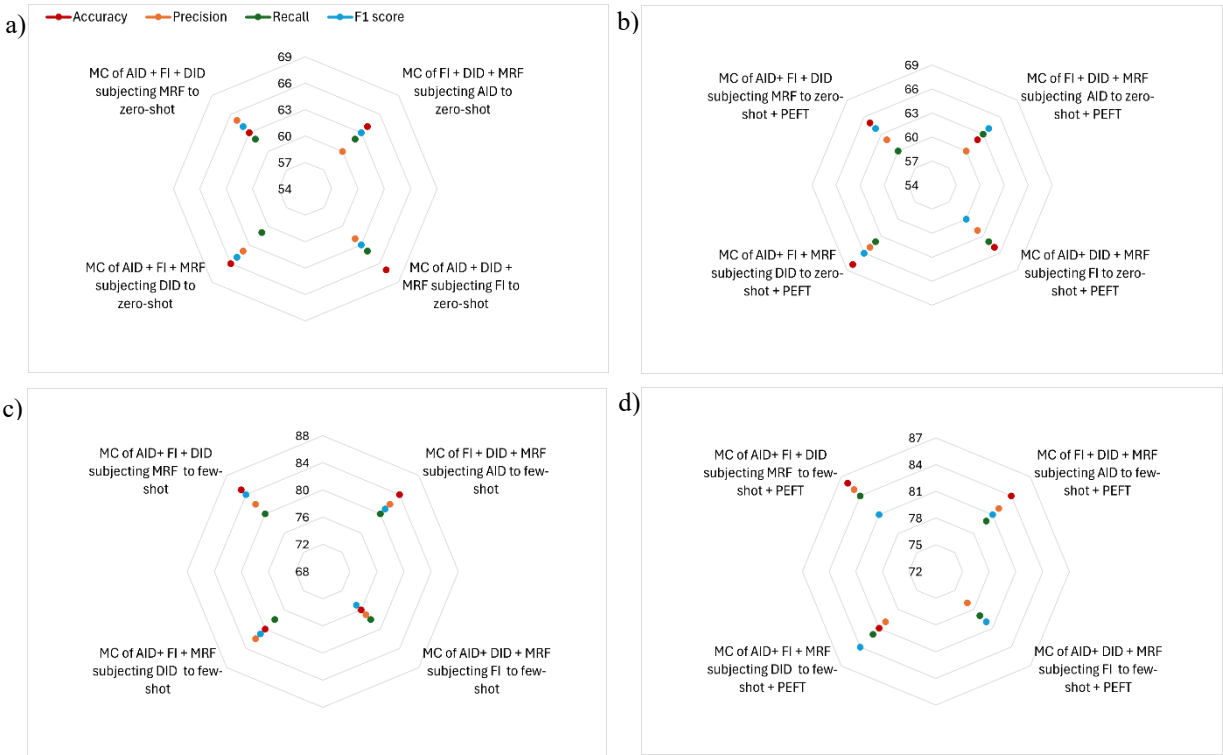


Figure 4. Performance of the measurement indicators for the other three multi-label classification tasks (a) when one clinical task was only trained with zero-shot learning, (b) when training on one clinical task with zero-shot learning and PEFT, (c) when one clinical task was only trained with few-shot learning, and (d) when training on one clinical task with few-shot learning and PEFT.
Note: 'MC' denotes with multi-label classification.

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

No significant difference is found in accuracy, precision, recall, and F1 score for the other group of clinical tasks between two training models, pure few-shot learning and few-shot learning with PEFT for a specific clinical task. (see Figure 4-c, Figure 4-d, p >0.05).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

3.6 Results of testing Hypothesis 6: Retrieval augmented generation may (or may not) improve both zero-shot and few-shot learning performance.

To evaluate Hypothesis 6, we compared the performance of the model of (1) zero-shot learning without RAG and with RAG; and (2) few-shot learning without RAG and with RAG for all four tasks.

