

Final Product: Memorial Sloan Kettering Cancer Center

Abstract

With the open release of the large language model (LLM) ChatGPT in 2022, artificial intelligence (AI) has grown extremely prevalent worldwide, in all industries. This includes healthcare, where LLMs aid in clinical tasks such as medical documentation, patient communication, diagnosis prediction, and decision-making, among other tasks. Nevertheless, these models are still young and prone to various inaccuracies that can lead to severe outcomes at the risk to patient safety and trust in AI tools. This research study explores how these inaccuracies can be combated by evaluating various LLMs' performance on healthcare-specific tasks.

This study compares and evaluates a general-purpose LLM, Llama 3.1, and a medical-domain-specific LLM, Asclepius. The models were trained using a publicly available dataset using GPU-memory efficient techniques to gauge their ability to handle domain-specific challenges. Three evaluation metrics were then employed to evaluate the accuracy of their outputs. The results of this study provide insights into LLMs' limitations in healthcare and potential strategies for tackling them in the future, paving the way for the successful adoption of advanced healthcare LLMs.

Introduction

With the integration of large language models (LLMs), such as ChatGPT, in the daily lives of every industry, it's no surprise that such artificial intelligence (AI) tools are also being increasingly used in the healthcare industry. LLMs can support healthcare professionals in tasks such as medical documentation, patient communication, and clinical decision-making to facilitate and ease their busy workflows. While promising, these LLMs are prone to inaccuracies, hallucinations, and bias in their outputs. These errors are crucial to address immediately as, in healthcare especially, they can lead to not only dangerous outcomes for patients but also to mistrust in AI-driven tools.

This research paper aims to investigate how such challenges can be overcome and provide direction to the improvements needed to enhance reliability and safety. This study uses publicly available, domain-specific datasets such as the MIMIC-III Clinical Database and MedQuAD to train LLMs for healthcare-specific tasks. Although our original objective was to perform domain-specific LLM fine-tuning using medical text data to improve performance and accuracy on healthcare tasks, we were unable to acquire sufficient GPU access to do so. Thus, we changed

our objective to align with GPU-memory efficient training instead of conducting full-scale fine-tuning. The objective is to investigate the next steps that could be taken to improve LLMs for effective adoption in the medical domain, so we trained existing LLMs on medical text data to observe their performance. We focused on metrics such as hallucination rates and semantic similarity to better understand how presently available models could be improved for medical purposes.

This project seeks to mitigate bias by informing hallucination reduction to improve the accuracy of healthcare- and medical-specific responses provided by LLMs. The research holds significant potential for advising future medical-specific fine-tuning of LLMs to enhance patient outcomes, create efficient clinical workflows, and cultivate ethical and trustworthy applications of AI in the healthcare field.

Literature Review

I. LLMs in Healthcare

LLMs are AI tools designed to understand and generate human-like text, such as ChatGPT, GPT-3, and LLaMA, among others. They are trained on vast amounts of text data to perform various natural language processing (NLP) tasks, ranging from text classification (assigning labels to text), language identification, creative writing, and summarization, to question-and-answering (Open AI, 2024). In a study by Cascell et al. (2023), the feasibility of ChatGPT being used in multiple healthcare situations was evaluated. While it demonstrated great potential for supporting clinical tasks, its accuracy and reliability were limited. ChatGPT was able to provide coherent and sensible responses when given medical-related prompts, but it struggled with producing precise and accurate medical information. Thus, there is a need for domain-specific LLMs that can better approach healthcare-specific tasks.

Specialized LLMs for healthcare settings can be invaluable for tasks such as clinical documentation, decision support, patient education, generating medical narratives, streamlining workflows, and reducing the cognitive load on healthcare providers. BioBERT and ClinicalBERT are fine-tuned on medical text data and are actively used in healthcare (Yang et al., 2023). BioBERT has been pre-trained on large-scale biomedical literature from the United States National Institutes of Health's (NIH) National Library of Medicine, specifically PubMed Central, a free, full-text archive. Such extensive and targeted training allows BioBERT to navigate through vast amounts of biomedical information to assist healthcare professionals and researchers in finding the right information immediately and accurately (Garon, 2024). ClinicalBERT is trained on anonymized clinical notes from the MIMIC-III Clinical Database, which contains real clinical data. This specific training makes ClinicalBERT ideal for tasks such as diagnosis classification and mortality prediction, aiding in decision-making processes

(Alsentzer et al., 2019). The integration of these models into clinical settings illustrates the growing application of NLP to improve healthcare processes in order to achieve more efficient healthcare experiences and delivery. Similarly, Nazi & Peng's (2023) review expressed that both BioBERT and ClinicalBERT have improved performance on medical-specific tasks, but many challenges existed in conjunction. The challenges included biases in training data, hallucinations, and ethical issues related to data privacy; they emphasized the need to create more robust and domain-specific training approaches to combat these inaccuracies in order to not only increase the performance of these specialized LLMs but also trust in them.

II. LLM Hallucinations in Healthcare

Hallucinations are when machine learning (ML) models, particularly LLMs such as GPT-3 and GPT-4, produce outputs that are coherent and grammatically correct, but factually inaccurate or nonsensical. For example:

Prompt: "How do I treat a water allergy?"

Hallucinated response: "A water allergy can be treated with antihistamines and regular exposure therapy to build tolerance to water."

While the response sounds legitimate and it is true that antihistamines can treat a water allergy, the latter half of the response is untrue and not a medically recognized treatment. This hallucination could lead to dire repercussions, highlighting the importance of domain-specific training and well-checked responses in LLMs. Such errors are especially common in medical and legal domains due to the need for precise, context-specific knowledge when operating within these domains' strict standards for accuracy and accountability.

LLMs prioritize generating plausible, fluent, human-sounding texts over factual correctness, leading to errors such as hallucinations (Li et al., 2024). Addressing these hallucinations lies in the design and training of LLMs, requiring foundational changes such as explicit fact-checking and focusing on factual accuracy and verifiability during fine-tuning. This is echoed by Orgad et al. (2024), who state that hallucinations are deeply ingrained in LLM's probabilistic nature. LLMs often "know" the correct information but have trouble expressing it in their outputs, so the inconsistencies between a model's internal representations and its generated text need to be re-examined.

In a study by Yang et al. (2023), transformer-based models such as BERT, GPT-3, and LLaMA are at the forefront of NLP in healthcare. Domain-specific LLMs such as BioBERT, PubMedBERT, and ClinicalBERT, which are trained on medical text data, have the potential to improve task-specific performance. Despite this, they are still prone to the "hallucination effect," which is when LLMs create false information in a guessing behavior. This occurs due to gaps in

domain-specific knowledge and models' lack of depth and nuance to handle such specialized topics due to their training on broad datasets, which can lead to significant consequences such as misdiagnosis or incorrect treatment recommendations. However, there have been recent advances in mitigating hallucinations such as methods like self-reflection (Ji et al., 2023). This is where models self-evaluate and correct their own responses to improve accuracy. This could be especially useful in the medical domain due to uncommon professional concepts and potential social risks involved where the results of misinformation can be extremely dangerous. Yet, self-reflection does not address the root causes of hallucinations which lie in LLM training architecture, and only mitigates them (Li et al., 2024). Despite such advancements, hallucination reduction in healthcare still remains largely unexplored.

