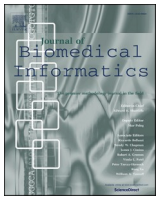
Journal of Biomedical Informatics 156 (2024) 104662

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records

Mohammad Alkhalaf [a,b], Ping Yu [a,*], Mengyang Yin [c], Chao Deng [d]

[a] School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia
[b] School of Computer Science, Qassim University, Qassim 51452, Saudi Arabia
[c] Opal Healthcare, Level 11/420 George St, Sydney NSW 2000, Australia
[d] School of Medical, Indigenous and Health Sciences, University of Wollongong, Wollongong, NSW 2522, Australia

ARTICLE INFO

Keywords:
Generative AI
Nursing notes
LLAMA
Malnutrition
Summarization
RAG

ABSTRACT

Background: Malnutrition is a prevalent issue in aged care facilities (RACFs), leading to adverse health outcomes. The ability to efficiently extract key clinical information from a large volume of data in electronic health records (EHR) can improve understanding about the extent of the problem and developing effective interventions. This research aimed to test the efficacy of zero-shot prompt engineering applied to generative artificial intelligence (AI) models on their own and in combination with retrieval augmented generation (RAG), for the automating tasks of summarizing both structured and unstructured data in EHR and extracting important malnutrition information.

Methodology: We utilized Llama 2 13B model with zero-shot prompting. The dataset comprises unstructured and structured EHRs related to malnutrition management in 40 Australian RACFs. We employed zero-shot learning to the model alone first, then combined it with RAG to accomplish two tasks: generate structured summaries about the nutritional status of a client and extract key information about malnutrition risk factors. We utilized 25 notes in the first task and 1,399 in the second task. We evaluated the model's output of each task manually against a gold standard dataset.

Result: The evaluation outcomes indicated that zero-shot learning applied to generative AI model is highly effective in summarizing and extracting information about nutritional status of RACFs' clients. The generated summaries provided concise and accurate representation of the original data with an overall accuracy of 93.25%. The addition of RAG improved the summarization process, leading to a 6% increase and achieving an accuracy of 99.25%. The model also proved its capability in extracting risk factors with an accuracy of 90%. However, adding RAG did not further improve accuracy in this task. Overall, the model has shown a robust performance when information was explicitly stated in the notes; however, it could encounter hallucination limitations, particularly when details were not explicitly provided.

Conclusion: This study demonstrates the high performance and limitations of applying zero-shot learning to generative AI models to automatic generation of structured summarization of EHRs data and extracting key clinical information. The inclusion of the RAG approach improved the model performance and mitigated the hallucination problem.

1. Introduction

Malnutrition is a significant health concern, which can lead to weight and muscle loss, and affect a person's physical health, cognitive functions, immune function, and overall quality of life. It is prevalent and extremely harmful for the frail older people living in RACFs, and can expose these older people to higher risk of developing chronic diseases, experiencing functional decline, increasing rate of falls, hospitalization and mortality [1,2]. In older people, malnutrition could be caused by a variety of factors, including physiological changes associated with aging, chronic diseases, social isolation, and poor appetite [3,4]. The complex nature of malnutrition demands attention to its risk factors and

* Corresponding author at: Centre for Digital Transformation, School of Computing and Information Technology, Faculty of Engineering and Information Sciences, Northfield Ave, University of Wollongong, Wollongong, NSW 2522, Australia.
E-mail address: ping@uow.edu.au (P. Yu).

https://doi.org/10.1016/j.jbi.2024.104662
Received 3 February 2024; Received in revised form 25 May 2024; Accepted 28 May 2024
Available online 14 June 2024

implement targeted prevention and management strategies. To date, the main methods for nutrition prevention and management are regular in-person screening for malnutrition risks, regular nutritional assessment, and specific interventions to address the risk factors [5,6].

To date, different data analytics have been applied to uncover malnutrition information from the structured EHR [7–9]. However, these methods are inadequate, often missing crucial health information about the nutritional health status of older people that are recorded in narrative text-based notes instead of structured tables in health services [10,11]. It has been found that approximately 80 % to 90 % of health data is recorded in unstructured format [12–14]. Consequently, it is imperative to develop an effective machine learning approach to retrieve key health information from the large volume of unstructured clinical data to understand a person's health status. An effective method to accomplish this is to transform unstructured clinical notes to structured summarization, which could turn patient information into clinical insights. [15,16].

Summarization of the important health information recorded in the free text notes of EHR could potentially enhance health data accessibility and analytics accuracy. This will assist nurses and doctors in efficiently retrieving clinical information to provide timely and informed treatment decisions [17].

In addition to summarization, extracting information about risk factors for a certain disease is also essential for providing timely support and preventive intervention to stop the progression of disease-related complications. Furthermore, understanding an individual's specific risk factors allows for the development of personalized, tailored strategies to address the person's disease-related issues [18–20], and is important for improving client's health outcomes. At the population level, knowledge about the prevalence of risk factors is important for government health departments and health organizations for assessing the health needs of a population, developing prevention programs, prioritizing resources to improve public health and planning policy changes, and creating the right health services [21,22].

Transforming unstructured nursing notes in EHR into structured summarization with key information has been a technical challenge for a variety of reasons, including format inconsistencies, length variations, typos, and specialized medical terms. Extracting risk factors is also a challenge because they are embedded into the clinical notes and may be subjective or depend on individual health history [23,24]. Another challenge is the high demand for domain expertise and time to develop large, annotated corpora [25]. These challenges have made automated extraction of health information through human or basic natural language processing (NLP) techniques a time-consuming and labor-intensive task [26–29].

The latest development in NLP and more specifically, the generative AI and large language models (LLMs) are revolutionizing the way humans interact with text data. These LLMs have the ability to comprehend content of a given text and perform different NLP tasks without the need for large amounts of data for fine-tuning [30]. These advancements can significantly enhance summarization and health risk factor extraction techniques. They can effectively address certain limitations of the aforementioned NLP technologies [27]. Furthermore, they bring in technical breakthroughs in NLP with ability to accurately comprehend content of a given text and perform different tasks without the need for extensive data for fine-tuning; even exceed human capabilities in certain tasks [30–32].