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

Although no statistically significant difference is found in accuracy, precision, recall, and F1 score between the zero-shot learning without RAG and with RAG in all four multi-label clinical classification tasks (Figure 5-a, $p > 0.05$), the results suggest a trend where zero-shot learning with RAG tends to perform better than without RAG.

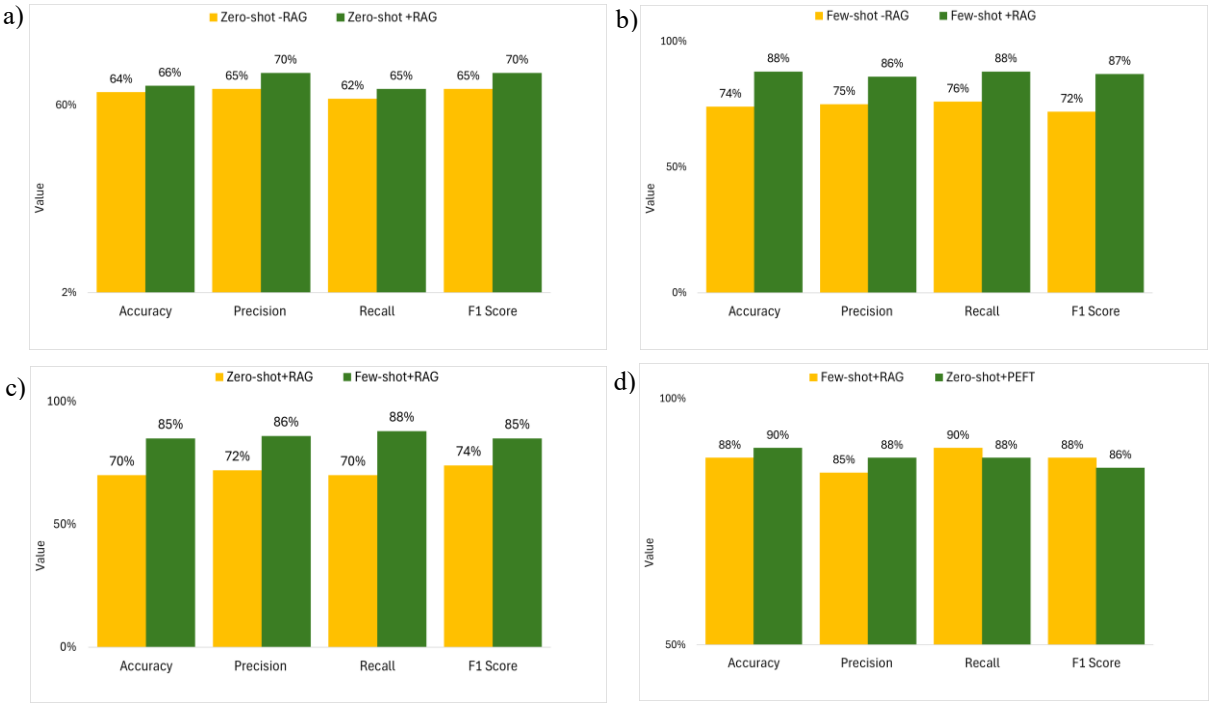


Figure 5. Comparative evaluation of model performance for four multi-label classification tasks with the following training methods: (a) zero-shot learning without RAG versus with RAG, (b) few-shot learning without RAG versus with RAG, (c) zero-shot learning with RAG versus few-shot learning with RAG, and (d) few-shot learning with RAG versus zero-shot learning with PEFT.
Note: '-RAG' denotes without RAG.

3.6.2 Comparing the performance of few-shot learning model without RAG and with RAG for all four tasks.

Few-shot learning with RAG significantly improves model accuracy, precision, recall and F1 score in all four multi-label clinical classification tasks (Figure 5-b, $p < 0.05$). The level of improvement is as follows: a 19% increase in model accuracy, a 15% increase in precision, a 16% increase in recall, and a 21% increase in F1 score.