III. Current State of Training Healthcare-Specific LLMs

Training healthcare-specific LLMs has greatly increased in recent years due to the demand and integration of AI tools in various industries. Unlike general-use LLMs such as ChatGPT, medical-domain-specific LLMs are designed especially to handle the unique tasks of medicine, such as understanding medical terminology syntax. Models such as BioBERT, ClinicalBERT, PubMedBERT, and Asclepius, to name a few, demonstrate the successful expertise of healthcare LLMs.

These LLMs have advanced immensely in recent years thanks to medical datasets such as MIMIC-III and MedQuAD (Yang et al., 2024). MIMIC-III is a dataset containing real, de-identified clinical notes and MedQuAD is an expert-verified medical question-and-answer database. These datasets are frequently used for training and fine-tuning LLMs as they have the contextual, healthcare-specific information necessary to turn a general LLM into a specialized LLM. MIMIC-III was used to train LlamaCare (Li et al., 2024) and ClinicalBERT (Alsentzer et al., 2019) to perform tasks such as diagnosis classification and medical summarization, and MedQuAD was used to train Baize, a medical chatbot (Xu et al., 2023).

Numerous strategies are being refined to train specialized medical LLMs for both efficiency and accuracy. For instance, SmallToLarge (S2L) is a scalable data selection method that can perform supervised fine-tuning by using the training trajectories of smaller models. This reduces the amount of data needed for fine-tuning instead of the full-scale dataset. In a study by Yang et al. (2024), S2L used only 50% of MIMIC-III to outperform full dataset training. This improved data efficiency without sacrificing performance, and also reduced computational costs. Additionally, LLMs are traditionally trained using retrieval-augmented generation (RAG) methods which separates retrieval model training from LLM training. However, a new approach has been proposed where LLMs are synchronized during training: Joint Medical LLM and Retrieval Training (JMLR) allows the retrieval model to retrieve relevant medical information which the LLM uses to generate accurate responses (Wang et al., 2024). This training method not only

improved accuracy performance on medical Q&A datasets — and surpassed previous models — but it also reduced hallucinations and had a more computationally efficient training process.

IV. Gaps

While specialized medical LLMs have great potential in ensuring the precision and ethical standards required in healthcare — which can support better patient outcomes and trust in AI tools — training them includes immense challenges. Data biases, unverified answers, and ethical concerns related to data privacy spotlight the imminent need for robust training and evaluation to ensure accuracy and reliability in these LLMs.

Despite significant progress made in specialized medical LLMs, there remain large gaps in literature and research. Many existing models are trained on limited datasets that fail to capture the nuances and highly dimensional nature of real-world clinical data, leading to hallucinations, generalization, and biases (Kweon et al., 2023). There is a need for proactive, fundamental architectural changes in LLMs to address hallucinations rather than only using mitigative techniques such as self-reflection (Li et al., 2024). Moreover, there have only been a handful of studies where LLMs are employed in real-life clinical environments. Ethical considerations such as patient privacy and data security are also ongoing challenges (Nazi & Peng, 2024). Another area that needs attention is the lack of resource-efficient and scalable fine-tuning techniques. The S2L technique has begun addressing this by optimizing training resource allocation, but this is still a space for major progress (Yang et al., 2024). By addressing these gaps in specialized LLMs, successful development and trust in LLMs for healthcare tasks can ensue.

Methods

I. Data

A. MIMIC-III Clinical Database

The MIMIC-III Clinical Database is a large-scale, freely available database that contains de-identified health-related data from over 40,000 unique critical care patients of the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001 and 2012 (Johnson et al., 2016). The data spans approximately 60,000 patient admissions, including both adult and neonatal patients. All the data was recorded in conjunction with routine hospital care, so daily workflow was not affected, from sources including critical care information systems, hospital electronic health record databases, and the Social Security Administration Death Master File. In the entire MIMIC-III database, there are 26 CSV files that contain different medical information, such as caregiver descriptions, codes for different procedures and diagnoses, patient prescriptions, patient descriptions, and patient transfers, among other information (Table 1). This database is especially valuable for three reasons: it is publicly available internationally, it

compromises a very large and diverse range of ICU patients, and it contains highly detailed data (e.g. vital signs, etc.). This database was created with the intention of scientific use for both researchers and healthcare professionals to improve patient care, build predictive models, and further the use of artificial intelligence in the healthcare field.

Core Tables	Clinical Tables	Diagnosis and Procedures	Free-Text and Notes	Meta and Lookup Tables
ADMISSIONS PATIENTS ICUSTAYS TRANSFERS	CALLOUT CHARTEVENTS LABEVENTS OUTPUVENTS INPUTEVENTS_CV INPUTEVENTS_MV PROCEDUREEVENTS_MV PRESCRIPTIONS MICROBIOLOGYEVENTS	DIAGNOSES_ICD DRGCODES PROCEDURES_ICD D_ICD_DIAGNOSES D_ICD_PROCEDURES D_CPT CPT_EVENTS	NOTEVENTS	CAREGIVERS D_ITEMS D_LABITEMS DATETIMEEVENTS SERVICES

Table 1: Content organization of MIMIC-III's 26 CSV files.

NOTEVENTS

We focused on the NOTEVENTS file, which contains all clinical notes and text documentation written by healthcare professionals during patients' hospital stays, with over two million clinical notes. This file includes physician and nursing notes, radiology reports, discharge summaries, and the exact time the notes were written and entered into the digital medical system. This file has each row representing a single clinical note. So, a row of information does not represent all of the notes and information for a single patient, but rather a single entry; patients can have numerous rows dedicated to their hospitalization. Additionally, the NOTEVENTS table is unique as it contains qualitative insights whereas many of the other tables include quantitative, numerical data. While the data is extremely suitable to train an LLM and is great for NLP tasks, it will be challenging to use due to the unstructured nature of its free text that includes unique abbreviations, spellings, etc. (Figure 1a, 1b), requiring substantial data cleaning.

ROW_ID	SUBJECT_ID	HADM_ID	CHARTDATE	CHARTTIME	STORETIME	CATEGORY	DESCRIPTION	CGID	ISERROR	TEXT	
0	174	22532	167853.0	2151-08-04	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2151-7-16**] Dischar...
1	175	13702	107527.0	2118-06-14	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2118-6-2**] Discharg...
2	176	13702	167118.0	2119-05-25	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2119-5-4**] D...
3	177	13702	196489.0	2124-08-18	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2124-7-21**] ...
4	178	26880	135453.0	2162-03-25	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2162-3-3**] D...