Since GPT-3.5 stormed the world in November 2022, there has been an ongoing surge of interest in exploring the application of the decoder-based generative AI and LLMs in healthcare. To date, the majority of GPT models for medicine are applied to medical questions and answering tasks [30]. Hu et al. applied GPT-3.5 and GPT-4 on clinical named entity recognition tasks on public data sets [25]. However, GPT-3.5 and GPT-4 are commercial cloud-based solutions with data hosted in the US. This has prohibited the researchers in other countries to apply them for clinical text processing due to violation of many countries' information privacy laws. To the best of our knowledge, there is no attempt in applying the open-source LLM to text summarization and health risk factor extraction tasks from real-world EHR in after-hospital care setting. This motivates us to pilot the implementation of prompt engineering techniques in the open-source Llama model for clinical text summarization and health risk factor extraction tasks.

Llama 2 is a state-of-the-art (SOTA) LLM model developed by Meta AI. It is an autoregressive language model that uses optimized transformer architecture. It has been trained on a vast amount of text data using supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to give helpful and safe output [33]. In addition to its impressive performance, what makes Llama 2 model suitable for this project than the commercial SOTA models such as GPT 3.5 and GPT 4 is the availability to run the model locally without sending sensitive personal data to US, an act that violates many countries' health information privacy laws [30,34–36].

Training Llama 2 and other autoregressive models involves the basic operation of prompting them with queries, questions, or tasks. Prompting involves crafting effective instructions or queries to stimulate the model to respond. There are different types of prompting strategies such as zero-shot, few shot learning and chain of thoughts. Zero-shot learning refers to the capability of a language model to generate relevant responses or outputs for tasks it hasn't been explicitly trained on, without specific examples or guidance during training [37].

Example of zero-shot prompting.

Zero-shot Prompt	Q: Name four risk factors of malnutrition in older people?
AI Output	1. Poor Appetite or Reduced Food Intake 2. Financial Constraints 3. Underlying Health Conditions 4. Limited Access to Nutrient-Rich Foods

To unlock the full potential of generative AI models, a recently developed technique known as Retrieval augmented generation has emerged. This technique combines elements of both retrieval and generation methods [38]. In this approach, the model uses a retrieval system to retrieve relevant information from a dataset or knowledge base and then generates responses or content based on that retrieved information [39]. Retrieval systems are efficient in finding relevant information from a large dataset, while generation models excel at creating coherent and contextually appropriate responses. By incorporating retrieval, the model can access factual information from a knowledge base and produce contextually relevant responses [40,41]. This integration process enhances model accuracy and mitigate issues like hallucination. Notably, the integration of RAG with generative models is relatively new, and there is limited empirical research assessing its effectiveness in health informatics applications [39].

This study makes three methodology innovations: first, it employs cutting-edge open-source Generative AI model Llama 2 for clinical text data processing. These models are known for their human-like ability to comprehend and generate text summary; yet their application in clinical NLP is still in the early days. Therefore, our study validates the feasibility of applying the generative AI models for clinical NLP. Second, we designed and implemented a zero-shot soft prompt for the summarization of clinical notes. This method diverges from traditional techniques that require significant amounts of training data for each clinical NLP task, presenting a more adaptable and efficient solution. Third, we piloted the integration of traditional generation techniques with information retrieval using RAG technique. This integration enables our model to not only generate text based on learned patterns but also retrieve pertinent information from a knowledge base or dataset. The result is enhanced accuracy and relevance in clinical note summarization.

Statement of purpose.

	Summary
Problem or Issue	Early identification of health deficits can aid timely intervention. Summarization of the important health information recorded in the free text notes of EHR along with health risk factor extraction could enhance health data accessibility and prevent health deterioration.
What is Already Known	Automation of summarization and health risk factor extraction is a challenging task. Previous solutions that require adequate training data are not optimal.
What this Paper Adds	This paper presents an innovative approach that applies soft prompts within an open-source LLM to automate clinical text summarization and extract health risk factors. It also pilots the RAG approach to mitigate the risk of LLM hallucination.

2. Methodology

The dataset was obtained from 40 residential aged care facilities in the state of New South Wales (NSW), Australia. The Human research ethics approval was granted by the Human Research Ethics Committee, the University of Wollongong and the Illawarra Shoalhaven Local Health District (Year 2020).

There is no publicly available malnutrition-labeled data; therefore, we first built a malnutrition-specific labeled dataset. To accomplish this, we developed a process to identify and label records with malnutrition [42]. This started with the construction of a rule-based model to identify all malnutrition and weight loss related notes in the dataset. Using these rules, we extracted 2,474 notes belonging to 1,283 clients. Manual analysis and screening of all extracted notes identified 196 notes that did not fit the malnutrition definition and were marked with negative labels. Ultimately, we manually labeled ground truth dataset that contain 2,278 progress notes with information about malnutrition. The entire process underwent inter-rater reliability assessment by three nursing domain experts, resulting in a 92 % agreement. In addition to nursing notes, structured data such as demographic and weight tables were utilized. Demographic table includes age and gender information while the weight table includes monthly weight records of each resident. This labeled data was only utilized as a gold standard to ensure model answers align with those of healthcare professionals, nurses and dietitians, reinforcing the importance of generating accurate and clinically relevant output.

We conducted the zero-shot experiments on two specific tasks: (1) Structured summarization of malnourished residents' notes, and (2) malnutrition risk factor extraction from the progress notes. We compared the outcome of both tasks with and without the integration of the RAG approach (Fig. 1).

2.1. Task 1: Structured summarization of text notes for malnourished residents

To achieve this objective, the task involves converting extensive text-based dietitian notes into a structured summarization that highlights crucial nutrition-related information (Table 1). Our dataset consists of

Table 1
Task 1 corpus.

Task 1	Structured summarization of text notes for malnourished residents This task uses zero-shot prompting to summarize nursing notes.
Input	Nursing note
Output	Age, gender, weight, BMI, weight history, medical history, list risk factors causing malnutrition or weight loss and list malnutrition interventions or recommendations
Number of notes	Without RAG integration: 25 notes. One note per resident With RAG integration: 52 notes. Number of notes per resident: Mean: 2.08,Standard Deviation (SD):1.09