3.7 Results of testing Hypothesis 7: Zero-shot learning may (or may not) reach the same level of performance as few-shot learning for the same clinical task with RAG.

To evaluate Hypothesis 7, we compared the performance of zero-shot and few-shot learning with RAG for all four clinical tasks. Few-shot learning with RAG significantly improves model accuracy, precision, recall and F1 score in all four multi-label clinical classification tasks than zero-shot learning (Figure 5-c, $p < 0.05$). The level of

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

improvement is as follows: a 17% increase in model F1 score, an 18% increase in model accuracy, an 18% increase in precision, and a 23% increase in recall.

3.8 Results of testing Hypothesis 8: Few-shot learning with RAG may (or may not) perform better than zero-shot learning with PEFT for all four tasks.

To evaluate Hypothesis 8, we compared the performance of few-shot learning with RAG and zero-shot learning with PEFT for all four clinical tasks. No significant difference is found in accuracy, precision, recall, and F1 score between the few-shot learning with RAG and zero-shot learning with PEFT (Figure 5-d, $p > 0.05$) in all four multi-label clinical classification tasks.

4 Discussion

This study explores the impact of zero-shot and few-shot prompt learning strategies, both with and without PEFT and RAG, 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, eight research hypotheses have been formulated, and experimental designs have been implemented to rigorously test these hypotheses.

Our first hypothesis proposed that zero-shot and few-shot learning with similar prompting templates exhibit comparable performance in multi-label classification across various clinical tasks. Our findings support the alternative hypothesis, confirming that there is no significant difference in performance between these two learning approaches with similar prompting templates. While we did not identify statistically significant differences across all tasks, a consistent pattern emerged: the agitation in dementia and malnutrition risk factor classification tasks showed slightly better performance in accuracy, precision, recall, and F1 score compared to the frailty index and depression in dementia tasks. This performance disparity may be attributed to Llama 3 being trained on more data related to the former two tasks, leading to enhanced knowledge and better classification outcomes for those specific tasks.

Our second hypothesis proposed that few-shot learning performs better than zero-shot learning for the multi-label classification of the same clinical task. Our findings support the alternative hypothesis, confirming that few-shot learning significantly enhances performance compared to zero-shot learning (without PEFT or RAG). This improvement indicates that few-shot domain adaptation effectively minimizes false positives and false negatives while increasing true positives in classification tasks. These results highlight that exposing an LLM to initial information from the target domain can significantly improve its performance [48-50] in classification tasks. Our third hypothesis suggested that PEFT can improve both zero-shot and few-shot learning performance. The results support the alternative hypothesis, demonstrating that PEFT significantly improves performance in both learning modes compared to the models without PEFT. Based on our findings, we observed that applying PEFT within a specific domain effectively reduces false positives and false negatives while increasing true positives in domain-specific information extraction tasks. This confirms that the PEFT approach facilitates targeted modifications to the model's parameters [27, 51, 52], leading to improved overall performance in multi-label classification tasks within that domain.

Our fourth hypothesis proposed that zero-shot learning can reach the same level of performance as few-shot learning for the same clinical task after PEFT. Our findings support the alternative hypothesis, confirming that there is no significant difference in performance between zero-shot and few-shot learning with PEFT. Based on our findings, this implies that further exposure through few-shot learning may not be necessary for models already fine-tuned within the domain. Our fifth hypothesis proposed that fine-tuning for one clinical task impacts model performance across other clinical tasks. However, the findings support the null hypothesis, indicating that fine-tuning for a specific task does not significantly affect the model's performance on other classification tasks. 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 [27, 51]. 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.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Our sixth hypothesis asserts that RAG can improve both zero-shot and few-shot learning performance. However, the findings support the null hypothesis, showing no significant difference in performance between zero-shot learning with and without RAG. In contrast, a significant improvement was observed in few-shot learning when RAG was utilized. The lack of enhancement in zero-shot learning with RAG may be due to the absence of task-specific examples, which limits the model's ability to effectively leverage retrieved information and generalize to unseen tasks [48, 49, 53]. Nevertheless, the results suggest that zero-shot learning with RAG consistently performs slightly better than without RAG, which could indicate that RAG offers a subtle advantage by providing additional context, though not enough to result in substantial gains. Conversely, the significant performance boost in few-shot learning suggests that RAG effectively leverages the limited examples provided to enrich model understanding, leading to improved classification accuracy. This disparity highlights RAG's potential effectiveness in contexts where the model can benefit from supplementary information, particularly in few-shot settings [54, 55].