Figure 1a: Screenshot of NOTEVENTS' columns.

TEXT

Admission Date: [**2151-7-16**] Discharge Date: [**2151-8-4**]Service:ADDENDUM:RADIOLOGIC STUDIES: Radiologic studies also included a chestCT, which confirmed cavitory lesions in the left lung apexconsistent with infectious process/tuberculosis. This alsomoderate-sized left pleural effusion.HEAD CT: Head CT showed no intracranial hemorrhage or masseffect, but old infarction consistent with past medicalhistory.ABDOMINAL CT: Abdominal CT showed lesions ofT10 and sacrum most likely secondary to osteoporosis. These canbe followed by repeat imaging as an outpatient. [**First Name8 (NamePattern2) **] [**First Name4 (NamePattern1) 1775**] [**Last Name (NamePattern1) **], M.D. [**MD Number(1) 1776**]Dictated By:[**Hospital 1807**]MEDQUIST36D: [**2151-8-5**] 12:11T: [**2151-8-5**] 12:21JOB#: [**Job Number 1808**]

Figure 1b: Screenshot of NOTEEVENTS' text column.

Data Cleaning

The text column for each clinical entry row contains unstructured, messy medical notes, with all identifiable information de-identified for confidentiality. It consists of a mixture of structured (e.g., dates, names, identifiers) information and unstructured free text (Figure 1b). The raw text is unusable for LLM training and NLP tasks as it is too noisy, inconsistent, and domain-specific. Cleaning the data will allow the content to be better understood and used by LLMs, and this was done by applying NLP techniques to preprocess the data systematically, ensuring its suitability for further analysis.

The goal of data cleaning was to segment the unstructured, free-text data into predefined sections (e.g., medical history, demographics, lab results, etc.) and organize them in an analyzable, structured format. Regex-based (regular expressions-based) matching was used to segment the notes into predefined categories, capturing text between section headers and organizing it into a dictionary. Regex patterns were designed for each predefined section based on common keywords that were found in the clinical notes. Admission and discharge dates, gender, and age were associated with “Demographics.” Chief complaint, social history, and family history were associated with “Medical History.” Physical exams, lab results, and radiology reports were associated with “Clinical Observations.” This was achieved by batch processing, where the clinical notes were processed in chunks, and then extracted sections were combined into a data

frame with one column per section. The cleaning process resulted in structured data tables for easier LLM analysis (Figure 2).

	Admission_Date	Discharge_Date	Age	Gender	Service	ChiefComplaint	Medical_History	Social_History	Family_History	Allergies ...	Physical_Exam
0	["2151-7-16"]n	["2151-8-4"]n			ADDENDUM:n					...	
1	["2118-6-2"]n	["2118-6-14"]n	Pn		MICU and then to ["Doctor Last Name "] Medic...	s, neurologic changes, rashes,ipalpitations, ...	This is an 81-year-old female/nwith a history ...	: Lives with her husband, Dr. ["Known lastna...	: Noncontributory/n	: or some other type of inflammatory process. A...	Blood pressure 142/76,nheart rate 100 and reg...
2	["2119-5-4"]n	["2119-5-25"]n			CARDIOTHORACICn	:n81 yo F smoker w/ COPD, severe TBM, s/p tra...	bronchoscopy 3/31,4/2,3,["5-12"],["5-17"]...	:n: The patient is married and worked as a cl...	:n: (+) FHx CAD; Father with an MI in his 40's...	:nAmlodipine/n/nAttending: ["Last Name (NameP...	Admit H+PnGeneral-lovely 81 yr old female in ...
3	["2124-7-21"]n	["2124-8-18"]n			MEDICINE/nDIV. OF GASTROENTEROLOGY/nWhen: WEDN...	:nCOPD exacerbation/Shortness of Breath/nMa...	Intubation/narterial line placement/nPICC line...	:nThe patient is married and worked as a clin...	:n(+) FHx CAD; Father with an MI in his 40's...	:nAmlodipine/n/nAttending: ["First Name3 (LF)...	On admission/nVitals: T: BP: 116/46 P: 92 Q2: ...
4	["2162-3-3"]n	["2162-3-25"]n	Mn		MEDICINE/n	:nMr. ["Known lastname 1829"] was seen at [...	1. Anterior cervical osteotomy, C6-C7, with de...	:nPatient recently discharged from ["Hospita...	reports patient was very sad and attempted to...	:nPatient recorded as having No Known/nTo Dr...	Phyical exam prior to surgery was not obtained... :n=====

Figure 2: Screenshot of cleaned NOTEEVENTS.

There were significant challenges to cleaning and using the NOTEEVENTS file. The data's format changes after the 100,000th row, leaving only 40,000 entries usable out of the entire 400,000 rows. This emphasizes the variability in the dataset's structure, which would require the exclusion of problematic rows or extra cleaning. Additionally, the clinical notes in the text column varied widely in section headers, abbreviations, formatting, and typos, which required advanced cleaning. However, such advanced cleaning for large-scale unstructured text required significant computational resources to process and utilize effectively, which we didn't have.

It is important to mention that although we intended to fine-tune an LLM and use the MIMIC-III database, it was not feasible due to limited computational resources and GPU memory. The text data in MIMIC-III's NOTEVENTS file was specifically designed for fine-tuning tasks, hence it could not contribute to our final output. However, it is included in this paper as we, the MSK team, have devoted effort to researching and cleaning it.

B. MedQuAD Dataset

MedQuAD, also known as the Medical Question Answering Dataset, is a specialized dataset designed for training and evaluating Q&A systems in the medical domain. The dataset contains 47,457 question-answer pairs from trusted, verified medical sources, spanning various NIH agencies as well as other credible health websites (Ben Abacha & Demner-Fushman, 2019). The topics covered include symptoms, diagnosis, treatment, causes, prevention, and healthcare services, among others, of various medical topics ranging from polycystic ovary syndrome to what anesthesia is. Questions are formatted in various ways, such as "What are the symptoms of ____?" and "What tests are used to diagnose ____?" The answers are written in a way that is easy to understand by both those in the medical field and those who are not (ex: Asthma is caused by inflammation and narrowing of the airways ...). This dataset was used for training and benchmarking purposes, so we needed to clean it, and that was done in two parts: question selection and cleaning.

Question Selection

We opted to double-check that the questions and answers were accurate and could be directly found on the websites, and we did this by hand. During this process, it was also discovered that many of the questions and answers were duplicated, so when we selected the 141 Q&A pairs to be used for benchmarking, we did it manually to ensure that we had a variety of questions. We only used 141 Q&A pairs due to limitations in computational capacity on our laptops.