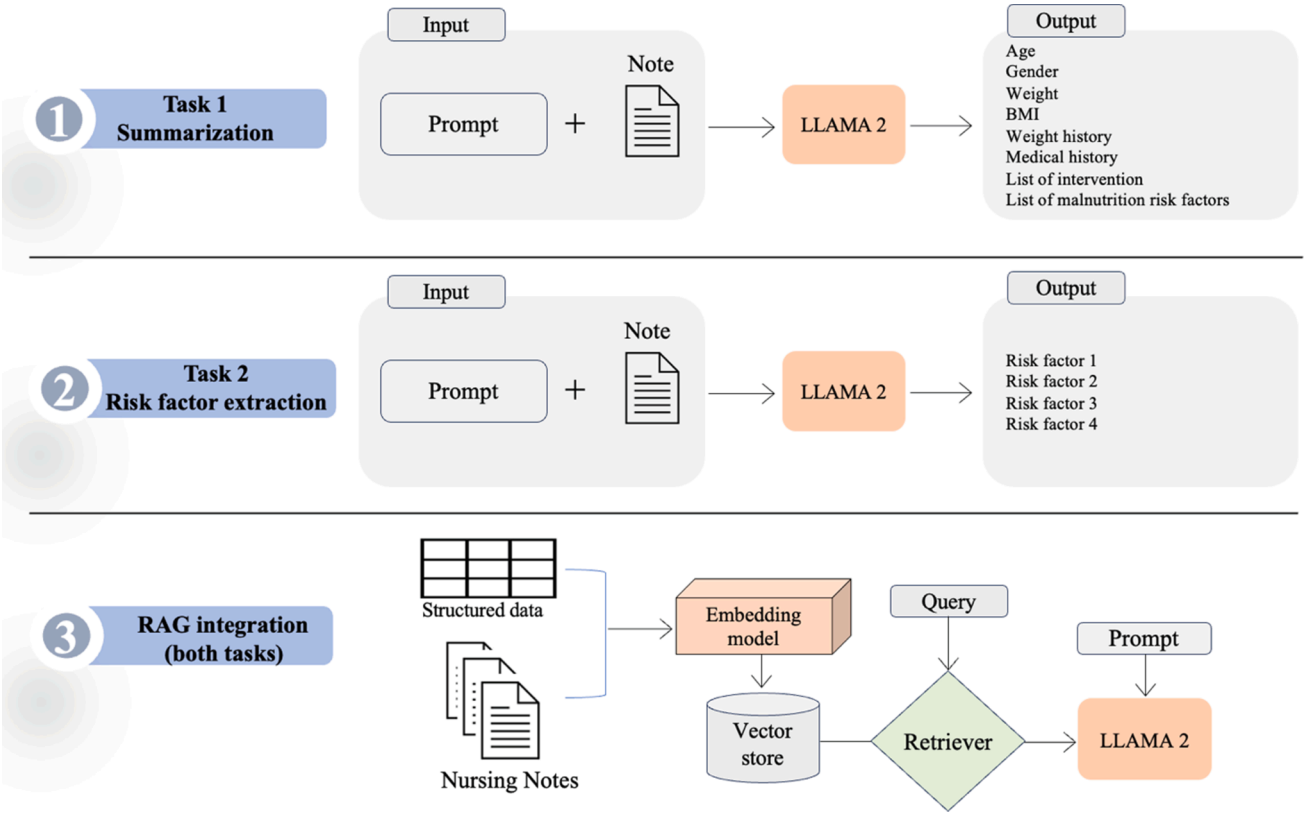


Fig. 1. Overview of the machine learning methods, including instruction tuning and RAG.

25 notes, each attributed to a specific resident, with an average sequence length of 850 tokens (95 % confidence interval: 740.7 – 984.3) for each note.

Additionally, in the case of RAG, we incorporated all available dietitian notes for the 25 residents, resulting in a total of 52 notes. The RAG system enabled us to effectively include more notes. The retrieval process in RAG concentrates on extracting essential information without requiring the inclusion of the entire note, addressing the challenge of maximum token length. For example, the number of tokens for all the dietitian notes of each resident could exceed the maximum limit of the model, and the retrieval system solves this issue by focusing only on retrieving queried information.

A detailed structured summary was made by humans for each text note and served as the "gold standard" for the model evaluation. We constructed the prompt for zero-shot learning in the following manner:

Prompt 1. Structured summary of notes.

Summary template	Summary_template = Age: Gender: Weight: BMI: Weight history: Medical history: List risk factors causing malnutrition or weight loss:List malnutrition interventions or recommendations:
Prompt template	Prompt_template = [INST]≪SYS≫you are a helpful assistant≪/SYS≫Given the following context, please extract and fill in the summary template with the relevant information. If there is no information for a specific part of the template, leave that specific part of the template unfilled. context: [nursing_note] summary template: [Summary_template]filled template: [/INST]
Input	Actual nursing note from the dataset
Output	Structured summary

[INST] and [/INST] are instruction tags to inform the model that the content between the two tags are user instruction(s) or input. ≪SYS≫ and ≪/SYS≫ are system tags to mark the message from the system, which inform the model to understand its role being system. The system needs to respond to the instructions given by the users. Summary template provides specific instruction to the model to present the output in the format defined by the template.

2.2. Task 2: Extracting malnutrition risk factors from the nurse notes

In this task, we analyzed text notes to extract malnutrition risk factors (Table 2). We considered our previous findings as the gold standard for this task evaluation [42]. The data used in this task comprises 1,399 notes. Each note describes a resident's weight and nutrition situation. Although most notes highlight one or more risk factors, there are 550 notes that do not specify the underlying factors causing weight loss. We intentionally retained these notes to assess whether the model adheres strictly to the given prompt, or if it generates responses based on prior knowledge, potentially causing hallucination, i.e., the model makes up facts in responses. In this step, we instruct the model to go through each nursing note to extract key risk factors causing malnutrition.

Table 2
Task 2 corpus.

Task 2	Extracting malnutrition risk factors from the nurse notes This task uses zero-shot prompting to extract malnutrition risk factors from nursing notes.
Input	Nursing note
Output	list of malnutrition risk factors causing malnutrition
Number of notes	Number of notes: 1,399 notes (550 with no risk factors mentioned.) Number of residents: 719 No. of notes per resident: Mean: 1.92, SD: 1.37.

Prompt 2. Extract risk factors of malnutrition and weight loss.