Our seventh hypothesis asserts that zero-shot learning can reach the same level of performance as few-shot learning for the same clinical task with RAG. Our findings support the null hypothesis, demonstrating that there is a significant difference in performance between zero-shot and few-shot learning with RAG. Zero-shot learning may not reach the same performance level as few-shot learning with RAG due to the absence of task-specific examples that help the model better understand the nuances of the clinical tasks. It relies solely on the model's pre-existing knowledge [48, 49], which may not cover the specific clinical details required for accuracy. In contrast, few-shot learning provides the model with critical initial information and context from the target domain, enabling it to make more informed predictions [54-56]. As a result, while RAG can enhance contextual relevance, the lack of direct examples in zero-shot learning limits its effectiveness compared to few-shot approaches.

Our eighth hypothesis proposed that few-shot learning with RAG and zero-shot learning with PEFT achieve the same performance levels in multi-label classification. The results support the alternative hypothesis, showing that both methods can deliver similar outcomes despite their distinct mechanisms. Zero-shot learning with PEFT selectively fine-tunes the most relevant parameters [27], enabling the model to adapt effectively to new domains with minimal labelled data [27, 51]. On the other hand, few-shot learning with RAG benefits from integrating relevant external information through retrieval, combined with task-specific examples [55, 56]. This enhances the model's ability to handle complex clinical data, leading to fewer classification errors, such as false positives and false negatives—an insight derived from our study's findings. Our findings demonstrate that both PEFT and RAG excel in adapting to domain-specific tasks, highlighting their complementary strengths [57, 58] in improving model performance for multi-label classification in real-world clinical settings.

The findings of this study have significant implications for integrating AI models, particularly LLMs, into clinical practice. When combined both zero-shot and few-shot learning approaches with RAG and PEFT on multi-label classification tasks, the findings, present efficient solutions for improving multi-label classification in various healthcare settings. Given the ongoing challenge of obtaining labelled data from domain experts, RAG, combined with few-shot learning, emerges as an efficient option. By leveraging external knowledge sources through retrieval, RAG minimizes dependence on large, annotated datasets, reducing the need for extensive fine-tuning specific to each domain. This makes RAG a quicker and less resource-intensive solution, as it doesn't require the same level of GPU resources or training time as PEFT, which do not require high technical expertise [37]. In contrast, PEFT demands more computational power and expertise, requiring substantial GPU resources and additional time for fine-tuning specific model parameters [37]. However, the enhanced performance that PEFT offers makes it useful in scenarios where the computational infrastructure is available.

This adaptability makes RAG equally valuable as PEFT for information retrieval in health and geriatric care, where clinical decision-making often relies on diverse and overlapping health conditions [59]. Therefore, this machine learning method can enable AI systems to integrate domain-specific knowledge in retrieving diagnostic and assessment data from the relevant but limited number of EHRs. This capability can improve diagnostic accuracy and support real-time health assessment, facilitating person-centred treatment and care plan development. Ultimately, the flexible deployment of this technique can help streamline care delivery in complex clinical contexts, reducing the cost and time associated with data labelling, requiring fewer resources and less training time, while maintaining robust model performance.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