Question Cleaning

Many of the original questions were formatted in a way so that the model would provide a “yes” or “no” answer, which wasn’t going to provide us any contextual information on how the model performs in terms of accuracy. For example, an original question would be: “Do you have information about pulling a patient up in bed (Also called: Moving a patient in bed)?” The model would likely provide a simple “yes” or “no” answer. So, we reformulated the question so that the model would have to provide a more descriptive answer: “What is the correct method for safely pulling or moving a patient up in bed, including steps to protect the patient’s skin and shoulders, prevent caregiver injury, and use appropriate equipment like slide sheets?”

II. Model Selection

We researched multiple LLMs in order to select the models that we would be using for our project, ranging from PubMedBERT to ClinicalT5. By compiling various LLMs and their succinct details, including accessibility, pros and cons, and possible use for medical text (Table 2), we were able to decide on two models: Llama 3.1 and Asclepius.

model name	Person Assigned	description	price	Pros	Cons	Related File/ Links	Possible Info on Medical Application (Not Compulsory)
K-Nearest Neighbors (KNN)	Wendy Ding(support)	machine learning algorithm	Free	- Simple to understand and implement - Effective for small datasets	- Slow prediction time for large datasets - Sensitive to irrelevant features	https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-sci-kit-learn	Can be used for diagnosing diseases based on similarity to known patient data
Convolutional Neural Networks (CNN)	Wendy Ding(support)	Convolutional Neural Networks (CNN) are deep learning models used for medical image analysis.	Google Colab: Free GPU/TPU access; upgrade to Colab Pro for \$10/month for more power and resources	Excellent for image analysis tasks. Automatically extracts features from images.	Requires large datasets and significant computational power. Prone to overfitting with small datasets.	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6515915/	Tumor detection in MRI/CT scans. X-ray classification for diseases. Retinal image analysis for diabetic retinopathy. Histopathology for cancer detection.
Autoencoders	Wendy Ding(support)	Autoencoders are used in medical imaging for noise reduction, data reconstruction, and anomaly detection.	Free with open-source libraries; \$10/month on Google Colab Pro for GPU use.	Effective for noise reduction and data compression. Improves image quality.	Prone to overfitting. Requires large datasets for optimal performance.	https://arxiv.org/abs/2004.05453	Autoencoders are used in medicine for noise reduction, image reconstruction, and data compression in MRI/CT scans.
BioGPT	Tang Chia Hui (support)	specialised for biomedical text	Free for research use, Free for finetuning but requires manual set-up (cost depends on what we use, e.g. GPU)	excels specifically in biomedical question answering and knowledge extraction	has a very specific purpose, not all encompassing	https://arxiv.org/abs/2210.10341	used for literature mining, question-answering on biomedical research and hypothesis generation
BioGPT (supplemental info)	Qinhao Chen	based on GPT-2 because the parameters of GPT-3 is too huge to operate		scores higher in six medical tasks than PubMedBert, GPT2, BioBert, and REBEL		ibdiem	
PubMedBERT	Qinhao Chen	based on BERT and trained with PubMed	open-source	outperforms BioBERT, SciBERT, Clinical BERT, BlueBERT in most BLURB NLP tasks	good at discriminating tasks but not generative tasks	https://arxiv.org/pdf/2007.15779	
BlueBERT-PubMed & MIMICIII	Qinhao Chen		open-source	good at Cancer and Smoke tasks	more inaccurate than PubMedBERT in most cases	https://arxiv.org/pdf/2310.13146	
PMC-LLaMA	Qinhao Chen	2023 paper, based on LLaMA-2, light-weighted on the parameters with 7B	open-source	outperforms GPT-4 on Medical QA Tasks		https://arxiv.org/pdf/2304.14454	

Table 2: Model selection organization.

A. Llama 3.1

Llama 3.1 is a generic LLM that was released as the second latest iteration of Meta’s Llama series. It is a collection of open-source LLMs ranging from 8 billion to 405 billion parameters (Meta, 2024). It is highly versatile, performing well in various natural language processing tasks such as question-answering, translation, and dialogue generation, among others. Llama 3.1 is also extremely customizable so it can be used in medical applications through fine-tuning with a medical dataset. We opted to use Llama 3.1 because it has some of the best robust general language understanding and conversational skills, which ensures user-friendly interactions. Additionally, it was extremely accessible, being free and open-source. We also chose the 8-billion parameter model due to resource constraints, less training time, and practicality.

B. Asclepius

Asclepius, named after the Greek god of medicine and healing, is the first and one of the best specialized medical LLMs that was developed specifically to address the challenges associated with using real clinical notes in training due to privacy regulations like HIPAA (Health Insurance Portability and Accountability Act). It is an adaptation of LLaMA-3.0, which served as its base model, and was pre-trained using synthetic clinical notes generated with GPT-3.5-turbo for domain adaptation before fine-tuning with clinical instructions. These synthetic clinical notes were created using public case reports from biomedical literature, such as MIMIC-III and PubMed so that patient confidentiality would not be breached (Kweon et al., 2023). Asclepius was evaluated using GPT-4 and tailored evaluation criteria to focus on accuracy, relevancy, and completeness. It was found that Asclepius performed just as well as LLMs that are trained on real clinical notes. We decided to use Asclepius as our second model, per Aleksandr’s suggestion, because it fills the gap in domain expertise that Llama 3.1 has, ensuring medical accuracy and reliability. Additionally, we were provided with the option of using the 8-billion or 13-billion parameter model, and we chose the former due to resource constraints.

III. Training

A. GPU-Memory Efficient Training

Due to the lack of sufficient GPU access, we adapted our approach to focus on GPU-memory efficient training rather than conducting full-scale LLM fine-tuning. We utilized quantization to make the training process more computationally efficient, which is when the parameters of an LLM are transformed into a more memory-efficient format without sacrificing performance significantly. Specifically, the model’s parameters were transformed into a 4-bit normal float format. This is when numbers are represented with 4 bits, instead of the usual 32 or 16, where 1 bit is allocated for the sign (positive or negative), 1-2 bits for the exponent (magnitude or range), and 1-2 bits for the fraction to represent precision (Liu et al., 2023). This dramatically reduces the memory required to store model parameters and speeds up training operations, while allowing for observation and evaluation of model performance on medical tasks.

B. Model Answer Generation

The training process was quite simple for Llama 3.1 and Asclepius. No full- or large-scale training was performed due to a lack of computational resources and also because the goal was to focus on the performance and functionality of the two models, rather than building full models. We used Unsloth AI, an open-source tool designed for “finetuning large language models like Llama-3 ... use 70% less memory, and with no degradation in accuracy” (Unsloth, 2024), to train a small, sample model. Instead of using the entire MedQuAD dataset, only a portion of the dataset was used — 141 Q&A pairs. Minimal preprocessing was needed to use MedQuAD for training as the Q&A pairs were split into questions and answers. The extracted questions served as input prompts to feed into our LLMs and their corresponding answers were used as the reference/target answers for the output. We guided our models by using predefined prompts to ensure that the model generates detailed answers, such as:

“You are an intelligent model. Below is a medical question.

