Department of Informatics
King's College London
United Kingdom

7CCSMPRJ Individual Project

C-QuAL – A Clinical Question Answering Benchmark for Long-Context Large Language Models

---

Name: **Alfie Lamerton**
Student Number: K23080018
Course: Artificial Intelligence

Supervisor: Prof. Yulan He

This dissertation is submitted for the degree of MSc in Artificial Intelligence.

Acknowledgements

I would like to thank Professor Yulan He and Tao Wang for encouraging and guiding me on the production of this report, Hainiu Xu for giving me feedback on my dataset generation approach, and David Codling and Dipen Patel for providing me with their insights and experience with the clinical setting.

Abstract

The question-answering capabilities of large language models (LLMs) is improving rapidly. These capabilities have opened opportunities for automation and assistance in healthcare. LLMs show promise in clinical decision reasoning – the organisation, summary, identification, and retrieval of clinical text from electronic health records (EHR). LLMs for clinical decision reasoning must be benchmarked on their capabilities to be selected for clinical question answering (QA). This report documents the development of C-QuAL – a relevant and representative clinical question answering benchmark for long-context LLMs and evaluates its effectiveness on a number of natural language tasks.

Nomenclature

LLM	Large language model
Q-A pair	Two strings containing a question and answer to that question
QA	Question answering

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Large language models (LLMs) have recently started showing impressive question answering (QA) capabilities [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11],. These capabilities have opened up the possibility of developing new applications in healthcare. In the clinical setting, clinicians spend and energy manually navigating, retrieving, and organising information in electronic health records (EHRs).

Many of these tasks can be automated thanks to the abundance of clinical data in corpora such as the n2c2 and MIMIC research datasets [12, 13]; specifically, the manual analysis of EHRs when working with patients. LLMs can be fine-tuned to assist clinicians with decision reasoning processes by automating many of the tasks related to information retrieval and organisation, reducing the amount of time clinicians need to spend processing the information themselves.

To select an LLM for the task of assisting clinicians with decision reasoning, the capabilities of the LLM must be evaluated. This enables the most appropriate model to be chosen from a group of candidate models via analysis and comparison. Benchmarking datasets effectively evaluate LLM capabilities by testing the LLM on representative examples of prompts from the use case for which the model is designed. In terms of clinical decision reasoning, that use case is question answering. This report documents the development of a novel benchmarking dataset for clinical question answering: **C-QuAL**.

The main contributions of this research project are:

1. C-QuAL – a **C**linical **Qu**estion **A**nswering **L**arge language model benchmarking dataset
2. A dataset generation framework for automatically constructing datasets like C-QuAL
3. This report on the rationale and methodology behind the creation of the dataset

1.2 Aims and Objectives

1.2.1 Problems

The broad problem this project aims to address is that clinicians use their time and cognitive energy to manually navigate, retrieve, and organise information in EHRs for decision reasoning, when many of these tasks can now be automated by LLMs. LLMs have been fine-tuned to assist with clinical decision reasoning by automating these tasks, and clinical QA benchmarking datasets have been created to effectively evaluate these LLMs. But the existing datasets have limitations; they are not linguistically rich, not

representative of clinical contexts, and do not leverage the full extent of the utility of LLMs in producing high-quality question-answer pairs (Q-A pairs).

1.2.2 Broad Project Aim

The aim of the LLM for clinical decision reasoning project is to create an LLM which can be deployed in clinical settings to reduce the time clinicians spend manually carrying out information retrieval tasks on EHRs. The hypothesis behind this aim is that saving clinicians' time could have a positive effect on the efficiency of healthcare systems.

1.2.3 The Aim of This Variant

A clinical QA benchmarking dataset is required to validate such an LLM. Benchmarking a QA LLM allows for the assessment of the strengths and limitations of the model before deployment. This assessment is useful for informing model choice. It therefore follows that the quality of the chosen QA benchmarking dataset is important for achieving an accurate model evaluation. Clinical QA benchmarking datasets exist, but have limitations.

This variant aims to provide as good an evaluation as possible to the LLM produced by the LLM for clinical decision reasoning project, so that model can be evaluated as effectively as possible prior to deployment. The variant also aims to contribute a new dataset which addresses the issues identified in the existing datasets, making it more clinically relevant in the following ways:

- Leveraging chat completion LLMs to generate questions and answers based on patient disharge summaries instead of using automatic logical-form-based generation templates, increasing the lexical richness and representativeness of Q-A pairs of the free-form nature of clinical settings;
- Drawing de-identified patient discharge summaries from MIMIC-III – a medical corpus more up to date and representative than corpora such as n2c2;
- Leveraging long-context LLMs to draw questions and answers from multiple discharge summaries, increasing the possible length of the discharge summaries used for Q-A pair generation, further increasing the representativeness of real clinical use cases with EHRs.

The broad aim of this variant is to contribute a new dataset to the natural language processing (NLP) community which can be used to benchmark existing or new clinical LLMs, with the goal of contributing to the improvement of healthcare systems by providing them with LLMs which have been evaluated as effectively as possible.

1.2.4 The Objective of This Variant

The objective of this variant is to produce a new clinical QA benchmarking dataset which is richer and more representative than the existing datasets, and use it to benchmark

the LLM being developed in the LLM for clinical decision reasoning project as well as QA LLMs more broadly. It is also an objective of this variant to make the dataset, and the methodology for creating and generating the dataset (or other QA datasets using other language data) with code, publicly available and accessible so that it can be used or improved further by the community.

1.3 Background and Literature Survey

The following section documents the relevant literature regarding the potential of LLMs in the assistance of clinical decision reasoning, the importance of benchmarking in the deployment of LLMs in real-world contexts, the existing solutions for benchmarking clinical LLMs, and the justification for the proposal of C-QuAL in light of this literature.

1.3.1 Clinical Question Answering

The aim of clinical QA is to assist with clinical processes by providing answers to clinical questions, a capability previously reserved for humans. Within this aim are two distinct aims; to identify and retrieve text from a patient note in answer to a question [10], and to automate the clinical analysis of this text in order to inform clinical decision reasoning.

1.3.2 The Potential of LLMs in Clinical Contexts

LLMs have immediate applications in clinical decision reasoning, but the applications are limited to the retrieval, organisation, and summarisation of text due to the current lack of explainability, interpretability, and validation methods.

[14] explore the application of LLMs in encoding clinical knowledge and helping with medical QA using medical QA benchmarks, finding that medical QA may be an emergent capability in LLMs. [15] provide an overview of the applications of LLMs in the domain of healthcare, suggesting QA LLMs applied in healthcare may help medical professionals with information retrieval in clinical notes. [16] develop an LLM for clinical tasks such as medical QA, suggesting that such LLMs can be utilised for clinical decision reasoning assistance. [17] claims LLMs could be an innovative solution for augmenting clinical decision-making processes due to their ability to analyse complex knowledge and big data, and that clinicians could use LLMs to assist with patient diagnosis.

[18] claim LLMs show promise for compiling patient notes and supporting clinical decision making, but do not express confidence in LLMs automating decision reasoning tasks without oversight. They claim the clinical application of LLMs remains problematic due to data bias, inaccurate information, and ethical concerns. Lastly, they suggest there is a lack of methods for evaluating the clinical utility of LLMs. [19] state the potential for LLMs in documentation management and literature summarisation, but express caution with applying LLMs in diagnosis, clinical decision support, and patient triage until a robust validation process is available. [20] critically analyse the potential application

of more agentic clinical LLMs in clinical workflows, suggesting LLMs are capable of influencing clinical decision-making, and that robust, representative evaluations of these LLMs are important.

The architecture of LLMs rely on neural networks which are inherently uninterpretable and unexplainable. There is a body of work attempting to interpret and explain LLMs and other machine learning models with architectures based on neural networks ([21, 22, 23], and some work has attempted to address the lack of interpretability in LLMs in clinical settings [24, 25]. While LLMs seem to be able to process and reason with expert clinical knowledge [26, 27], clinical safety measures make it difficult to deploy unexplainable and uninterpretable models in clinical contexts [28].