Risk factor template	Risk_factor_template= risk factor 1: risk factor 2: risk factor 3:risk factor 4:
Prompt template	Prompt_template = [INST]≪SYS≫You are a helpful assistant ≪/SYS≫Use the provided context to fill in the risk factors template with 4 words or less, describing each malnutrition or weight loss risk factor mentioned in the context. If the context does not mention any risk factors, just say no risk factor mentioned. Do not add extra information. context: [nursing_note] Risk factors template: [Risk_factor_template]filled template: [/INST]
Input	Nursing note
Output	List of risk factors

2.3. Retrieval augmented generation

To enable the model to access an external knowledge base including more nursing notes and structured data, we utilized the RAG approach. This approach involves taking nursing notes and breaking them into manageable chunks of a fixed size (600 characters). This step is essential to facilitate subsequent processing and analysis. Similarly, when dealing with structured data, this approach parses the information row by row. Next we utilized the sentence transformer model from Huggingface library [43]. This model is designed to encode sentences into dense vectors, capturing semantic meaning and relationships between words. By embedding the segmented data, we transform information into a numerical representation that preserves the contextual information and relationships within the data. The embedded vectors are then stored in a vector store along with their metadata, providing an efficient approach for retrieval and analysis. Vector stores allow for fast similarity searches, enabling quick access to relevant information [44]. Then, we utilized maximum marginal relevance retrieval (MMR), a well-established search algorithm provided by the LangChain library to retrieve related documents. MMR operates by maximizing the diversity of the selected documents while ensuring their relevance to the query. This selection leverages the reliability and proven performance of MMR over alternative methods such as similarity search. The parameter for the search operation (k) was set at 20, the limitation of the available GPU memory for processing documents simultaneously without triggering out-of-memory errors. To retrieve information for a specific resident on a specific period of time, a retriever takes a resident ID, date and a query as input and searches the vector store for documents relevant to the query. The retrieved documents contain the embedded vectors representing the segmented data related to the resident and query. After that, retrieved documents are sent to the generative model along with specific prompt instruction to produce a concise and informative summary (Fig. 2).

We chose Llama 2 model since it is the leading open-source model. Given that Llama 2 was regarded as the best available model for the task, especially considering its origins from a reputable research team like Meta AI, it was selected as the primary model for the study. This decision was driven by the desire to utilize a high-quality, SOTA model from a reliable source, ensuring the credibility and efficacy of the research outcomes. It was also accessible for researchers to run locally, ensuring that sensitive data remains private and not shared. We also selected the 13 billion parameter version of the model instead of the 7 billion parameter one since it did not significantly increase memory requirements, allowing it to run efficiently on standard GPU hardware. This meant that the computational infrastructure needed for running the larger model was readily available and feasible within the project's resource constraints. Moreover, Meta AI's tests and evaluations had demonstrated that the 13 billion parameter model exhibited superior performance compared to the 7 billion parameter version in all

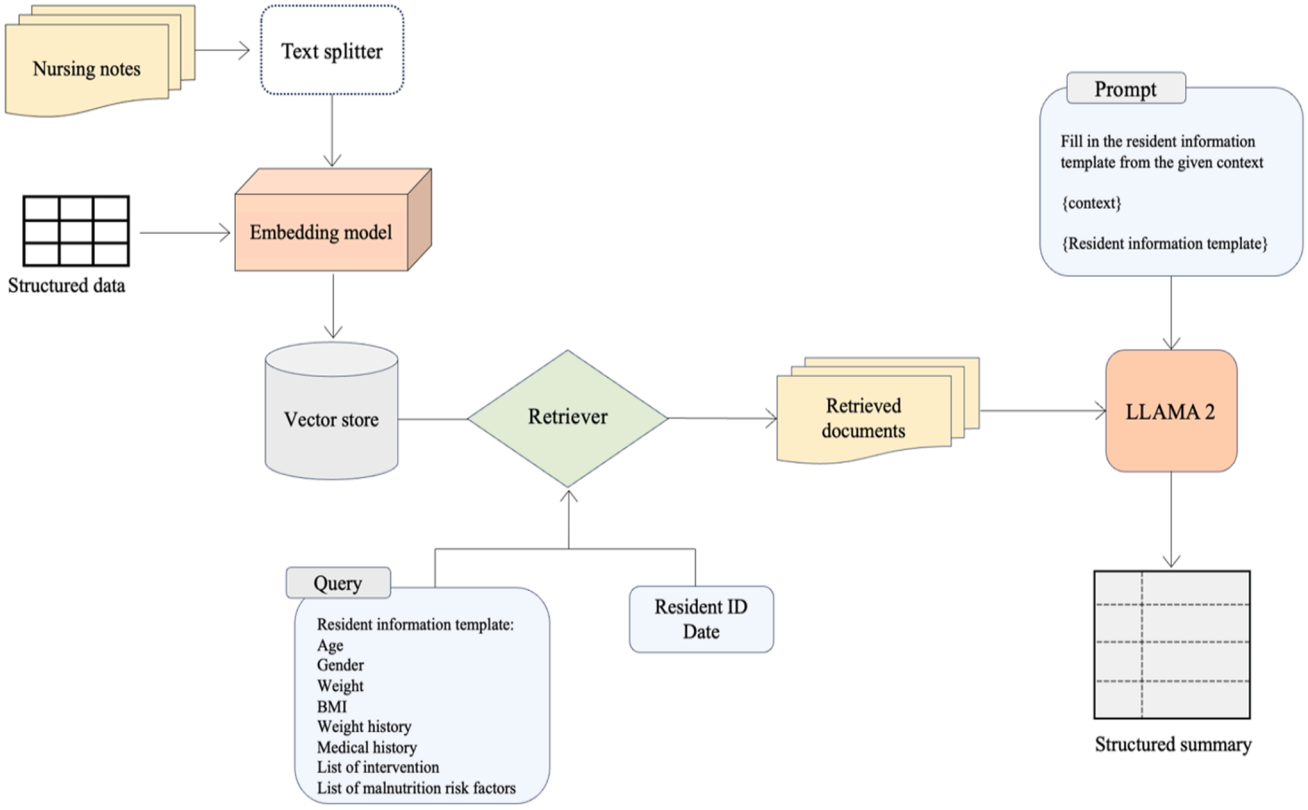


Fig. 2. Overview of RAG approach.

benchmarks. This evidence provided confidence that selecting the larger model would likely result in better outcomes for the task at hand.

We acquired permission to access the Llama 2 model's weight from Meta [45] and Huggingface library. We used a quantization process to reduce the number of bits used to represent the weights and activations of a neural network, while minimizing the impact on the model's accuracy. We used BitsandBytes library to quantize the model weights (4 bits) to fit into our accessible GPU memory [46]. Tesla T4 GPU with 15 GB memory was used for all experiments. LangChain library was used for the RAG approach [47].

3. Results

3.1. Performance of Task1: Summarization of nutrition care with zero-shot prompt engineering alone

Human evaluation was conducted by reading each generated note and comparing its meaning with the gold standard summary generated by human experts, with special focus on the key variables and values. The results showed a high level of alignment in the machine generated summaries with the gold standard summaries in all of the 25 notes. However, there were some hallucinations observed, specifically with the age (11 notes), weight (one note) weight history (one note) and list of risk factors (one note).