Write a detailed response that appropriately completes the answer. The answer should provide all necessary and accurate information, while being concise.”

The training process pipeline consisted of tokenizing, or breaking down, the input (the questions from MedQuAD), feeding the tokenized input into the quantized Llama 3.1 and Asclepius models, and then generating outputs based on the predefined prompts. Below is an example of the resulting model-generated answers:

Question: What is (are) Polycystic ovary syndrome? (Also called: Polycystic ovaries; Polycystic ovary disease; Stein-Leventhal syndrome; Polyfollicular ovarian disease)

Llama-3.1: Polycystic ovary syndrome (PCOS) is a common health condition affecting women of childbearing age. PCOS affects how the ovaries function, causing the ovaries to produce an abnormal amount of hormones. This can lead to infertility and other health problems. PCOS is a chronic condition that can be managed with lifestyle changes and medication. Treatment options include birth control pills, metformin, and other medications. With proper treatment, women with PCOS can have healthy pregnancies and live normal lives.<|end_of_text|>

Asclepius: Polycystic ovary syndrome is a condition characterized by the presence of multiple cysts in the ovaries, which can cause pain, bloating, and other symptoms. It is also known by several other names, including Stein-Leventhal syndrome and Polyfollicular ovarian disease.<|end_of_text|>

IV. Benchmarking & Evaluation

To evaluate the performance of the trained Llama 3.1 and Asclepius models on MedQuAD data, we utilized three methods: benchmarking with BERTScore, DeepEval evaluation, and manual evaluation.

A. BERTScore Benchmarking

Benchmarking is a process done to evaluate a model’s performance against predefined metrics that are industry standards to assess the effectiveness of a model, as well as identify its strengths and weaknesses and determine areas of improvement. For our Llama 3.1 and Asclepius models which were trained on MedQuAD, we used data from MedQuAD — the reference answers — to compare with the model’s outputs as it is verified by experts, making it the ideal benchmark to evaluate the models’ semantic accuracy.

The benchmarking metric we decided on was BERTScore to evaluate the quality of the text-generation tasks. BERTScore uses contextual embedding from a pre-trained BERT model to capture semantic similarities between the model-generated and reference texts. BERTScore calculates a semantic similarity score for each token in the model-generated answer and for each token in the reference answer using cosine similarity. Precision, recall, and F1 scores are calculated by BERTScore to evaluate the similarity for each of Llama 3.1 and Asclepius’ answers. Precision measures how much of the model-generated answer is relevant compared to the reference answer; recall measures how much of the reference answers that the model-generated answers capture; F1 score measures the average of precision and recall. This allows us to not only compare Llama 3.1 and Asclepius’ outputs to the reference answers but also how Llama 3.1, a general-purpose LLM, performs compared to Asclepius, a medical domain-specific LLM.

B. DeepEval Evaluation

DeepEval is an open-source framework used to evaluate and test LLM outputs and encompasses various evaluation metrics, including metrics such as G-Eval, hallucination detection, and answer relevancy. With GPT-4o API, we used the G-Eval metric, to determine the correctness of the model-generated answers by determining factual accuracy and completeness with the reference answers. A correctness score is calculated for each model-generated answer based on the overlap of accurate and relevant details. This is done using the Chain of Thought (CoT) technique, which uses step-by-step reasoning to determine if there are contradictions or omissions in the model-generated answers compared to the reference answer. Especially, the G-Eval allows us to modify the CoT process in order to better fit our aims. We used the following criteria: “Determine whether the actual output is factually correct based on the expected output. Provide 1 for correct and 0 for incorrect. Omission of details is acceptable and should not be deemed as incorrect.” Then the following CoT prompt is used:

[“Check whether the facts in ‘actual output’ contradicts any facts in ‘expected output’”,

“You should only penalize omission of important detail”,
“Vague language are OK”]

C. Manual Evaluation

For manual evaluation, we created our own evaluation metric by calculating two scores: a hallucination score and a factor score.

The hallucination score measures if there are any inaccuracies or factual errors in the model-generated answers. This is done by checking for the disease on Mayo Clinic’s website, and investigating if the answer generated by the model aligns with the information verified and provided by Mayo Clinic. If there is any incoherence or incorrect information, no matter how seemingly insignificant, the model-generated answer receives a factor score of 1. If the model-generated answer is coherent and entirely correct, then it receives a factor score of 0.

The factor score is calculated by counting the number of factoids in the reference answer and the model-generated answers, and then comparing them. Factoids are defined as small, specific pieces of information, such as facts or short statements, that the LLM retrieves or generates to answer a question. Here is an example of how to count factoids:

Question: What are some reasons the Eiffel Tower was built, and how tall is it?

Answer: The Eiffel Tower was built as the entrance arch for the 1889 World’s Fair (1), was designed by Gustave Eiffel’s company (2), and is 330 meters tall (3).

The answer above has 3 factoids.

The factor score is as follows:

factoids from reference answer present in model-generated answer

factoids from reference answer

The number of factoids in the MedQuAD answer is the number of facts, or short statements. The number of factoids in the model-generated answer is the # of MedQuAD factoids present in the model-generated answer. Below is an example of calculated hallucination and factor scores (Table 3).

ANSWERID	Question	Answer	URL	AsclepiusAnswer	Hallucination Score	Factor Score
ADAM_0003147_Secl.txt	What is (are) Polycystic	Polycystic ovary syndrome is a	https://www.nlm.	Polycystic ovary syndrome is a condition characterized	0	2/5

	ovary syndrome? (Also called: Polycystic ovaries; Polycystic ovary disease; Stein-Leventhal syndrome; Polyfollicular ovarian disease)	condition in which a woman has an imbalance of female sex hormones. This may lead to changes in the menstrual cycle, cysts in the ovaries, trouble getting pregnant, and other health problems.)	nih.gov/medlineplus/ency/article/00369.htm	by the presence of multiple cysts in the ovaries, which can cause pain, bloating, and other symptoms. It is also known by several other names, including Stein-Leventhal syndrome and Polyfollicular ovarian disease.< end_of_text >		
--	---	--	--	--	--	--

Table 3: Example of hallucination and factor scores calculation.

The hallucination score is a 0 because Asclepius’ answer is coherent with the information found on Mayo Clinic. The final factor score is 2/5. The MedQuAD answer has 5 factoids: Polycystic ovary syndrome is a condition in which a woman has an imbalance of female sex hormones (1). This may lead to changes in the menstrual cycle (2), cysts in the ovaries (3), trouble getting pregnant (4), and other health problems (5). The Asclepius answer has 2 out of the 5 MedQuAD factoids: Polycystic ovary syndrome is a condition characterized by the presence of multiple cysts in the ovaries (3), which can cause pain, bloating, and other symptoms (5). It is also known by several other names, including Stein-Leventhal syndrome and Polyfollicular ovarian disease.