1.3.3 The Importance of Benchmarking an LLM on QA Tasks

In the context of LLMs, benchmarking refers to the process of evaluating and comparing the performance of LLMs on specific NLP tasks (such as QA, or language understanding). The purpose of this benchmarking is to assess the capability of an LLM in relation to the given tasks. Benchmarks illuminate the strengths and weaknesses of LLMs. Benchmarks can also be used reliably to test different LLMs, meaning the performance of multiple LLMs can be compared based on one benchmark, allowing for the comparison of multiple LLMs on specific task types.

Benchmarks are typically structured as datasets containing prompts designed to evaluate specific capabilities. The structure of the dataset depends on the language task being benchmarked. For QA benchmarking, datasets usually contain rows of Q-A pairs consisting of questions designed to target specific model capabilities followed by an exemplar answer. Some QA benchmarking datasets also contain source material given to the LLM as context for answering the question.

Benchmarking LLMs for clinical decision reasoning tasks is important because of the potential scale of the impact LLMs may have on healthcare. A large number of clinicians may use LLMs to support their decision reasoning in the future, so small differences in the quality of the LLMs that are deployed may have significant consequences on patient outcomes. Generally, benchmarking is useful for informing the future development of models because:

- Future model development can be informed by benchmarking, as benchmark results highlight areas where model capabilities can be improved;
- Benchmark results are somewhat repeatable, such as through results averaging, although not completely [29];
- Model results can be used to compare the performance of multiple models to inform the appropriate model choice for a given task.

index	Discharge Summary	Question	Expected Answer	LLM Answer
0	...	Did the patient require...	No, the patient...	No, the patient...
1	...	Was the patient...	Yes, the patient...	Yes...
2	...	What was the reason for...	The patient underwent...	The patient...
3	...	What were the treatment...	The patient was...	The treatment...
4	...	Did the patient have...	No, the patient...	Yes, there was...

Figure 1: Example QA dataset structure

In the context of clinical QA using EHRs, LLMs should be evaluated on their ability to answer questions related to the content of the EHR in order to assess their ability to summarise, identify, and arrange the text and data within the EHRs.

1.3.4 Existing Solutions

There is a substantial body of datasets designed to contribute to solving the broad problem identified above: clinicians spending time and energy manually navigating, retrieving, and organising information in EHRs for decision reasoning. However, these datasets have limitations.

1.3.4.1 emrQA

In their 2018 paper, [1] present a QA benchmarking dataset for QA tasks using EHRs derived from the n2c2 NLP research datasets [12], emrQA, to address the problem of physicians manually seeking answers to questions from electronic medical records (EMRs). emrQA was the first publicly available EMR QA dataset, created by repurposing existing annotations for other NLP tasks. The authors provide the dataset, their methodology, a generation framework for constructing a QA dataset in any domain, and two QA challenges.

Produced in 2018 before the advent of LLMs, emrQA is now outdated. The dataset generation framework relies on logical form templates and automated QA generation scripts. This process can now be automated by LLMs, which pose an advantage in terms of the lexical richness of the QA pairs they are capable of generating without the need for the manual logical programming of QA generation.

The emrQA dataset was also built using outdated research challenge data that is not representative of real-world clinical contexts. The data is mainly drawn from medical exams and online forums, which does not accurately represent clinical use cases. More relevant clinical corpora are available for the construction of QA datasets.

1.3.4.2 EHRNoteQA

[2] echo this criticism, claiming current clinical LLM evaluation methods are not realistic or representative of the clinical context because they usually focus on single-note information and limited topics. They present EHRNoteQA, the first clinical QA dataset to leverage LLMs for the generation of QA pairs, addressing the issues highlighted by providing a dataset, methodology, and generation framework. The dataset uses a more up-to date medical corpus, MIMIC-IV [30], leverages LLMs for the production (and some of the annotation) of Q-A pairs, and creates 8 Q-A pair categorisations.

The authors also highlight the current challenge of LLMs processing contexts longer than 8,000 tokens. This limits the number and size of discharge summaries an LLM can be given to generate a Q-A pair or answer a question. They address this by removing whitespace and categorising discharge summaries on length to leverage LLMs with different context lengths to reduce generation costs.

The authors suggest LLMs with larger context windows could be leveraged to produce Q-A pairs based on multiple, longer discharge summaries. This could pose an advantage to their existing dataset as the ability of an LLM to process multiple long notes would be more representative of the clinical use case.

1.3.4.3 EHR-DS-QA

[3] also leverage LLMs for the generation of a clinical Q-A benchmarking dataset using the MIMIC-IV corpus. However, they use a smaller LLM than [2] (Llama-2-13b) which decreases the complexity and diversity of the Q-A pair, limit the discharge summary length to 6,000 tokens which reduces the number of discharge summaries that can be used to generate Q-A pairs, and only pass a single discharge summary to the LLM for generating a Q-A pair, limiting the representativeness of the dataset of real-world clinical contexts where patient information spans multiple EHR notes.

1.3.4.4 emrKBQA

[4] present emrKBQA, a clinical QA dataset for answering physician questions based on EHRs using MIMIC-III data (Johnson et al. 2016), re-purposing the emrQA logical form templates and extending them to include an SQL query which retrieves an answer from a knowledge base. Their dataset aims to improve on the emrQA dataset by providing a solution for extracting QA pairs from the structured part of EHRs, not just the clinical

notes. They also provide better question quality and diversity than emrQA thanks to entity filtering and updated paraphrase groups. However, the question set is limited to the knowledge of the physicians polled, which has the potential to introduce sample bias. Also, the automated question generation creates unrealistic questions compared with the capability of modern LLMs on the same task. Lastly the generated QA pairs are limited in diversity due to the slot filling nature of the automated generation process.

1.3.4.5 DrugEHRQA

[5] present the first QA dataset for multi-model EHRs, containing QA pairs from structured tables and unstructured clinical notes from MIMIC-III. The diversity of their questions are limited by the type of relations extracted from the dataset.

1.3.4.6 EHRXQA

[6] introduce another multi-modal QA dataset using EHRs, drawing from structured EHR data from MIMIC-III and X-ray images.

1.3.4.7 MedQA

[7] present MedQA, the first free-form multiple choice medical QA dataset with versions in English, Chinese, and Simplfied Chinese. They draw data from medical board exams, causing their dataset to lack representativeness of clinical contexts. The multiple choice QA format also limits the capacity of the dataset to benchmark LLMs as it is possible for the LLM to return the correct answer accidentally.

1.3.4.8 PubMedQA

[8] draw from biomedical literature abstracts to produce a biomedical QA dataset, PubMedQA. The dataset contains questions, contexts, a long answer, and a yes/no/maybe answer, against which models are scored.

1.3.4.9 RxWhyQA

[9] produce a dataset focusing on 'why questions', targeting the reading comprehension capabilities of QA models. They include questions with zero and multiple answers, which test the ability of the model to respond that a question is unanswerable, increasing the robustness of the benchmark. They draw from the n2c2 challenge dataset [12] which is not as representative of clinical contexts as corpora such as MIMIC-III and MIMIC-IV [30] and thus limits the quality of their QA pairs which rely on the quality of the origin data. They also identified that their drug-reason relations are not linguistically diverse.

1.3.4.10 CliniQG4QA

[10] find that QA models do not generalise well to clinical texts from different insitutions or groups of patients. They aim to address the lack of diversity in existing QA benchmarking datasets, using question phrase prediction to predict potential question phrases based on answer evidence to gather the kind of question that may be typically asked by a clinician about the evidence. Their questions improve clinical QA model performance using their dataset, highlighting the importance of question diversity in generating QA pairs for benchmarking. This dataset preceded the advent of LLMs, suggesting the potential for the creation of diverse question sets by leveraging the generative potential of LLMs.