Regarding age, in 11 instances where age was not mentioned, the model filled out the age part of the template with wrong numbers rather than following the prompt and indicating "not specified". In other five instances where age was not mentioned, it correctly followed the prompt and stated that age was not specified. Regarding gender, it was explicitly mentioned in one note which was correctly extracted by the model. In the other 16 instances, the model correctly inferred gender from the context while it stated not specified in eight instances. For both the weight and weight history when they were not specifically mentioned in one instance, the model filled their parts with inaccurate information instead of indicating " not specified". In terms of the risk factors list, the model added an extra risk factor in one instance. The overall accuracy was 93.25 % (Table 3).

Table 3
Comparison of accuracy of zero-shot learning alone versus zero-shot learning in conjunction with RAG.

Field	Accuracy without RAG (%)	Error type	Accuracy with RAG (%)	P-value* (0.05)
Age	56	Intrinsic and extrinsic hallucination*	100	< 0.05
Gender	100	–	100	< 0.05
Weight	96	Intrinsic hallucination	100	> 0.05
BMI	100	–	100	> 0.05
Weight history	96	Intrinsic hallucination	100	> 0.05
Medical history	100	–	98	> 0.05
Intervention list	100	–	98	> 0.05
Risk factor list	98	Intrinsic hallucinations	98	> 0.05
Overall accuracy	93.25		99.25	

*__Intrinsic hallucinations__: generated content is based on information present in the input data but is misrepresented or synthesized incorrectly.
__Extrinsic hallucinations__: generated content is not supported by the source material at all [48].
__BMI__: Body Mass Index.
*__P-value__ considers both accuracy of the model's output and the instances where the model couldn't provide an output ('not specified' outputs). By comparing the performance of the model with and without the RAG using the p-value, we assess whether the inclusion of RAG significantly improves the model's overall output.

3.2. Performance of task 1: Summarization of nutrition care combining zero-shot prompt with retrieval augmented generation

After giving the model access to more notes (52 notes) and to structured data through RAG approach, the model performance improved noticeably. It only overlooked one diagnosis in medical history (one note), one recommendation (one note) and one risk factor (one note). The overall accuracy achieved through the RAG approach was 99.25 % (Table 3). Refer to Supplementary Tables S1 and S3 for more details and examples of the results.

3.3. Performance of task 2: Extracting malnutrition risk factors from the nurse notes

We compared the model's identified malnutrition risk factors with those that were manually extracted by human experts, i.e., gold standard in two ways. First, we randomly selected notes and checked if risk factors extracted are correct or not. Second, we compared the prevalence of risk factors identified by the model to the prevalence of risk factors identified by the gold standard.

3.3.1. The model's level of accuracy in malnutrition risk factor identification for each progress note

We purposely sampled 200 nursing notes to verify the model's output accuracy in extracting the correct malnutrition risk factors. Malnutrition risk factors were explicitly mentioned in 100 notes and were not specified in another 100 notes. The purpose was to determine if the model strictly followed the prompt instructions or used its pre-existing malnutrition knowledge to generate general output regarding malnutrition. For the 100 notes with risk factors mentioned, the model correctly extracted the risk factors in 95 notes. However, it missed the risk factors in 5 notes, resulting in an accuracy rate of 95 %. For the notes that did not record specific risk factors, the model correctly followed the prompt and stated that "no risk factor mentioned" in 85 notes. In contrast, it generated risk factors that were not mentioned in the notes (extrinsic hallucination) in 15 notes, resulting in an accuracy rate of 85 %. Therefore, the overall accuracy of the 200 notes is 90 %. We also compared the result after the integration of RAG. However, in this task, the RAG approach did not have any effect on the model performance.

3.3.2. The model's accuracy in identification of the prevalence of malnutrition risk factors

We compared the model's performance in extracting prevalence of risk factors based on the gold standard [42]. In this step, we removed all notes that did not mention any cause. This was done to ensure a focused comparison between the model's output and the gold standard, specifically emphasizing instances where causes were explicitly mentioned. The model has suggested that poor appetite is the highest malnutrition risk factor. It was followed by low intake and dementia. The statistical analysis revealed differences between AI output and the gold standard across various diagnoses such as poor appetite, poor intake, dysphagia, surgery, nausea, confusion, pneumonia, and stroke, while showing no statistical difference for other diagnoses (Fig. 3). Refer to Supplementary Tables S2, S4 and S5 for more details and examples of Task 2 results.

4. Discussion

In this study we tested the ability of generative AI model Llama 2 in summarization and extracting risk factors for malnutrition and weight loss from residential aged care nursing progress notes. Consistent with the findings of previous studies [31,49], our experiment has achieved promising results. Given the increasing complexity of health data, these LLMs are promising to significantly improve the secondary use of data captured in EHR, specifically the free text notes, which are the most flexible and frequently updated information in healthcare.