Results

I. Benchmarking Results

The BERTScore F1 scores for both Llama 3.1 and Asclepius were quite high (Figure 3). For Llama 3.1, the average precision score was 0.8517, average recall was 0.8258, and average F1 score was 0.8377 (Appendix Table 1). For Asclepius, the average precision score was 0.8340, average recall was 0.8321, and average F1 score was 0.8325 (Appendix Table 2). The high precision scores for both models highlight that their generated answers were generally quite strong in semantic similarity to the reference answers. The high recall scores for both also suggest that both models capture a significant portion of the information in the reference answers. Both of the high F1 scores indicate that Llama 3.1 and Asclepius maintain a rather good balance between answer quality and completeness.

II. Model Evaluation Results

The average DeepEval scores for Llama 3.1 and Asclepius are 0.5234 and 0.6119, respectively (Figure 3). The individual scores for both models varied widely for answer evaluation, ranging from 0.0 - 0.9 (Appendix Table 3). However, the DeepEval scores were on average higher for almost every Asclepius-generated answer and varied less widely.

III. Manual Evaluation Results

Llama 3.1’s model-generated answers resulted in 94 hallucination scores of 0 and 47 hallucination scores of 1 ($M = 0.6667$) (Appendix Table 4). Asclepius’ model-generated answers resulted in 110 hallucination scores of 0 and 31 hallucination scores of 1 ($M = 0.7801$) (Appendix Table 4). Thus, Llama 3.1 did not have any hallucinations 67% of the time, while Asclepius did not have any hallucinations 78% of the time (Figure 3). In terms of factor score, Llama 3.1 had an average factor score of 0.3769 while Asclepius had an average factor score of 0.4314 (Appendix Table 4).

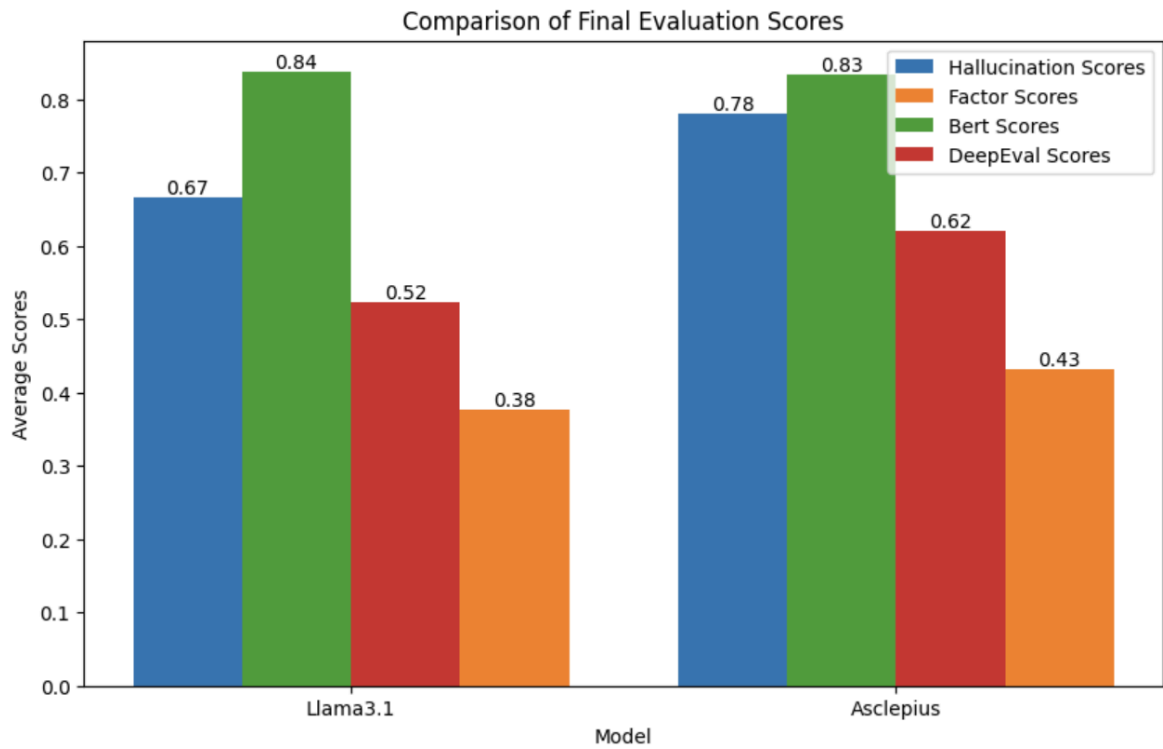


Figure 3: Final results of Llama 3.1 and Asclepius’ three evaluation metric scores (BERTScore, DeepEval, manual evaluation).

The hallucination scores here depict the score of 0, meaning no hallucination occurred.

Conclusions & Takeaways

This study investigated the performance of Llama 3.1 and Asclepius on their effectiveness and accuracy when answering medical-specific questions, using three metrics: BERTScore, DeepEval, and manual evaluations. The results provide insights into the strengths and limitations of a general-purpose LLM vs. a domain-specific LLM in the medical domain and offer direction for future steps to improving LLMs for healthcare functions.

BERTScore measured the semantic similarity between the model-generated answers and the reference answer, with higher scores expressing better alignment with the reference answers. The F1 scores for the two models were about the same, at ~ 0.83 . Such high scores suggest that both models perform well in generating answers that convey a similar meaning to the reference answer. The lack of difference between the scores implies that there is not a significant benefit of domain-specific tuning for Asclepius when it comes to semantic similarity.

The DeepEval score assessed the overall correctness of the model-generated answers compared to the reference answers. Asclepius demonstrates a higher degree of correctness and factual reliability compared to Llama 3.1. This is likely due to Asclepius being pre-trained on medical-specific information, giving it an advantage in generating domain-specific answers for healthcare settings. However, the moderate scores for both LLMs highlight that correctness/factuality is an area for refinement.

Regarding the manual evaluation's hallucination metric, Asclepius has fewer hallucinations compared (31) to Llama 3.1 (47), making it more reliable in factual accuracy and logical coherence. This suggests that such domain-specific LLMs are more effective in reducing hallucinations. For the factor score, since Asclepius had an average factor score greater than Llama 3.1 (0.43 vs. 0.38), Asclepius captures more relevant information in its answers than Llama 3.1 does, illustrating more alignment with the reference answer. This indicates that Asclepius is better at generating more factually complete answers, likely due to being fine-tuned on domain-specific datasets. However, both models' relatively low factor scores express that there is still room for improvement in capturing completeness and relevant information.

Overall, Asclepius outperformed Llama 3.1 on most evaluation metrics (DeepEval score, hallucination score, factor score) except BERTScore, where it performed about the same as Llama 3.1. While Llama 3.1 did not fall too far behind Asclepius for the most part, it being a general-use LLM has limited performance when used in specialized tasks. The results here shed light on the benefits of domain-specific fine-tuning of LLMs, as seen in Asclepius, for healthcare applications.