1.3.4.11 MedMCQA

[11] provide a multi-subject multiple choice medical question answering dataset designed to be representative of real-world medical entrance exam questions. While it is another good QA dataset it is less relevant as question types are more representative of exam questions than real-world clinical questions.

In light of the existing contributions, the objectives of this project become clear: there is a gap in the research for a clinical QA benchmarking dataset that leverages long-context LLMs to test the QA capabilities of LLMs for answering clinically relevant and representative questions drawing from multiple patient discharge summaries.

Leveraging LLMs with longer context windows is a natural next step following EHRNoteQA's utilisation of LLMs and multiple discharge summaries. The data shows that due to clinical safety concerns, the clinical field is not prepared to entrust decisions to LLMs directly, and so the optimal application is applying LLMs to the manual tasks which cost a significant amount of clinician time.

2 Objectives, Specification and Design

The objective of this project is to produce a new clinical QA benchmarking dataset which is and more representative and clinically relevant than the existing datasets, with the broad goal of saving the time of clinicians, thus contributing to the improvement of the efficiency of healthcare systems. In light of the existing solutions, it becomes clear that there is a gap in the research for a clinical QA benchmarking dataset that leverages long-context LLMs to test the QA capabilities of LLMs for questions regarding treatment, assessment, diagnoses, problems and complications, abnormalities, etiology, and medical history, drawing from multiple discharge summaries of any length. Such an implementation could result in the production of a dataset that is maximally clinically relevant and representative.

2.1 Evaluation of Existing Solutions

One project objective is for the proposed dataset to be more representative of real world clinical use cases than the existing solutions. emrQA, emrKBQA, drugEHRQA, EHRXQA, MedQA, RxWhyQA, CliniQG4QA, and MedMCQA all fall short of this objective due to being constructed from logical templates or other implementations that do not focus on generative machine learning models. EHRNoteQA and EHR-DS-QA are the only solution which improves on this by using LLMs to produce QA pairs. The other datasets each address representativeness, but LLMs provide a significant improvement on this objective compared with other automated implementations.

Dataset	QA Pairs	Source Corpus	Released	Creation Method	Availability
emrQA	400,000	i2b2 annotations	2018	Question templates	Public
EHRNoteQA	962	MIMIC-IV Discharge Summaries	2024	LLM	Credentialed
EHR-DS-QA	156,599	MIMIC-IV	2024	LLM	Credentialed
emrKBQA	940,000	MIMIC-III	2021	Question templates	Public
DrugEHRQA	70,000	MIMIC-III	2022	Question templates	Credentialed
EHRXQA	36,174	MIMIC-IV and MIMIC-CXR	2023	Question templates	Credentialed
MedQA	61,097	Medical board exams	2020	Manual	Public
PubMedQA	273,500	PubMed abstracts	2019	Question templates	Public
RxWhyQA	96,939	n2c2 corpus	2022	Question templates	Public
CliniQG4QA	967,514	MIMIC-III	2021	Answer evidence extractor algorithm	Proprietary
MedMCQA	193,155	Medical entrance exams	2022	–	Public

Table 1: Summary of the specifications of existing clinical and medical QA datasets

2.2 Interviews

Details of the clinical use case must be examined for the derivation of the requirements of a clinical QA benchmarking dataset that are relevant, as clinical LLMs must be

benchmarked on their capabilities relative to the demands of the use case. To conduct this examination, two clinicians were interviewed about the details of clinical decision reasoning. The interviews resulted in information that aligns with the literature:

- Data quality and style can vary based on the experience of the clinicians creating it
- Discharge summaries are the most relevant source for evaluating AI models
- From the perspective of clinical safety, we are not ready to let an LLM automate diagnosis and decision making, but we are ready to use it to summarise and organise information
- Multiple choice questions are not representative of the clinical use case. Most questions are open-ended.
- Multiple document retrieval would be more useful than single, because on a single document you can find information quickly (e.g., with CTRL + F)
- What matters is how well the LLM interprets the question. That's what the benchmark should be assessing. How well the LLM answers questions with nuanced answers, that leverage its competitive advantage over keyword searching. CTRL + F 'blood test' may not bring up all the blood tests the patient has had, because that is not how EHR is structured.
- Crucial information includes diagnoses, medications, what worked or did not work in the past, and any documented risks

2.3 Requirements

The literature and interviews highlighted the importance of clinical relevance and representativeness. It follows that a set of requirements must be produced in order to inform the development of specifications for the proposed dataset, which inform the development of the dataset itself. Requirements are specific, granular objectives that a software artefact must meet in to meet the broad goal of the project. The broad goal of this project is to produce a clinical QA benchmarking dataset that effectively evaluates clinical QA models. For the resulting dataset must be clinically relevant and representative to meet this goal.

2.3.1 Validating Clinical Relevance and Representativeness

For the proposed dataset to be clinically relevant, it must effectively benchmark models that are likely to be utilised in the clinical use case. For a model to be used, it must pose an advantage to clinicians without also posing safety concerns. If a clinical QA model is unsafe, it will not be adopted, and thus a benchmarking dataset that evaluates models on capabilities deemed as not yet clinically safe, such as automating clinical decision

reasoning, it will not be relevant to the clinical use case.

It has been identified in the interviews and analysis of existing solutions that the proposed dataset should be designed to test an LLM's ability to summarise, identify, and arrange text, and answer specific questions, related to **treatments**, **assessments**, **diagnoses**, **problems**, **complications**, **abnormalities**, **etiology**, **medical history**; but not test the LLM's ability to **interpret medical data**, **make recommendations**, or **make diagnoses** to ensure that the dataset is only testing for capabilities that are clinically relevant at present. LLMs could become more involved with the clinical decision reasoning process in future if progress in interpretability and explainability pave the way for the safe application of black box models in clinical use cases. But designing a dataset which benchmarks models for these capabilities may unfairly bias relevant models under selection for application now.

Another method for ensuring clinical relevance is expert annotation. This process involves the evaluation of a subset or complete set of Q-A pairs by qualified medical professionals. The purpose of clinical expert annotation is to check that the Q-A pairs being used to evaluate models are accurate, relevant, and representative. Variations of this process have been implemented in the emrQA, EHRNoteQA, emrKBQA, EHRXQA, MedQA, PubMedQA, and CliniQG4QA datasets. Providing the dataset generation process is consistent, clinical experts validating the quality of a subset of Q-A pairs can provide a useful evaluation of the the quality of the overall dataset in terms of accuracy, clinical relevance, and clinical representativeness.

2.3.2 Clinical Corpus

To generate Q-A pairs for benchmarking clinical capabilties, it is a requirement that the Q-A pairs must be drawn from some existing clinical data source to be useful. Some datasets, such as emrQA and RxWhyQA, draw from the i2b2 n2c2 National NLP Clinical Challenges dataset, but this corpus has since been criticised for not capturing the diversity and complexity of EHRs.

More recent datasets such as EHRNoteQA, emrKBQA, EHRXQA, CliniQG4QA, and DrugEHRQA, draw from MIMIC databases. All except EHRNoteQA draw specifically from MIMIC-III. EHRNoteQA uses MIMIC-IV, which provides a more modular data organisation structure, larger quantity of patient patients and admissions, and more recent data entries. The dataset selected for the Q-A pair generation of this project is MIMIC-III. The corpus contains 59,652 discharge summaries from the Beth Israel Deaconess Medical centre from 2001 to 2012. A discharge summary is part of a the EHR for a patient. A discharge summary is a clinical document produced by clinicians at the end of a patient visit. The purpose of a discharge summary is to communicate all of the important information regarding the care and treatment of the patient to future clinicians.

Discharge summaries are useful for generating Q-A pairs. Discharge summaries contain patient information that is relevant to the clinical use case as identified previously, i.e., treatments, assessments, diagnoses, problems, complications, abnormalities, etiology, and medical history, because they are a record of all notes collated over the span of the visits of a patient. Discharge summaries are also highly representative of EHRs because they cover all of the relevant information that clinicians look for when processing them, and contain timestamps for temporal reasoning. The proposed dataset is designed to benchmark the capabilities of models processing information in EHRs, so the more representative the source corpus is of the intended use case, the more effective the resulting benchmarking dataset will be.