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. In the future, we will broaden our scope by incorporating additional clinical classification tasks into our study. Secondly, although we have examined four multi-label clinical classification tasks, additional tasks, such as question answering, summarization, and relation extraction are yet to be explored to enable a more comprehensive knowledge about the LLM's performance in EHR data. The third limitation relates to the selection of model performance evaluation metrics, which utilizes accuracy, precision, recall, and F1 score as the primary evaluation metrics. In future studies, we will broaden our evaluation metrics to encompass calibration, robustness, fairness, bias, toxicity, and efficiency [60]. Diversifying 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 and RAG on their performance. Our findings reveal that the same prompting template—whether zero-shot or few-shot, with or without PEFT or RAG—yields comparable performance across different multi-label classification tasks. Few-shot learning consistently outperforms zero-shot learning in classification tasks, even without PEFT or RAG. However, when PEFT is integrated with both zero-shot and few-shot learning, zero-shot learning achieves performance levels similar to few-shot learning in classification tasks. On the other hand, with RAG, few-shot learning consistently surpasses zero-shot learning, indicating RAG's superior ability to provide relevant external knowledge in few-shot learning. Our analysis also shows that fine-tuning LLMs for a particular clinical task does not significantly impact on the model's performance when applied to other clinical tasks. Notably, few-shot learning with RAG achieves performance comparable to zero-shot learning with PEFT, further highlighting the adaptability and efficiency of these advanced methods in clinical applications.

Acknowledgements: This research is part of a PhD project supported by the University of Wollongong, Australia and the University Grants Commission, Sri Lanka. It is also supported by Round 6 Research Grant from Aged Care Research and Innovation Grant Australia (ARIIA).

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], [Zhenyu Zhang], [Yunshu Zhu], [Alan Christy Soewargo].

Declarations

Availability of Data and Material

Data is not available to comply with the Australian laws in privacy protection.

Funding Statement

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

Competing Interest

The authors have no conflicts of interest to declare.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Consent for Publication

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

Supplementary Material

The Supplementary Material for this article can be found in 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] L. Y. Jiang *et al.*, "Health system-scale language models are all-purpose prediction engines," *Nature,* vol. 619, no. 7969, pp. 357-362, 2023.

[2] J. Ge, M. Li, M. B. Delk, and J. C. Lai, "A comparison of a large language model vs manual chart review for the extraction of data elements from the electronic health record," *Gastroenterology,* vol. 166, no. 4, pp. 707-709. e3, 2024.

[3] N. J. Bhate, A. Mittal, Z. He, and X. Luo, "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.

[4] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature,* vol. 620, no. 7972, pp. 172-180, 2023.

[5] Z. Afzal, E. Pons, N. Kang, M. C. Sturkenboom, M. J. Schuemie, and J. A. Kors, "ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus," *BMC bioinformatics,* vol. 15, pp. 1-12, 2014.

[6] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: historical perspective and recent advances," *Journal of the American medical informatics association,* vol. 17, no. 3, pp. 229-236, 2010.

[7] S. Khan and J. A. Shamsi, "Health Quest: A generalized clinical decision support system with multi-label classification," *Journal of King Saud University-computer and information sciences,* vol. 33, no. 1, pp. 45-53, 2021.

[8] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "MedEx: a medication information extraction system for clinical narratives," *Journal of the American medical informatics Association,* vol. 17, no. 1, pp. 19-24, 2010.

[9] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system," *BMC medical informatics and decision making,* vol. 6, pp. 1-9, 2006.

[10] R. Farkas and G. Szarvas, "Automatic construction of rule-based ICD-9-CM coding systems," in *BMC bioinformatics*, 2008, vol. 9: Springer, pp. 1-9.

[11] F. Dabek and J. J. Caban, "Leveraging big data to model the likelihood of developing psychological conditions after a concussion," *Procedia computer science,* vol. 53, pp. 265-273, 2015.

[12] L. S. Dhingra, M. Shen, A. Mangla, and R. Khera, "Cardiovascular care innovation through data-driven discoveries in the electronic health record," *The American Journal of cardiology,* vol. 203, pp. 136-148, 2023.

[13] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *proceedings of the 25th international conference on machine learning*, 2008, pp. 160-167.