The model performance in generating structured summaries of

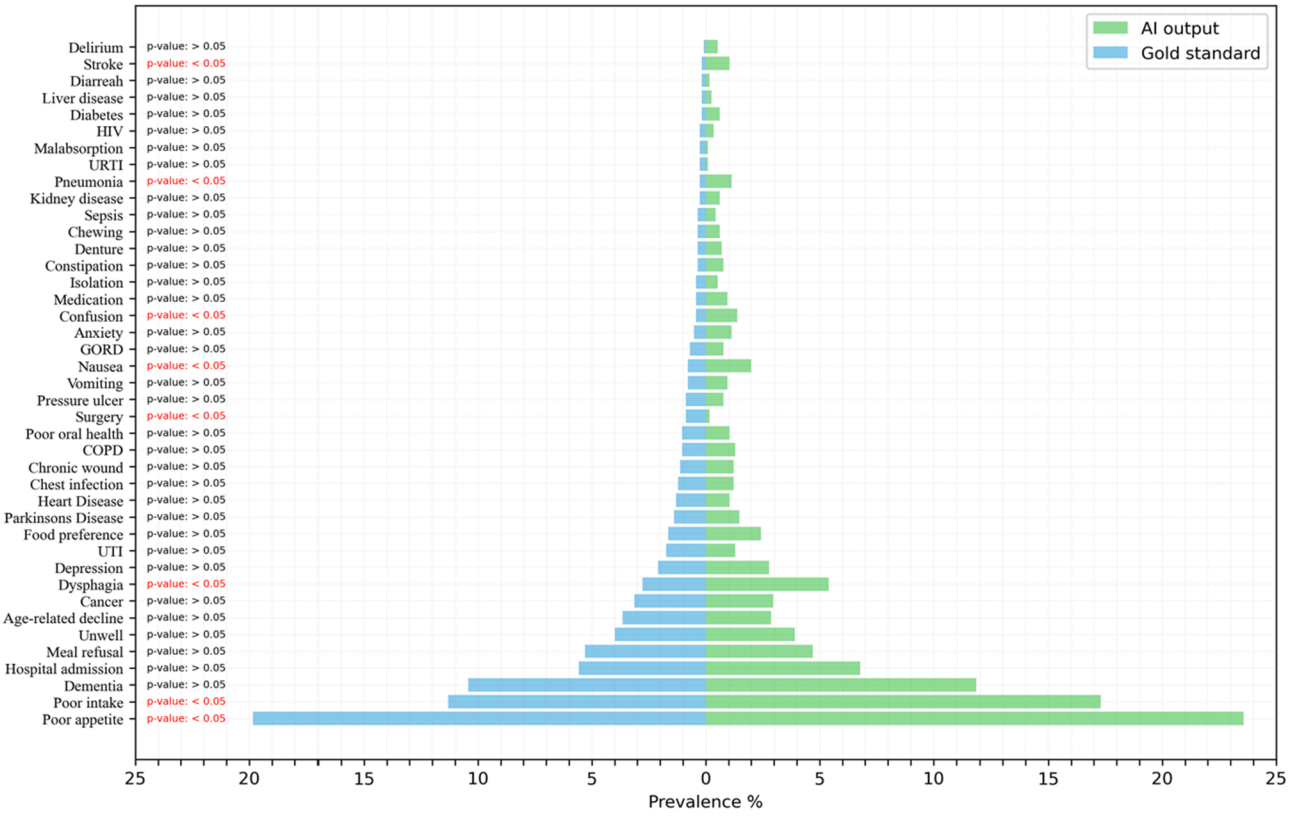


Fig. 3. Task 2 output: prevalence of malnutrition risk factors identified by the AI method in comparison with the gold standard. A p-value < 0.05 indicates that there is a statistically significant difference between the AI output and the gold standard in risk factors extracted.

malnutrition notes via zero-shot learning was noteworthy. It was able to accurately follow the prompt instructions to capture key information about malnutrition from each note. However, some challenges emerged, specifically when information of any part of the template was not explicitly mentioned in the note and required reasoning from the contextual information to derive the answer. For example, when age was not mentioned in the note, the model hallucinated and filled the age part of the template with the resident's weight (Supplementary Table S3). In another similar incident when the weight was not specified, the model hallucinated and filled the weight with a wrong number. Similarly, in a case where weight history was not mentioned, the model mistakenly substituted it with medical history instead of indicating that weight history was not specified.

After applying RAG, the model had access to an extended knowledge base, including additional nursing notes, demographic and weight tables. This access enabled the model to improve accuracy in certain generated responses, in age and gender where the model achieved 100% accuracy (Table 3, Supplementary Table S3). Additionally, the model was able to output more detailed information regarding weight, BMI and medical history, benefiting from access to more notes. However, it overlooked one intervention, one risk factor and one diagnosis. It is crucial to note that in these cases, the RAG retrieval has accurately retrieved the correct documents. The incomplete information output was caused by the limitations of the model itself, rather than the retrieval capability of RAG. The incorporation of RAG evidently enhances the model's accuracy. This is evident as removing RAG approach results in the model reverting to the same hallucinatory errors.

The model achieved satisfactory results in extracting the key risk factors for malnutrition from the free text nursing notes at individual client level. Similar to the use case of summarization, the model achieved outstanding results when risk factors were explicitly mentioned in the notes (95 %).

However, when the notes did not mention any risk factors, the performance dropped (85 %). The model tended to hallucinate and generate risk factors that were not mentioned in the note, such as "poor appetite" and "poor intake", rather than adhering to the prompt and stating, "no risk factor mentioned" (Supplementary Table S4). Otherwise, the identified prevalence of risk factors was similar, although not identical between the Llama 2 and gold standard (Fig. 3). We hypothesized that the observed variation may be attributed to our method of analyzing nursing notes. The gold standard is derived from the clinician judgement of dietary information documented in the clinical notes. Conversely, the AI model analyses the clinical documentation and identify all plausible risk factors, independent of clinician assessment. It is likely that the AI output will be further improved by refining the prompt and providing further specific instructions utilizing chain of thought prompt.

While the results were highly encouraging, these models still exhibited some limitations. A major concern in introducing generative AI and LLMs into high safety stake healthcare and medical settings is model hallucination, that is, generating plausible, yet unverified output [50–52]. We observed hallucinations in both tasks, similar to other's findings [31,53,54]. Generative models may face difficulties in generating outputs that are both coherent and contextually accurate. It may struggle with nuanced understanding, resulting in inconsistent and sometimes meaningless responses [55]. We have found that incorporating RAG can mitigate hallucination by providing access to more data, including structured data. However, this approach doesn't fundamentally solve the underlying issue within the generative model. When access to external data was revoked, the model reverted to incorrect outputs, indicating a reliance on that information. Another concern is inconsistency, as the model fails to produce consistent output when presented with minor variations in the prompt. This observation is consistent with what others have reported [49,56]. Moreover, even with the same prompt, the model's output varies from one note to another based on the information provided in each note. However, we found that giving the model a template to follow mitigated this issue. Alternative prompting methods such as few-shot learning, chain of thoughts or fine-tuning may address zero-shot learning limitations. These methods allow the model to learn from a few examples and refine its understanding and reasoning, which could potentially enhance the model's ability to generate more logical output while minimizing hallucinations [57,58].

In this study we only explored the zero-shot strategy for its simplicity. It is likely that other prompt engineering techniques could further improve the model performance. Incorporating few-shot examples into model generation is challenging for two reasons. First, the length of the notes posed a significant constraint, given the model's maximum length of 4096 tokens. Additionally, reliance on few-shot examples led to occasional errors in certain instances, making it difficult to create few-shot instructions that adequately cover all possible situations [59]. This struggle with contextual understanding and generalization suggests that the model may have memorized the specific examples provided, leading to a failure in its ability to generalize effectively across diverse contexts. To overcome this problem, the LLM might benefit from more explicit instructions through fine-tuning. Moreover, we used the smaller Llama 2 model (13B), which, as expected, exhibits lower performance than the larger Llama model with a higher number of parameters, e.g., Llama 2 70B [45]. Moreover, our exploration was limited to just one aspect of the RAG approach. It is noteworthy that the functionalities of a RAG system extend beyond what we have demonstrated here. There are diverse RAG capabilities and applications, such as creating chatbots for medical professionals and developing agents for various medical purposes [60].