This section decomposes the broad aim of the proposed dataset, identifying the objectives the dataset must achieve to fulfil that aim and the features that will implement the objectives

2.4 Specifications

The requirements decompose the broad aim of the proposed dataset into granular features. The specifications can thus be created from the requirements. The creation of specifications involves defining the features of the proposed dataset in terms of the features derived from the specifications.

2.5 Design

2.5.1 Structure

The format of QA benchmarking datasets typically follows a {'Evidence', 'Question', 'Correct Answer'} format. In this case, the LLM will also be prompted to provide a reason for the answer it gave based on the evidence (discharge summary). Therefore, the relevant resulting structure for the proposed dataset is {'Discharge Summary', 'Question', 'Correct Answer', and '[Model Name] Answer'}. The QA data can be parsed from rows of this structure and nested in a prompt verbally detailing the requirements of the benchmark to the LLM.

The question types and categories selected for the dataset were informed by primary research and interviews. The questions aim to be most representative of the questions clinicians would ask in real-world contexts. They also aim appropriately challenge the model being benchmarked, meaning the answers cannot be too easy to extract from the discharge summary, making the benchmark ineffective for LLM evaluation. Lastly, the questions must mitigate spurious cues and correlations, to prevent the LLM from identifying correlations which unfairly improve the chances of the LLM to answer correctly.

Your task is to answer a clinical question based on the following discharge summary:
{Discharge Summary}

You should give an answer and a reason for your answer in the following format:

Answer: [your answer]
Question: {Question}

Answer:

Figure 2: An Example of How the Dataset Columns are Parsed into the LLM Prompt

2.5.2 Question Types

The question types were selected to robustly test the capabilities of clinical LLMs to provide an optimally effective benchmark. In this context, 'type' refers to the semantic format of a question in terms of the kind of answer it entails as a response. Question type is an orthogonal question specification to question category: it does not entail anything about the clinical detail of the question, only the lexical properties of the question. The following question types were selected for the proposed dataset: **1) yes/no/maybe**, **2) unanswerable**, **3) temporal**, **4) factual**, **5) Summarisation**, and **6) Identification**

2.5.2.1 Yes/No/Maybe Questions Questions of this type allow for robust evaluation by targeting questions at specific details from discharge summaries. The inclusion of 'maybe' as an answer increases representativeness for questions where the answer is not conclusive from the information available in the discharge summary.

2.5.2.2 Unanswerable Questions The decision to include unanswerable questions was inspired by [9] and originated in SQuAD 2.0 [31]. The reason for including unanswerable questions is to verify the LLM is not responding with an equivalent of a guess answer, thus potentially achieving correct answers without successfully processing the prompt.

2.5.2.3 Temporal Questions Temporal questions are included to test the ability of the LLM to identify events in the patient's EHR and answer questions which require the understanding of time. An LLM that cannot reason accurately about information

related to events across time may respond with information that is irrelevant or no longer true about the patient.

2.5.2.4 Factual Questions Factual questions evaluate the ability of the LLM to retrieve facts about a patient from a discharge summary. This ability enables clinicians to retrieve facts about a patients without manually reading discharge summaries. The question is also important for testing the ability of the LLM to demonstrate understanding of synonyms in clinical data. For example, if the LLM is asked if the patient has had tests, and the discharge summary contains records of 'assessments', the question tests whether the LLM is capable of identifying that the patient has had tests.

2.5.2.5 Summarisation Questions These questions evaluate the ability of the LLM to condense vast data into concise summaries. This ability is crucial for LLMs deployed in clinical settings where clinicians will primarily use the LLM for organising information. Therefore evaluating this ability is required for a representative benchmark.

2.5.2.6 Identification Questions Lastly, identification questions are included in the proposed dataset to evaluate the ability of the LLM to identify important information.

2.5.2.7 Excluded Question Types Diagnostic questions were not chosen for the proposed dataset because they would test the LLM on its ability to simulate and automate clinical decision reasoning, which has been identified as irrelevant for the clinical use case at present.

These question types were chosen to maximise the robustness of the dataset for benchmarking various LLM abilities. The inclusion of diverse, relevant question types increases the range of abilities the LLM is tested for. The selected question types help maximise the effectiveness of the dataset by making its evaluation of LLMs as well-rounded as possible.

2.5.3 Question Categories

Question categories were chosen to target specific clinical reasoning and processing abilities of LLMs. Question categories specify the clinical details targeted by the question, and thus focus on the clinical capabilities of the LLM rather than the question-answering capabilities. The following question categories were selected for the proposed dataset: **1) treatment**, **2) assessment**, **3) diagnosis**, **4) problem or complication**, **5) abnormality**, **6) etiology**, and **7) medical history questions**.

2.5.3.1 Treatment Questions

Questions regarding treatments that a patient has had based on their discharge summary

are included for the importance of past treatments for informing future clinical decisions regarding a patient.

2.5.3.2 Assessment Questions
Questions related to patient assessments, e.g., blood tests are also important for informing clinical decisions.

2.5.3.3 Diagnosis Questions
Questions related to the past diagnoses of a patient help clinicians understand the context of a patient visit.

2.5.3.4 Problem or Complication Questions
Questions about past problems and complications noted in the discharge summary of a patient.

2.5.3.5 Abnormality Questions
Questions referring to any abnormalities in the medical history of a patient based on their discharge summary.

2.5.3.6 Etiology Questions
Questions related to the causes of clinical conditions in a patient discharge summary.

2.5.3.7 Medical History Questions
Questions related to information in a discharge summary regarding past patient visits and events.

These question categories were selected to provide the LLM with a balanced evaluation of clinical abilities. Regulating the proportions of questions in each clinical question category allows for the dataset to target the capabilities of an LLM equally. The inclusion of question types and categories allows for the specification of optimal questions for the evaluation of clinical QA LLMs.

2.5.4 Answer Types

Typically, QA benchmarking datasets have multiple-choice or open-ended answer types. It is clear from the literature and interviews that multiple-choice questions are not representative of the clinical use case. Clinicians do not typically give a set of possible answers when asking questions. Therefore, to maximise the representativeness of the proposed dataset, open-ended answer types were selected only.

2.5.5 Benchmarking Metrics

Metrics for comparing the answers of LLMs to the expected answers in the dataset are required for evaluating LLMs. The choice of benchmarking metrics depends on the requirements of the dataset and the types of questions and answers. If a dataset consists only of yes/no questions, the choice of benchmarking metric would only need to reflect the proportion of the correct answers provided by the LLM. However, because the proposed dataset tests the ability of an LLM to answer a diverse range of question types and categories, the chosen benchmarking metrics must reflect this diversity.

1. **Exact Match**
   Exact match is especially useful for factual questions. Great if the LLM exactly matches the correct answer, but not most probable outcome, and more nuanced should be tested for. Syntactic and grammatical divergence can happen without the LLM being less correct.

2. **F1 Score**
   F1 score provides a more balanced evaluation of open-ended answers which do not have a short or simple answer.

3. **Semantic Answer Similarity**
   Evaluates semantic similarity, shifting the focus from textual likeness.

4. **ROUGE-L**
   Focuses on recall by analysing the 1-word overlap between generated text and reference text.

5. **BLEU**
   Typically used for generated text, focuses more on precision.

2.5.6 Dataset Statistics

The attributes of the dataset are also analysed using the following statistics 1) number of Q-A Pairs, 2) proportion of question types, 3) lexical richness, 4) topic distribution, and 5) coverage of clinical concepts.

2.5.6.1 Number of Q-A Pairs The Q-A pair size is measured to illustrate the robustness of the evaluation of LLMs using the dataset. Larger datasets provide more statistical power when benchmarking, so knowledge of the dataset size can inform the significance of model evaluations.