[14] B. Ji, "Vicunaner: Zero/few-shot named entity recognition using vicuna," *arXiv preprint arXiv:2305.03253,* 2023.

[15] H. Yu, Z. Yang, K. Pelrine, J. F. Godbout, and R. Rabbany, "Open, closed, or small language models for text classification?," *arXiv preprint arXiv:2308.10092,* 2023.

[16] C. Zakka *et al.*, "Almanac—retrieval-augmented language models for clinical medicine," *NEJM AI,* vol. 1, no. 2, p. AIoa2300068, 2024.

[17] P. Lee, S. Bubeck, and J. Petro, "Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine," *New England Journal of medicine,* vol. 388, no. 13, pp. 1233-1239, 2023.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[18] A. Goel *et al.*, "Llms accelerate annotation for medical information extraction," in *machine learning for health (ML4H)*, 2023: PMLR, pp. 82-100.
[19] M. Wornow *et al.*, "The shaky foundations of large language models and foundation models for electronic health records," *npj digital medicine,* vol. 6, no. 1, p. 135, 2023.
[20] P. Yu, H. Xu, X. Hu, and C. Deng, "Leveraging generative AI and large language models: A comprehensive roadmap for healthcare integration," in *Healthcare*, 2023, vol. 11, no. 20: MDPI, p. 2776.
[21] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys,* vol. 55, no. 9, pp. 1-35, 2023.
[22] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, "Gptscore: evaluate as you desire," *arXiv preprint arXiv:2302.04166,* 2023.
[23] 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.
[24] T. Brown *et al.*, "Language models are few-shot learners advances in neural information processing systems 33," 2020.
[25] Y. Lee, P. Atreya, X. Ye, and E. Choi, "Crafting in-context examples according to LMs' parametric knowledge," *arXiv preprint arXiv:2311.09579,* 2023.
[26] O. Rubin, J. Herzig, and J. Berant, "Learning to retrieve prompts for in-context learning," *arXiv preprint arXiv:2112.08633,* 2021.
[27] N. Ding *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature machine intelligence,* vol. 5, no. 3, pp. 220-235, 2023.
[28] A. Abdallah, M. Abdalla, M. Elkasaby, Y. Elbendary, and A. Jatowt, "Amurd: annotated multilingual receipts dataset for cross-lingual key information extraction and classification," *arXiv preprint arXiv:2309.09800,* 2023.
[29] C. P. Chai, "Comparison of text preprocessing methods," *Natural language engineering,* vol. 29, no. 3, pp. 509-553, 2023.
[30] X. Piao, D. Synn, J. Park, and J.-K. Kim, "Enabling large batch size training for dnn models beyond the memory limit while maintaining performance," *IEEE Access,* 2023.
[31] S. Qiao, H. Wang, C. Liu, W. Shen, and A. Yuille, "Micro-batch training with batch-channel normalization and weight standardization," *arXiv preprint arXiv:1903.10520,* 2019.
[32] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, "Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records," *Journal of biomedical informatics,* p. 104662, 2024.
[33] Y. Zhu *et al.*, "COVID-19 and its influence on prevalence of dementia and agitation in Australian residential aged care: A comparative study," *Psychiatry International,* vol. 5, no. 4, pp. 642-659, 2024.
[34] Y. Zhu, T. Song, Z. Zhang, M. Yin, and P. Yu, "A five-step workflow to manually annotate unstructured data into training dataset for natural language processing," in *MEDINFO 2023—The future is accessible*: IOS press, 2024, pp. 109-113.
[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," (in eng), *Technol health care,* vol. 31, no. 6, pp. 2267-2278, 2023, doi: 10.3233/thc-230229.
[36] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," *Journal of cheminformatics,* vol. 6, pp. 1-15, 2014.
[37] S. R. Dandavolu. "Fine-tuning and RAG: Which one is better?" https://www.analyticsvidhya.com/blog/2024/05/fine-tuning-vs-rag/ (accessed.
[38] Y. Liu, "The application of RAG technology in traditional chinese medicine," in *2024 international conference on artificial intelligence and communication (ICAIC 2024)*, 2024: Atlantis press, pp. 402-408.
[39] K. Menon, "Utilizing open-source AI to navigate and interpret technical documents: leveraging RAG models for enhanced analysis and solutions in product documentation," 2024.
[40] J. A. Kumar and S. Abirami, "Ensemble application of bidirectional LSTM and GRU for aspect category detection with imbalanced data," *Neural computing and applications,* vol. 33, no. 21, pp. 14603-14621, 2021.
[41] P. Pramokchon and P. Piamsa-nga, "A feature score for classifying class-imbalanced data," in *2014 International computer science and engineering conference (ICSEC)*, 2014: IEEE, pp. 409-414.
[42] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge," *Cureus,* vol. 15, no. 6, 2023.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.24.24309441; this version posted October 23, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[43] J. Hao and T. K. Ho, "Machine learning made easy: a review of scikit-learn package in python programming language," *Journal of educational and behavioral statistics,* vol. 44, no. 3, pp. 348-361, 2019.