Despite the prompting technique choice and LLM's smaller size, we were able to successfully replicate the results achieved in our previous work, which took weeks to complete [42]. The replication was achieved in just a few days, with a reasonable level of accuracy. The efficiency gains and practicality for use in real-world applications make this exploration worthwhile.

5. Conclusion

The latest advancements in large language models have opened many opportunities for health informatics in the near future [30]. This study demonstrates that by utilizing LLM and generative AI, we can successfully extract nutrition summaries and identify risk factors for malnutrition from unstructured nursing notes within aged care EHR, a task that might otherwise be difficult to complete. It also demonstrates that the integration of the RAG approach optimizes the utilization of data by LLMs in healthcare settings. This, in turn, will improve data accessibility and streamline the data analysis process. Furthermore, the application of NLP using LLMs has great potential for improving the quality of care and ensuring timely interventions. It will transform the way we address malnutrition and other health problems within healthcare and aged care settings.

CRediT authorship contribution statement

Mohammad Alkhalaf: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Ping Yu:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Mengyang Yin:** Formal analysis, Data curation. **Chao Deng:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The first author, Mohammad Alkhalaf, is supported by a full PhD scholarship from Qassim University, Saudi Arabia. The authors are grateful for the aged care organization that shared the de-identified electronic health records, which provided the opportunity to conduct this significant research project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2024.104662.

References

[1] M.I.T.D. Correia, D.L. Waitzberg, The impact of malnutrition on morbidity, mortality, length of hospital stay and costs evaluated through a multivariate model analysis, Clin. Nutr. 22 (3) (2003) 235–239, https://doi.org/10.1016/S0261-5614(02)00215-7.

[2] R.J. Stratton, et al., 'Malnutrition Universal Screening Tool' predicts mortality and length of hospital stay in acutely ill elderly, Br. J. Nutr. 95 (2) (2006) 325–330, https://doi.org/10.1079/bjn20051622.

[3] T. Ahmed, N. Haboubi, Assessment and management of nutrition in older people and its importance to health, Clin. Interv. Aging 5 (2010) 207–216, https://doi.org/10.2147/cia.s9664.

[4] E. Agarwal, et al., Malnutrition in the elderly: A narrative review, Maturitas 76 (4) (2013) 296–302, https://doi.org/10.1016/j.maturitas.2013.07.013.

[5] J. Kellett, R. Bacon, A. Simpson, Malnutrition prevalence in aged care residences, Nutr. Diet. 69 (2012) 72–164, https://doi.org/10.1111/j.1747-0080.2012.01611.x.

[6] K. Flint, et al., Mealtime care and dietary intake in older psychiatric hospital inpatient: A multiple case study, J. Adv. Nurs. 77 (3) (2021) 1490–1500, https://doi.org/10.1111/jan.14728.

[7] E. Fashho, et al., Investigating the prevalence of malnutrition, frailty and physical disability and the association between them amongst older care home residents, Clin. Nutr. ESPEN 40 (2020) 231–236, https://doi.org/10.1016/j.clnesp.2020.09.014.

[8] S. Sahin, et al., Prevalence of anemia and malnutrition and their association in elderly nursing home residents, Aging Clin. Exp. Res. 28 (5) (2016) 857–862, https://doi.org/10.1007/s40520-015-0490-5.

[9] L. Robb, et al., Malnutrition in the elderly residing in long-term care facilities: A cross sectional survey using the Mini Nutritional Assessment (MNA®) screening tool, South Afr. J. Clin. Nutr. 30 (2) (2017) 34–40, https://doi.org/10.1080/16070658.2016.1248062.

[10] K. Lind *et al.*, "Measuring the prevalence of 60 health conditions in older Australians in residential aged care with electronic health records: a retrospective dynamic cohort study," pp. 1-9, 2020, doi: DOI: 10.21203/rs.2.21384/v1.

[11] H. Kharrazi, et al., The value of unstructured electronic health record data in geriatric syndrome case identification, J. Am. Geriatr. Soc. 66 (8) (2018) 1499–1507, https://doi.org/10.1111/jgs.15411.

[12] T.B. Murdoch, A.S. Detsky, The inevitable application of big data to health care, JAMA – J. Am. Med. Assoc. 309 (13) (2013) 1351–1352, https://doi.org/10.1001/jama.2013.393.

[13] H.-J. Kong, Managing unstructured big data in healthcare system, Healthcare Inform. Res. 25 (1) (2019) 1–2.

[14] A. Mustafa, M. Rahimi Azghadi, Automated Machine Learning for Healthcare and Clinical Notes Analysis, Computers 10 (2) (2021) pp, https://doi.org/10.3390/computers10020024.

[15] A. Laxmisan *et al.*, "Clinical Summarization Capabilities of Commercially-available and Internally-developed Electronic Health Records," (in En), *Appl. Clin. Inform.*, vol. 03, no. 01, pp. 80-93, 2017/12/16 2012, doi: DOI: 10.1055/s-0037-1618556.

[16] G. Adams *et al.*, "What's in a summary? laying the groundwork for advances in hospital-course summarization," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, 2021, vol. 2021: NIH Public Access, p. 4794.

[17] J.S. Hirsch, et al., HARVEST, a longitudinal patient record summarizer, J. Am. Med. Informat. Assoc. 22 (2) (2014) 263–274, https://doi.org/10.1136/amiajnl-2014-002945.

[18] N.C. Favaro-Moreira, et al., Risk Factors for Malnutrition in Older Adults: A Systematic Review of the Literature Based on Longitudinal Data, Adv. Nutr. 7 (3) (2016) 507–522, https://doi.org/10.3945/an.115.011254.

[19] J.M. Steinkamp, et al., Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes, J. Biomed. Informat. 102 (2020) 103354, https://doi.org/10.1016/j.jbi.2019.103354.

[20] C. Serón-Arbeloa, et al., Malnutrition screening and assessment, Nutrients 14 (12) (2022) 2392, https://doi.org/10.3390/nu14122392.

[21] J.M.M. Meijers, et al., Malnutrition prevalence in The Netherlands: results of the Annual Dutch National Prevalence Measurement of Care Problems, Br. J. Nutr. 101 (3) (2008) 417–423, https://doi.org/10.1017/S0007114508998317.

[22] L.A. Barker, B.S. Gout, T.C. Crowe, Hospital Malnutrition: Prevalence, Identification and Impact on Patients and the Healthcare System, Int. J. Environ. Res. Public Health 8 (2) (2011) 514–527. https://www.mdpi.com/1660-4601/8/2/514.