2.5.6.2 Proportion of Question Types A breakdown of the question types included provides insight into the structure of the dataset for analysis by illustrating the model capabilities being evaluated and their proportion in the dataset.

2.5.6.3 Lexical Richness Lexical richness analysis provides insight into the detail and diversity of the Q-A pairs, which are important features for a representative dataset. Higher lexical richness is more representative of real world contexts where language is dynamic and contains high lexical variance.

2.5.6.4 Topic Distribution Topic distribution analysis shows the diversity of clinical topics in the Q-A pairs, which also lends to representativeness. Real-world clinical situations are complex, and multiple clinically relevant topics can be involved in the medical history, diagnosis, and treatment of a patient.

2.5.6.5 Coverage of Clinical Concepts Lastly, measuring the coverage of clinical concepts further illuminates the structure of the dataset in terms of the concepts (separated by question category) it contains, which offers insight into the clinical capabilities being evaluated.

2.5.7 The Role of LLMs in Dataset Creation

[2] show that LLMs can be leveraged for the creation and annotation of Q-A pairs to produce an effective QA benchmarking dataset. LLMs pose a unique solution to the task of generating clinical questions with high lexical diversity, with the trade-off being that the questions must be rigorously verified as accurate, relevant, and representative. LLMs also pose the advantage of reducing the need for logical specifying the structure of the Q-A pairs, using a method like logical form templates. Again, however, this requires the contents of the Q-A pairs is verified for containing the correct information.

[2] also identified that the context length of frontier LLMs at the time was not enough for the inclusion of all MIMIC-IV discharge summaries. As a result, they only selected discharge summaries from MIMIC-IV under 8,000 tokens, omitting 30% of the patients in the MIMIC-IV corpus. This technological constraint limits the representativeness of EHRNoteQA, but can be remedied by leveraging long-context LLMs for the production of Q-A pairs.

The proposed dataset maximally leverages the generative capabilities of LLMs for the creation and annotation of Q-A pairs under the hypothesis that an effective dataset must only be verified as clinically relevant and representative to be adopted. Therefore, the focus for evaluating the effectiveness of the dataset relies on the annotation method, not the generation method. To leverage the resources available for this project, 80% of the Q-A pairs are annotated by an LLM. To emulate the process of expert oversight in annotating LLM-generated Q-A pairs, Q-A pairs are annotated by a more capable LLM than the one that generates it, in a self-supervised learning setup. This setup optimises dataset quality while minimising expert involvement and annotation efforts.

2.6 Mitigating Spurious Cues and Correlations

WikiQA [32], SelQA [33], and ANTIQUE are examples of QA datasets that have been showed to contain spurious cues and correlations that models can exploit to answer questions correctly. CoQA [34] attempt to limit spurious correlations in their dataset by avoiding exact word matches between questions and answers. Therefore, the proposed dataset aims to mitigate spurious cues and correlations in the same way, by removing Q-A pairs with 3 or more matching words in a processing step, as well as by prompting the annotation model to exclude any questions that give clues to the Q-A model.

Dataset	QA Pairs	Source Corpus	Creation Method	Availability
C-QuAL	1,742	MIMIC-III	LLM	Credentialed

Table 2: Specifications for the proposed dataset derived from the requirements

3 Methodology and Implementation

3.1 Dataset Generation Framework

The dataset was created using Python [35] and Microsoft Azure Cloud Services [36]. Python was used for the encoding of the dataset specifications and automation of the dataset generation process. Microsoft Azure Cloud Services was used for deploying LLMs for Q-A generation in accordance with the Responsible use of MIMIC data guidelines [37]. The code is hosted on GitHub and publicly available under the MIT license. The framework is written in loose accordance with PEP-8 guidelines and typical Python project structure conventions.

The framework has been produced to effectively manage the production of datasets across multiple iterations and to make the production process transparent and shareable. The purpose of this decision is to enable and encourage further work on the proposed datasets and other Q-A benchmarking datasets, in the clinical domain or in other domains.

3.2 Dataset Generation Process

The production of the dataset followed a workflow to ensure rigorous development. The workflow is based on agile principles and designed to allow for the continuous development of datasets over multiple iterations. The structure of the data generation process can be defined as a 7-step development workflow as follows: **1) gathering input data**, **2) data pre-processing**, **3) integrating dataset specifications**, **4) generating dataset**, **5) evaluating resulting dataset**, **6) updating dataset specifications**, and **7) iterating on process**. The following sections detail the specific actions taken to achieve each of the tasks in the workflow for the production of the proposed dataset.

It should be noted that the specific details can be varied for the production of other datasets.

3.2.1 Gathering Input Data

This first step involves sourcing data to be used to generate Q-A pairs using an LLM. Critically, the data gathered must be clinically relevant and representative. The chosen data for the generation of this dataset has been identified as MIMIC-III due to its availability, and improved representativeness and diversity compared with the n2c2 i2b2 corpus.

3.2.2 Data pre-processing

This is the task of selecting the optimally relevant and representative data from the gathered input data. For this dataset, the gathered data is MIMIC-III. MIMIC-III contains multiple rows of patient data. This step requires reasoning about which subset of the data is most useful for producing Q-A pairs. MIMIC-III contains 26 tables containing various types of clinical patient data, but not all of this data is useful for the production of clinical Q-A pair generation. Some tables are more useful than others. For example, the *CPTEVENTS* table contains patient billing information which, while still containing information about procedures, contains information less relevant to clinical decision reasoning such as costs and procedure codes. Using this table may introduce noise into the Q-A pair generation process and give the LLM less information to generate a Q-A pair from. Meanwhile, the *NOTEEVENTS* table contains discharge summaries, which summarise clinically relevant information for other caregivers. This table contains more data that is diverse and representative than other tables, making it a more effective data source.

Another data processing consideration is text pre-processing. This is the process of removing whitespace, stop words, and any other extraneous textual features that do not provide relevant information to the LLM. [2] feature this step for reducing the number of tokens passed to the LLM when generating Q-A pairs, citing a 10% reduction in the number of tokens. This step improves the relevance of the information given to the LLM for Q-A pair generation while reducing the number of tokens used, thus increasing the length of the discharge summary that can be passed to an LLM. Given the limited context lengths of LLMs highlighted by [2] and [3], this poses a significant advantage in the production of representative Q-A pairs by providing the LLM with amounts of information closer to the realistic setting of EHRs.

3.2.3 Integrating Dataset Specifications

This step involves taking the dataset specifications elicited from the requirements analysis, and constructing dataset features to meet these specifications, to achieve the objectives, and thus broader project aim, that the requirements were designed to fulfil.

To integrate the specifications, the following sub-steps were followed using the dataset generation framework: **1) Data Parsing**, **2) Prompt Design**, **3) Algorithm Design**, **4) Cloud Infrastructure Setup**, and **5) Parsing LLM Responses**. These sub-steps were implemented as modules of Python code in the dataset generation framework. Steps 1, 3, 4, and 5 were implemented only once and not modified in subsequent iterations. Step 1 should be modified for generating Q-A pairs from different data. Step 2 was modified to generate Q-A pairs of different types, categories, and other specifications across multiple iterations. Step 3 can be altered to create datasets of different sizes (number of Q-A pairs), and include different columns (in the framework, there is a boolean variable for the inclusion of the 'reason' column). Step 4 can be modified to generate Q-A pairs using different LLMs. Step 5 can be modified to create dataset in different formats, such as JSON.

3.2.4 Generating Dataset

To generate the dataset, the generation framework is run on a local machine with a pre-defined number of Q-A pairs. The framework calls the cloud service provider for Q-A pair generation, and saves the dataset in a specified format on the local machine.