[44] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research,* vol. 12, pp. 2825-2830, 2011.

[45] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Aaai*, 2006, vol. 6, no. 2006, pp. 775-780.

[46] D. Ramage, A. N. Rafferty, and C. D. Manning, "Random walks for text semantic similarity," in *Proceedings of the 2009 workshop on graph-based methods for natural language processing (TextGraphs-4)*, 2009, pp. 23-31.

[47] N. Breslow, "A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship," *Biometrika,* vol. 57, no. 3, pp. 579-594, 1970.

[48] V. D. Gowda, S. Suneel, P. R. Naidu, S. Ramanan, and S. Suneetha, "Challenges and limitations of few-shot and zero-shot learning," in *applying machine learning techniques to bioinformatics: few-shot and zero-shot methods*: IGI global, 2024, pp. 113-137.

[49] Y. Labrak, M. Rouvier, and R. Dufour, "A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks," *arXiv preprint arXiv:2307.12114,* 2023.

[50] K. Peng *et al.*, "Exploring few-shot adaptation for activity recognition on diverse domains," *arXiv preprint arXiv:2305.08420,* 2023.

[51] Z. Han, C. Gao, J. Liu, and S. Q. Zhang, "Parameter-efficient fine-tuning for large models: A comprehensive survey," *arXiv preprint arXiv:2403.14608,* 2024.

[52] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, "Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment," *arXiv preprint arXiv:2312.12148,* 2023.

[53] T. Abdullahi, R. Singh, and C. Eickhoff, "Retrieval augmented zero-shot text classification," in *Proceedings of the 2024 ACM SIGIR international conference on theory of information retrieval*, 2024, pp. 195-203.

[54] M. Bhattarai, J. E. Santos, S. Jones, A. Biswas, B. Alexandrov, and D. O'Malley, "Enhancing code translation in language models with few-shot learning via retrieval-augmented generation," *arXiv preprint arXiv:2407.19619,* 2024.

[55] R. Singhal, P. Patwa, P. Patwa, A. Chadha, and A. Das, "Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs," *arXiv preprint arXiv:2408.12060,* 2024.

[56] G. Izacard *et al.*, "Few-shot learning with retrieval augmented language models," *arXiv preprint arXiv:2208.03299,* vol. 2, no. 3, 2022.

[57] H. Soudani, E. Kanoulas, and F. Hasibi, "Fine tuning vs. retrieval augmented generation for less popular knowledge," *arXiv preprint arXiv:2403.01432,* 2024.

[58] A. Salemi and H. Zamani, "Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models," *arXiv preprint arXiv:2409.09510,* 2024.

[59] D. L. Frosch and R. M. Kaplan, "Shared decision making in clinical medicine: past research and future directions," *American journal of preventive medicine,* vol. 17, no. 4, pp. 285-294, 1999.

[60] P. Liang *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110,* 2022.