By using all three metrics — BERTScore, DeepEval, and manual — it seems that DeepEval's G-Eval metric is a suitable evaluation strategy for measuring the correctness of model-generated output to its reference. Future research can focus on addressing factor scores by enhancing models' ability to generate more factual and comprehensive answers and reducing hallucinations through fine-tuning using specialized datasets.

Appendix

AnswerID	Question	Reference_Answer	Model_Answer	Precision	Recall	F1_Score
ADAM_0003147_Sec1.txt	What is (are) Polycystic	Polycystic ovary syndrome is a condition in which	Polycystic ovary syndrome (PCOS) is a condition i	0.9066802263	0.9282924533	0.9173591137
ADAM_0003147_Sec2.txt	What causes Polycystic	PCOS is linked to changes in hormone levels that	Polycystic ovary syndrome (PCOS) is a common h	0.8736370802	0.818074286	0.8449432254
ADAM_0002818_Sec7.txt	What are the complicat	Buildup of fluid in tissues of body (lymphedema, c	1. Webbed neck	0.8608127832	0.7869035602	0.8222005367
ADAM_0002818_Sec9.txt	How to prevent Noonan	Couples with a family history of Noonan syndrom	The Noonan syndrome is a rare genetic condition.	0.826259315	0.8393058181	0.8327314854
GARD_0004375_Sec1.txt	What are the symptoms	The Human Phenotype Ontology provides the foll	Neurofibromatosis-Noonan syndrome (NFNS) is a	0.837351203	0.7556208968	0.7943894267
GARD_0004450_Sec3.txt	Is Noonan syndrome in	Noonan syndrome is inherited in an autosomal dc	Yes, Noonan syndrome is inherited. It is inherited i	0.8881638646	0.8607999682	0.8742678761
GARD_0004450_Sec4.txt	What are the treatment	Management generally focuses on the specific sig	Treatment of Noonan syndrome focuses on mana	0.8965555429	0.8535968661	0.8745490313
GHR_0000804_Sec2.txt	How many people are a	Polycystic kidney disease is a fairly common gen	There are about 600,000 people affected by polycy	0.9202393293	0.8572892547	0.8876496553
ADAM_0002818_Sec1.txt	What is (are) Noonan s	Noonan syndrome is a disease that can be passe	Noonan syndrome is a genetic disorder that affect	0.8623281717	0.8792276978	0.8706959486
GARD_0004450_Sec1.txt	What is (are) Noonan s	Noonan syndrome is a genetic disorder that caus	Noonan syndrome is a genetic condition that affect	0.9176362157	0.8569838405	0.8862735629
GHR_0000804_Sec1.txt	What is (are) polycystic	Polycystic kidney disease is a disorder that affect	Polycystic kidney disease is a genetic disorder cha	0.9427757859	0.8196823001	0.8769304752
GHR_0000804_Sec3.txt	What are the genetic ch	Mutations in the PKD1, PKD2, and PKHD1 genes c	PKD is caused by a mutation in a gene called PKD	0.8867934942	0.8178171515	0.8509097695
GHR_0000804_Sec4.txt	Is polycystic kidney dis	Most cases of polycystic kidney disease have an	Polycystic kidney disease is a genetic disorder cha	0.8784015775	0.8173872828	0.846796751
ADAM_0000719_Sec1.txt	What is (are) Celiac dis	Celiac disease is an immune disorder passed dow	Celiac disease is a digestive condition that damag	0.9210230112	0.8930125237	0.9068015218
ADAM_0000721_Sec1.txt	What is (are) Celiac dis	Celiac disease is a condition that creates inflam	Celiac disease is an autoimmune disorder that ca	0.8864002824	0.8954411149	0.8908977509
ADAM_0000721_Sec3.txt	What are the symptoms	The symptoms of celiac disease can be different	Celiac disease is a digestive disorder. It is caused	0.8368906975	0.7602493167	0.7967311144
ADAM_0000721_Sec7.txt	What are the complicat	You must carefully continue to follow the gluten-f	Celiac disease - sprue is a disorder that causes da	0.8363660574	0.8075355291	0.82169801
ADAM_0000721_Sec8.txt	Do I need to see a doct	Call your health care provider if you have sympto	No, you don't need to see a doctor for Celiac disea	0.8386246562	0.8651857376	0.8516981602
ADAM_0000721_Sec9.txt	How to prevent Celiac c	Because the exact cause is unknown, there is no	The answer is:	0.8251634836	0.7919573188	0.8082194924

Appendix Table 1: Llama 3.1 Benchmarking Results

AnswerID	Question	Reference_Answer	Model_Answer	Precision	Recall	F1_Score
ADAM_0003147_Sec1.txt	What is (are) Polycystic ova	Polycystic ovary syndrome is a condition in which	Polycystic ovary syndrome is a condition characte	0.8511602879	0.9045917988	0.8770629764
ADAM_0003147_Sec2.txt	What causes Polycystic ova	PCOS is linked to changes in hormone levels that	The exact cause of Polycystic ovary syndrome is	0.8420497179	0.8230688572	0.8324511647
ADAM_0002818_Sec7.txt	What are the complications	Buildup of fluid in tissues of body (lymphedema, c	Noonan syndrome is a genetic disorder that can k	0.8137360811	0.8069462776	0.8103269339
ADAM_0002818_Sec9.txt	How to prevent Noonan syn	Couples with a family history of Noonan syndrom	Based on the given information, there is no knowr	0.8190930486	0.8705252409	0.8440263271
GARD_0004375_Sec1.txt	What are the symptoms of N	The Human Phenotype Ontology provides the foll	The symptoms of Neurofibromatosis-Noonan syn	0.830129087	0.765296042	0.7963952422
GARD_0004450_Sec3.txt	Is Noonan syndrome inheri	Noonan syndrome is inherited in an autosomal dc	Yes, Noonan syndrome is inherited. It is a genetic	0.8590089679	0.8360194564	0.847358346
GARD_0004450_Sec4.txt	What are the treatments for	Management generally focuses on the specific sig	Based on the given information, there is no specif	0.8428688049	0.8448588848	0.8438627124
GHR_0000804_Sec2.txt	How many people are affect	Polycystic kidney disease is a fairly common gen	According to the PKD Foundation, approximately 2	0.843309164	0.8456176519	0.8444618583
ADAM_0002818_Sec1.txt	What is (are) Noonan syndr	Noonan syndrome is a disease that can be passe	Noonan syndrome is a genetic disorder that affect	0.8360034823	0.8641431928	0.8498404622
GARD_0004450_Sec1.txt	What is (are) Noonan syndr	Noonan syndrome is a genetic disorder that caus	Noonan syndrome is a genetic condition that typi	0.8816077113	0.8534111977	0.8672803044
GHR_0000804_Sec1.txt	What is (are) polycystic kidn	Polycystic kidney disease is a disorder that affect	Polycystic kidney disease is a condition in which	0.8603815436	0.8223837614	0.8409537077
GHR_0000804_Sec3.txt	What are the genetic change	Mutations in the PKD1, PKD2, and PKHD1 genes c	The genetic changes related to polycystic kidney	0.8321173191	0.8116126657	0.8217371702
GHR_0000804_Sec4.txt	Is polycystic kidney disease	Most cases of polycystic kidney disease have an	Yes, polycystic kidney disease is inherited. It is a d	0.8442038894	0.8122913241	0.8279402256
ADAM_0000719_Sec1.txt	What is (are) Celiac disease	Celiac disease is an immune disorder passed dow	The nutritional considerations for patients with ce	0.846409142	0.8589583635	0.8526375294
ADAM_0000721_Sec1.txt	What is (are) Celiac disease	Celiac disease is a condition that creates inflam	Celiac disease - sprue is a condition where the sm	0.8564501405	0.8930826783	0.8743829131
ADAM_0000721_Sec3.txt	What are the symptoms of C	The symptoms of celiac disease can be different	The symptoms of Celiac disease - sprue include a	0.8380746245	0.7770239711	0.8063954115
ADAM_0000721_Sec7.txt	What are the complications	You must carefully continue to follow the gluten-f	The complications of Celiac disease - sprue inclu	0.8294327855	0.8173906207	0.8233677149
ADAM_0000721_Sec8.txt	Do I need to see a doctor fo	Call your health care provider if you have sympto	Based on the information provided, it is recomme	0.8287696838	0.8835759759	0.8552957773
ADAM_0000721_Sec9.txt	How to prevent Celiac disea	Because the exact cause is unknown, there is no	To prevent Celiac disease-sprue, it is important to	0.8270891905	0.8599405885	0.8431950212