3.2.5 Evaluating Resulting Dataset

Once produced, the dataset can be analysed and evaluated using the evaluation framework contained in the same repository as the dataset generation framework. The framework analyses the dataset using the dataset evaluation metrics and statistics identified in the design section. Exact match between questions and answers is also analysed to reduce the chances of clinical QA models exploiting spurious cues and correlations. These methods illuminate the details and quality of the dataset produced, and can inform any subsequent specifications or requirements for future datasets.

3.2.6 Iterating on the Process

Once the dataset has been generated and evaluated it may be the case that another iteration of the generation process is required to build new specifications or features into the dataset. Depending on the results of the evaluation step, it may be appropriate to begin from step 3, building in the new dataset specifications and generating a new dataset. Or it may be the case that the source data is not appropriate for generating Q-A pairs, in which case the process may start again from step 1 or 2.

3.3 Q-A Pair Annotation

The existing datasets typically leverage expert annotation to validate the quality of Q-A pairs. In this case, LLMs are maximally leveraged for annotation to reduce annotation costs. For the production of C-QuAL, the Q-A pairs are generated by gpt-35-turbo, and annotated by gpt-4o, to emulate real-world annotation. This self-supervised process informs the quality of the Q-A pairs without involving experts for this project.

3.4 Justification of Methods

This methodology was chosen to optimally meet the objectives of the dataset (being clinically relevant and representative). The iterative process allowed for a continuous update of dataset specifications allowing for the development of a dataset that meets the requirements with minimal expert involvement. Future iterations of this dataset will leverage expert annotation to further increase the accuracy, relevance, and representativeness of the Q-A pairs

3.4.1 Novel Contribution

This methodology contributes C-QuAL – a new dataset, a dataset generation workflow, and a publicly available dataset generation framework. The main distinction of the dataset is the fact that it leverages long-context LLMs for generating questions based on multiple discharge summaries.

3.4.2 Limitations

The methodology also has limitations. Only two clinicians were interviewed to gather secondary data on the details of the clinical use case, and they were both psychiatrists. These factors limit the potential representativeness of the resulting dataset. If more clinicians were interviewed, a larger pool of clinical expertise would be sampled, informing more specific dataset requirements. If other clinical professionals were interviewed, the perspectives of clinicians in other domains would inform the dataset requirements, potentially informing more representative dataset requirements.

Also, the selected corpus, MIMIC-III, was selected partly due to availability. MIMIC-III is now outdated, containing clinical data from 2001 to 2012. MIMIC-IV may pose marginal advantages in size and representativeness. MIMIC-IV contains data from 225,000 more patients than MIMIC-III, collected from 2008 to 2022. MIMIC-IV also has a more modular structure. Therefore, MIMIC-IV may contribute to the production of more clinically relevant and representative Q-A pairs than MIMIC-III.

Lastly, the distribution of question types was not controlled when generating the dataset, so the number of question types in the dataset is not even. If the question types were generated regularly, their distribution could be balanced or engineered for specific QA capability evaluations.

4 Results, Analysis and Evaluation

The dataset generation and evaluation framework was implemented as described in the methodology, resulting in a dataset containing 11,451 rows. The results were obtained by first generating the initial Q-A pairs using **gpt-3.5-turbo** and MIMIC-III discharge summaries. Q-A pairs with 3 or more exactly matching words were also removed to

mitigate against spurious correlations. This resulted in a dataset containing 4,246 rows. The Q-A pairs were then annotated using **gpt-4o**, reducing the length of the resulting dataset to 1,742.

```
Your task is to evaluate the provided model output in response to
a specific question associated with the given discharge
summaries.

By using the correct answer also provided, you must score the
answer as 0 or 1, based on the following scoring instructions.

Scoring Instructions:
1. Assign 0 points if the answer is either incorrect, or
if it falsely claims there is no answer when one exists
according to the discharge summaries.\n\n
2. Assign 0 points if the question gives away the answer
to the model. Be strict about this.
3. Assign 1 otherwise
4. You should only assign the output 1 if you think it is a very
good question-answer pair for benchmarking a clinical large
language model

Please do not include any other information than the score
number.

Output format: Score: [your score of either 0 or 1]

Discharge Summaries:
{discharge_summary}

Question: {question}

Correct Answer: {expected_answer}

Score:
```

Figure 3: Q-A pair annotation prompt for **gpt-4o**.

Q-A pairs marked with '0's were then removed from the resulting annotated dataset, resulting in a final dataset of 1,743 rows.

4.1 Results

C-QuAL is presented in two formats: **C-Qual-XL** and **C-QuAL-small**. C-QuAL CL is a concatenation of all of the Q-A pairs intially generated for the creation of the dataset (totalling 11,451 rows). C-QuAL-small is the reduced, higher-quality dataset of 1,742 rows.

Model	Exact Match	F1 Score	Semantic Answer Similarity	Rouge	BLEU
GPT-4o	0.0	0.3711	0.8264	0.3714	0.4067
GPT-3.5-Turbo	0.0	0.4017	0.8941	0.3984	0.4830
Llama-3-70b-Instruct	0.0	0.3427	0.7093	0.3366	0.2902
Mistral-large	0.01212	0.3501	0.8722	0.3560	0.4564

Table 3: Performance metrics for different LLMs.

4.1.1 Model Benchmarks

4 LLMs were benchmarked on C-QuAL-small to provide preliminary results on its effectiveness as a benchmark. 3 illustrates the quality of answers generated by GPT-4o, GPT-3.5-Turbo, LLama-3-70b-Instruct, and Mistral-large. It is clear that GPT-3.5-Turbo gained the highest overall scores for the metrics, but this result is biased by the fact that GPT-3.5-Turbo is the model which initially generated the Q-A pairs.

The data shows that C-QuAL-small is an effective benchmark for LLMs with long context windows, and can aid in the selection of models to be deployed in clinical settings to assist with clinical decision reasoning.

4.2 Statistical Analysis

The proportion of question types present in C-QuAL-small, showing that the largest proportion of question types is **yes/no/maybe questions**. Of the question types specified to the LLM when generating the Q-A pairs, only 4 out of the 6 specified question types were included, and the distribution of the question types in the dataset is not balanced.

4.3 Limitations

Not all of the Q-A pairs were used when benchmarking the LLMs due to resource constraints, so the benchmarking results are not fully representative of the performance of the LLMs are the effectiveness of C-QuAL-small as a benchmark.

5 Legal, Social, Ethical and Professional Issues

A chapter gives a reasoned discussion about legal, social ethical and professional issues within the context of your project problem. You should also demonstrate that you are aware of the Code of Conduct & Code of Good Practice issued by the British Computer Society (BSC) (https://www.bcs.org/membership/become-a-member/bcs-code-of-conduct/) for computer science project and Rule of Conduct issued by The Institution of Engineering and Technology (IET) (https://www.theiet.org/about/governance/rules-of-conduct/) for engineering project. You should have applied their principles, where appropriate, as you carried out your project. You could consider aspects like: the effects of your project on the public well-being, security, software trustworthiness and risks, Intellectual Property and related issues, etc.