[23] J. Song, et al., Uncovering hidden trends: identifying time trajectories in risk factors documented in clinical notes and predicting hospitalizations and emergency department visits during home health care, J. Am. Med. Inform. Assoc. (2023) p. ocad101.

[24] M. Topaz, et al., Home health care clinical notes predict patient hospitalization and emergency department visits, Nurs. Res. 69 (6) (2020) 448.

[25] Y. Hu, et al., Improving large language models for clinical named entity recognition via prompt engineering, J. Am. Med. Inform. Assoc. (2024), https://doi.org/10.1093/jamia/ocad259.

[26] R. Pivovarov, N. Elhadad, Automated methods for the summarization of electronic health records, J. Am. Med. Inform. Assoc. 22 (5) (2015) 938–947, https://doi.org/10.1093/jamia/ocv032.

[27] I. Li, et al., Neural Natural Language Processing for unstructured data in electronic health records: a review, Comput. Sci. Rev. 46 (2022), https://doi.org/10.1016/j.cosrev.2022.100511.

[28] Y. Wang, et al., Clinical information extraction applications: a literature review, J. Biomed. Inform. 77 (2018) 34–49, https://doi.org/10.1016/j.jbi.2017.11.011.

[29] J. A. Banan, F. Chia Aziz, and H. Mzhda Yasin, "A Review of the Role and Challenges of Big Data in Healthcare Informatics and Analytics," (in English), *Computational Intelligence and Neuroscience : CIN,* vol. 2022, 2022 2022, doi: DOI: 10.1155/2022/5317760.

[30] P. Yu, et al., Leveraging generative ai and large language models: a comprehensive roadmap for healthcare integration, Healthcare 11 (20) (2023) pp, https://doi.org/10.3390/healthcare11202776.

[31] D. Van Veen *et al.*, "Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts," p. arXiv:2309.07430doi: DOI: 10.48550/arXiv.2309.07430.

[32] J. Liu, C. Wang, S. Liu, Utility of ChatGPT in Clinical Practice, J. Med. Internet Res. 25 (2023) e48568.

[33] Hugo Touvron *et al.*, "Llama 2- Open Foundation and Fine-Tuned Chat Models," *arXiv,* vol. 2307, 2023, doi: DOI: 10.48550/arXiv.2307.09288.

[34] V. K. Cody Bumgardner *et al.*, "Local Large Language Models for Complex Structured Medical Tasks," p. arXiv:2308.01727doi: DOI: 10.48550/arXiv.2308.01727.

[35] A. Toma *et al.*, "Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding," p. arXiv:2305.12031doi: DOI: 10.48550/arXiv.2305.12031.

[36] H. Wang *et al.*, "DRG-LLaMA : Tuning LLaMA Model to Predict Diagnosis-related Group for Hospitalized Patients," p. arXiv:2309.12625doi: DOI: 10.48550/arXiv.2309.12625.

[37] T. Brown, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[38] P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Adv. Neural Inf. Process. Syst. 33 (2020) 9459–9474.

[39] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," p. arXiv:2312.10997doi: DOI: 10.48550/arXiv.2312.10997.

[40] Y. Mao *et al.*, "Generation-Augmented Retrieval for Open-domain Question Answering," p. arXiv:2009.08553doi: DOI: 10.48550/arXiv.2009.08553.

[41] W. E. Thompson *et al.*, "Large Language Models with Retrieval-Augmented Generation for Zero-Shot Disease Phenotyping," *arXiv e-prints,* p. arXiv:2312.06457, 2023, doi: DOI: 10.48550/arXiv.2312.06457.

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

[43] T. Wolfe, et al., Transformers: state-of-the-art natural language processing, EMNLP (systems Demonstrations) (2020) 38–45, https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[44] X. Liu *et al.*, "Prompting Frameworks for Large Language Models: A Survey," *arXiv e-prints,* p. arXiv:2311.12785, 2023, doi: DOI: 10.48550/arXiv.2311.12785.

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

[46] T. Dettmers *et al.*, "QLoRA: Efficient Finetuning of Quantized LLMs," p. arXiv:2305.14314doi: DOI: 10.48550/arXiv.2305.14314.

[47] C. Harrison. "LangChain." https://github.com/langchain-ai/langchain (accessed 10 OCT, 2023).

[48] J. Maynez *et al.*, "On faithfulness and factuality in abstractive summarization," *arXiv preprint arXiv:2005.00661,* 2020.

[49] N. Bhate *et al.*, "Zero-shot Learning with Minimum Instruction to Extract Social Determinants and Family History from Clinical Notes using GPT Model," p. arXiv:2309.05475doi: DOI: 10.48550/arXiv.2309.05475.

[50] J. Kaddour *et al.*, "Challenges and Applications of Large Language Models," p. arXiv:2307.10169doi: DOI: 10.48550/arXiv.2307.10169.

[51] Y. Bang *et al.*, "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity," p. arXiv:2302.04023doi: DOI: 10.48550/arXiv.2302.04023.

[52] V. Rawte, A. Sheth, and A. Das, "A survey of hallucination in large foundation models," *arXiv preprint arXiv:2309.05922,* 2023.

[53] H. Alkaissi and S. I. McFarlane, "Artificial hallucinations in ChatGPT: implications in scientific writing," *Cureus,* vol. 15, no. 2, 2023.

[54] J. Wang *et al.*, "NoteChat: A Dataset of Synthetic Doctor-Patient Conversations Conditioned on Clinical Notes," p. arXiv:2310.15959doi: DOI: 10.48550/arXiv.2310.15959.

[55] K. Singhal, et al., Large language models encode clinical knowledge, Nature 620 (7972) (2023) 172–180, https://doi.org/10.1038/s41586-023-06291-2.
[56] Y. Hu *et al.*, "Zero-shot Clinical Entity Recognition using ChatGPT," p. arXiv: 2303.16416doi: DOI: 10.48550/arXiv.2303.16416.
[57] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403,* 2022.
[58] J. Wei, et al., Chain-of-thought prompting elicits reasoning in large language models, Adv. Neural Inf. Process. Syst. 35 (2022) 24824–24837.
[59] M. Mosbach *et al.*, "Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation," *arXiv preprint arXiv:2305.16938,* 2023.
[60] C. Wang, et al., Potential for GPT technology to optimize future clinical decision-making using retrieval-augmented generation, Ann. Biomed. Eng. (2023), https://doi.org/10.1007/s10439-023-03327-6.