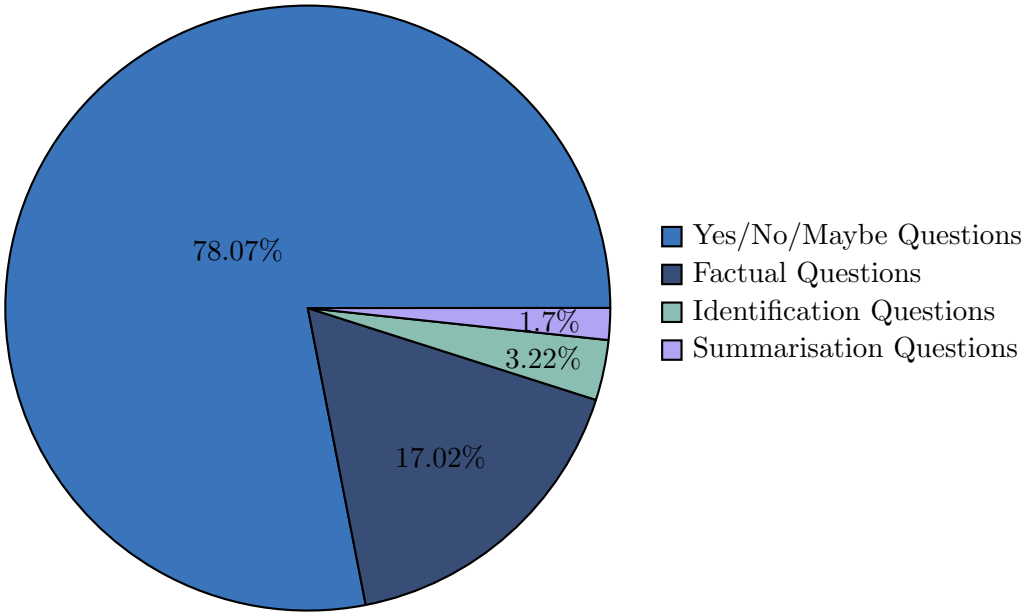

Figure 4: Proportion of question types in generated dataset

5.1 Implications of BCS Code of Conduct

In accordance with the BSC code of conduct, this project takes appropriate measures to ensure the privacy and legitimate rights of MIMIC-III data in the production of the dataset. No patient is identifiable from MIMIC-III patient records. The production of this dataset is compliant with all relevant legislation and takes measures to mitigate harmful outcomes.

5.2 Effects of Project

5.2.0.1 Potential Positive Effects The broad intended effect of this project is to contribute to the efficiency of healthcare systems. Namely to improve their efficiency by reducing the amount of time that clinicians need to spend manually reading and organising information in EHRs when diagnosing patients. One potential effect of this succeeding is improved patient outcomes in contexts such as hospitals where LLMs are deployed to assist with clinical decision reasoning. Mapping the causal pathway between reduced clinician time and improved patient outcomes, it can be argued that if clinicians spend less time reading and organising information in patient EHRs, they will need to spend less time treating patients. This may free the clinicians up to care for more patients at once, or it may allow them to take more time to rest, which could improve the quality with which they deliver care. Another way LLMs deployed for assisting with clinical reasoning could contribute to improved patient outcomes is by organising, arranging, and processing information more effectively than clinicians currently do. If this happens, then it can be argued that the quality of patient care would be increased

overall, which could lead to improved patient outcomes.

5.2.0.2 Potential Negative Effects

The automation of processes related to clinical decision reasoning could result in the lack of need for human clinicians in healthcare settings. It is not clear whether LLMs will replace clinicians, but work such as this is a step towards the automation of healthcare. While this may result in improved efficiency in the healthcare domain, it may also result in the redundancy of clinicians.

The introduction of black-box LLMs in clinical settings also raises safety concerns. Although benchmarks provide a realistic summary of the clinical capabilities of models, the models can make mistakes and this can lead to negative outcomes in healthcare. It could be the case that an LLM deployed in a clinical setting may not do what the developers or clinician intend, and as a consequence, produce an output that leads to unhelpful or harmful outcomes.

5.3 Privacy and Security

While the dataset is released publicly, it contains MIMIC-III discharge summaries which are protected under the HIPPA restrictions of the MIMIC-III dataset. For this reason, the dataset can only be made public to individuals with credentialed access to MIMIC-III data. The data within the discharge summaries is de-identified, meaning no patient data is visible to any individual working with C-QuAL, the C-QuAL generation framework, or MIMIC-III.

6 Conclusion

This report documents the production of a new clinical QA benchmarking dataset for large language models deployed to assist with clinical decision reasoning. The report details the evaluation of existing approaches and interviews with clinicians, finding that while solutions exist, the solutions are not all representative of clinical contexts. It also finds that the existing solutions do not all contain clinically relevant Q-A pairs, and thus may not be readily adopted in clinical settings. Lastly, it finds a gap in the existing research for the application of recent LLMs with long context windows than previously available.

Taking these findings into account, the report details the production of a dataset that aims to address the identified flaws and gaps in the existing literature. The dataset produced, C-QuAL, is presented in two formats: C-QuAL-XL, a large dataset with variance in Q-A pair quality, and C-QuAL-small, a smaller, higher-quality dataset. A generation and evaluation framework, and a generation methodology is also presented.

The report documents the evaluation of the dataset and benchmarking of long-context LLMs with the dataset, finding that it is an effective benchmark for the QA capabilities of LLMs in clinical settings. The report highlights the limitations of the methodology and dataset, such as that it lacks human annotation, is not as representative of clinical contexts as it could be, and that the benchmarking was not fully representative of model abilities. Lastly, the report illuminates potential directions for further work that may enable the development of a more effective benchmark that can build on the dataset presented in the report.

References

[1] A. Pampari, P. Raghavan, J. Liang, and J. Peng, "emrQA: A Large Corpus for Question Answering on Electronic Medical Records," Sept. 2018. arXiv:1809.00732 [cs].

[2] S. Kweon, J. Kim, H. Kwak, D. Cha, H. Yoon, K. Kim, J. Yang, S. Won, and E. Choi, "EHRNoteQA: An LLM Benchmark for Real-World Clinical Practice Using Discharge Summaries," June 2024. arXiv:2402.16040 [cs].

[3] K. Kotschenreuther, "EHR-DS-QA: A Synthetic QA Dataset Derived from Medical Discharge Summaries for Enhanced Medical Information Retrieval Systems."

[4] P. Raghavan, J. J. Liang, D. Mahajan, R. Chandra, and P. Szolovits, "emrKBQA: A Clinical Knowledge-Base Question Answering Dataset," in *Proceedings of the 20th Workshop on Biomedical Language Processing* (D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, eds.), (Online), pp. 64–73, Association for Computational Linguistics, June 2021.

[5] J. Bardhan, A. Colas, K. Roberts, and D. Z. Wang, "DrugEHRQA: A Question Answering Dataset on Structured and Unstructured Electronic Health Records For Medicine Related Queries," May 2022. arXiv:2205.01290 [cs].

[6] S. Bae, D. Kyung, J. Ryu, E. Cho, G. Lee, S. Kweon, J. Oh, L. Ji, E. Chang, T. Kim, and E. Choi, "EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images," *Advances in Neural Information Processing Systems*, vol. 36, pp. 3867–3880, Dec. 2023.

[7] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams," May 2021.

[8] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A Dataset for Biomedical Research Question Answering," Sept. 2019. arXiv:1909.06146 [cs, q-bio].

[9] S. Moon, H. He, H. Liu, and J. Fan, *RxWhyQA: a clinical question-answering dataset with the challenge of multi-answer questions.* Jan. 2022.

[10] X. Yue, X. F. Zhang, Z. Yao, S. Lin, and H. Sun, "CliniQG4QA: Generating Diverse Questions for Domain Adaptation of Clinical Question Answering," Dec. 2021. arXiv:2010.16021 [cs].

[11] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering," in *Proceedings of the Conference on Health, Inference, and Learning*, pp. 248–260, PMLR, Apr. 2022. ISSN: 2640-3498.

[12] "i2b2: Informatics for Integrating Biology & the Bedside."

[13] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, May 2016. Publisher: Nature Publishing Group.

[14] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Scharli, A. Chowdhery, P. Mansfield, B. A. y. Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, "Large Language Models Encode Clinical Knowledge," Dec. 2022. arXiv:2212.13138 [cs].

[15] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, and E. Cambria, "A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics," June 2024. arXiv:2310.05694 [cs].

[16] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu, "A large language model for electronic health records," *npj Digital Medicine*, vol. 5, pp. 1–9, Dec. 2022. Publisher: Nature Publishing Group.

[17] A. Andrew, "Potential applications and implications of large language models in primary care," *Family Medicine and Community Health*, vol. 12, p. e002602, Jan. 2024. Publisher: BMJ Specialist Journals Section: Communication.

[18] Y.-J. Park, A. Pillai, J. Deng, E. Guo, M. Gupta, M. Paget, and C. Naugler, "Assessing the research landscape and clinical utility of large language models: a scoping review," *BMC Medical Informatics and Decision Making*, vol. 24, p. 72, Mar. 2024.

[19] M. Cascella, F. Semeraro, J. Montomoli, V. Bellini, O. Piazza, and E. Bignami, "The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives," *Journal of Medical Systems*, vol. 48, p. 22, Feb. 2024.

[20] N. Mehandru, B. Y. Miao, E. R. Almaraz, M. Sushil, A. J. Butte, and A. Alaa, "Evaluating large language models as agents in the clinic," *npj Digital Medicine*, vol. 7, pp. 1–3, Apr. 2024. Publisher: Nature Publishing Group.

[21] U. Anwar, A. Saparov, J. Rando, D. Paleka, M. Turpin, P. Hase, E. S. Lubana, E. Jenner, S. Casper, O. Sourbut, B. L. Edelman, Z. Zhang, M. Günther, A. Korinek, J. Hernandez-Orallo, L. Hammond, E. Bigelow, A. Pan, L. Langosco, T. Korbak, H. Zhang, R. Zhong, S. hÉigeartaigh, G. Recchia, G. Corsi, A. Chan, M. Anderljung, L. Edwards, Y. Bengio, D. Chen, S. Albanie, T. Maharaj, J. Foerster,

F. Tramer, H. He, A. Kasirzadeh, Y. Choi, and D. Krueger, "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," Apr. 2024. arXiv:2404.09932 [cs].

[22] L. Bereska and E. Gavves, "Mechanistic Interpretability for AI Safety – A Review," Apr. 2024. arXiv:2404.14082 [cs].

[23] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1424, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1424.

[24] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1312.

[25] S. Hong, L. Xiao, X. Zhang, and J. Chen, "ArgMed-Agents: Explainable Clinical Decision Reasoning with LLM Disscusion via Argumentation Schemes," June 2024. arXiv:2403.06294 [cs].

[26] V. Liévin, C. E. Hother, A. G. Motzfeldt, and O. Winther, "Can large language models reason about medical questions?," *Patterns*, vol. 5, p. 100943, Mar. 2024.

[27] T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen, "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine," *NPJ digital medicine*, vol. 7, p. 20, Jan. 2024.

[28] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, pp. 1930–1940, Aug. 2023. Publisher: Nature Publishing Group.

[29] M. Ailem, K. Marazopoulou, C. Siska, and J. Bono, "Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks," June 2024. arXiv:2404.16966 [cs].

[30] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, p. 1, Jan. 2023. Publisher: Nature Publishing Group.

[31] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," Oct. 2016. arXiv:1606.05250 [cs].

[32] Y. Yang, W.-t. Yih, and C. Meek, "WikiQA: A Challenge Dataset for Open-Domain Question Answering," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (L. Màrquez, C. Callison-Burch, and J. Su, eds.),

(Lisbon, Portugal), pp. 2013–2018, Association for Computational Linguistics, Sept. 2015.

[33] "SelQA: A New Benchmark for Selection-Based Question Answering | IEEE Conference Publication | IEEE Xplore."

[34] Z. Zhong, M. Yang, and R. Xu, "Reducing Spurious Correlations for Answer Selection by Feature Decorrelation and Language Debiasing," in *Proceedings of the 29th International Conference on Computational Linguistics* (N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, eds.), (Gyeongju, Republic of Korea), pp. 1753–1764, International Committee on Computational Linguistics, Oct. 2022.

[35] "Welcome to Python.org," July 2024.

[36] "Cloud Computing Services | Microsoft Azure."

[37] "Responsible use of MIMIC data with online services like GPT."

A Appendix

A.1 Dataset Analysis Tables

Dataset	Question Types	Question Categories	Reasoning Types
emrQA	Logical form entity extraction	Relations, medications, heart disease, obesity, and smoking	Lexical variation (synonym), lexical variation (world/medical knowledge), syntactic variation, multiple sentence, arithmetic, temporal, incomplete context, and class prediction
EHRNoteQA	Not specified	Treatment, assessment, problem, etiology, sign/symptom, vitals, test results, history, instruction, and plan	Not specified
EHR-DS-QA	Not specified	Not specified	Not specified
emrKBQA	Yes/no, factual, and temporal	Test results, medications, conditions, and other (e.g., allergies, tobacco use)	Not specified
DrugEHRQA	Easy, medium, and hard	Drug-dosage, drug strength, route, form of medicine, problems	Multi-modal (table and text)
EHRXQA	Image-based, table-based, and image + table-based	Modality-based and patient-based	Uni-modal and cross-modal (image, table, and image + table)
MedQA	Text based factual questions based on a source document	Single piece of knowledge, and question based on patient condition description	One-step and multi-hop
PubMedQA	Yes/no/maybe	Does a factor influence the output? Is a therapy good/necessary? Is a statement true? Is a factor related to the output?	Inter-group comparison, interpreting subgroup statistics, interpreting group statistics, and other
RxWhyQA	Single-drug, multi-drug, answerable, and unanswerable	Unanswerable, 1-to-1 drug and reason, 1 drug and n reasons, multiple drugs and n reasons	Not specified
CliniQG4QA	What, when, has, was, why, how, is, did, can, any, and does	Not specified	Not specified
MedMCQA	Multiple-choice	Anaesthesia, anatomy, biochemistry, dental, ENT, FM, O&G, medicine, microbiology, ophthalmology, orthopaedics, pathology, pediatrics, pharmacology, physiology, psychiatry, radiology, skin, PSM, surgery, unknown	Question logic, explanation/definition, cause of events (diagnosis), treatment, mathematical, teleology/purpose, multi-hop reasoning, fill in the blanks, analogy, natural language inference, factual, and comparison

Figure 5: Comparative analysis of question specifications in existing datasets

Dataset	Provide Benchmarking Results	Evaluation Method	Models Evaluated
emrQA	Yes	Exact match and F1	DrQA and Class Prediction
EHRNoteQA	Yes	GPT-4, BLEU, ROUGE-L	GPT4, GPT4-Turbo, GPT3.5-Turbo, Llama3-70b-Instruct, Llama2-70b-chat, qCammel-70, Camel-Platypus2-70b, Platypus2-70b-Instruct, Mixtral-8x7b-Instruct, MPT-30b-Instruct, Llama2-13b-chat, Vicuna-13b, WizardLM-13b, qCammel-13, OpenOrca-Platypus2-13b, Camel-Platypus2-13b, Synthia-13b, Asclepius-13b1, Gemma-7b-it, MPT-7b-8k-instruct, Mistral-7b-Instruct, Dolphin-2.0-mistral-7b, Mistral-7b-OpenOrca, SynthIA-7b, Llama2-7b-chat, Vicuna-7b, Asclepius-7b
EHR-DS-QA	No	N/A	N/A
emrKBQA	No	N/A	N/A
DrugEHRQA	Yes	Exact match and F1	TREQS, RAT-SQL, BERT-QA, and ClinicalBERT
EHRXQA	Yes	Accuracy, F1 and AUC_rel	Prior, PubMedCLIP, MedViLL, and M^3AE
MedQA	Yes	Accuracy	Chance, PMI, IR-ES, IR-Custom, Max-out, BERT-Base-En, clinicalBERT-Base, BioRoBERTa-Base, BioBERT-Base, RoBERTa-Large, and BioBERT-Large
PubMedQA	Yes	Accuracy, and F1	Majority, human, Shallow Features, BiLSTM, ESIM w/ BioELMo, and BioBERT
RxWhyQA	No	N/A	N/A
CliniQG4QA	Yes	Exact match and F1	DocReader and ClinicalBERT
MedMCQA	Yes	Accuracy	BERT-Base, BioBERT, SciBERT, PubMedBERT

Figure 6: Comparative analysis of the benchmarking setups for existing clinical QA benchmarking datasets