Appendix Table 2: Asclepius Benchmarking Results

✓	AnswerID	CorrectedQuestions	Reference Answer	URL	Llama3.1 Answer	Asclepius Answer	DeepEval_Llama3.1_score	DeepEval_Llama3.1_reason	DeepEval_Asclepius_score	DeepEval_Asclepius_reason
0	ADAM_0003147_Sec1.txt	What is (are) Polycystic o	Polycystic ovary	https://www	Polycystic ovar	Polycystic ovary	0.8035267205	The actual output accurately desc	0.814025923	The actual output correctly identifi
1	ADAM_0003147_Sec2.txt	What causes Polycystic o	PCOS is linked t	https://www	Polycystic ovar	The exact cause	0.2757246356	The actual output describes symp	0.7521135767	The actual output correctly states t
2	ADAM_0002818_Sec7.txt	What are the medical, dev	Buildup of fluid i	https://www	Noonan syndro	Noonan syndro	0.3484835781	The actual output mentions Noon	0.8798186781	The actual output mentions medic
3	ADAM_0002818_Sec9.txt	How to prevent Noonan s	Couples with a f	https://www	Noonan syndro	There is current	0.1643023755	The actual output does not addre	0.8909020483	The output correctly states there is
4	GARD_0004375_Sec1.txt	What are the detailed sign	The Human Phe	https://rar	Neurofibromat	Neurofibromat	0.6123258713	The actual output correctly identifi	0.6280188949	The actual output mentions key fe
5	GARD_0004450_Sec3.txt	How is Noonan syndrome	The Human Phe	https://rar	Noonan syndro	Noonan syndro	0.9	The actual output accurately desc	0.285158117	The actual output incorrectly descr
6	GARD_0004450_Sec4.txt	What are the treatment op	Management ge	https://rar	Noonan syndro	The treatment o	0.1748852519	The actual output describes Noon	0.3546616254	The actual output acknowledges th
7	GHR_0000804_Sec2.txt	How many people are affe	Polycystic kidne	https://ghr	There are an es	The prevalence	0.9679178708	The actual output provides global	0.2687686243	The actual output acknowledges th
8	ADAM_0002818_Sec1.txt	What is (are) Noonan syn	Noonan syndro	https://www	Noonan syndro	Noonan syndro	0.9115897588	The actual output provides a com	0.6708498846	The actual output correctly describ
9	GARD_0004450_Sec1.txt	What is (are) Noonan syn	Noonan syndro	https://rar	Noonan syndro	Noonan syndro	0.9163393784	The actual output accurately desc	0.6792344229	The actual output correctly identifi
10	GHR_0000804_Sec1.txt	What is polycystic kidney	Polycystic kidne	https://ghr	Polycystic kidn	Polycystic kidn	0.9002981766	The output accurately describes P	0.7199983618	The actual output correctly identifi
11	GHR_0000804_Sec3.txt	What are the genetic char	Mutations in the	https://ghr	Polycystic kidn	The genetic cha	0.1152824463	The output repeats the same infor	0.5961555086	The actual output correctly mentio
12	GHR_0000804_Sec4.txt	Is polycystic kidney disea	Most cases of p	https://ghr	Polycystic kidn	Yes, polycystic k	0.4738053215	The output correctly states PKD is	0.6320673116	The actual output correctly states t
13	ADAM_0000719_Sec1.txt	What is (are) Celiac disea	Celiac disease i	https://www	Celiac disease	Celiac disease i	0.2520376251	The actual output provides detaile	0.8053452281	The actual output correctly identifi
14	ADAM_0000721_Sec1.txt	What is (are) Celiac disea	Celiac disease i	https://www	Celiac disease	Celiac disease -	0.8262005554	The actual output accurately desc	0.894231208	The actual output correctly describ
15	ADAM_0000721_Sec3.txt	What are the symptoms of	The symptoms	https://www	Celiac disease	The symptoms	0.1456092392	The actual output does not list an	0.8108552605	The actual output mentions key sy
16	ADAM_0000721_Sec7.txt	What are the complication	You must carefu	https://www	Celiac disease	The complicatio	0.188189325	The actual output discusses celiac	0.7709826974	The actual output mentions compli
17	ADAM_0000721_Sec8.txt	When should I see a doctor	Call your health	https://med	* When should	You should see	0.08515463007	The actual output is a repetition o	0.6783709171	The output correctly suggests see
18	ADAM_0000721_Sec9.txt	How to prevent Celiac dis	Because the exa	https://www	Celiac disease	The best way to	0.2418945186	The actual output incorrectly state	0.3154585317	The actual output suggests preven

Appendix Table 3: Llama 3.1 and Asclepius DeepEval Evaluation Results

✓	AnswerID	CorrectedQuestions	Reference Answer	URL	Llama3.1 Answer	Llama3.1 Hallucination Score	Llama3.1 Factor Score	Asclepius Answer	Asclepius Hallucination Score	Asclepius Factor Score
0	ADAM_0003147_Sec1.txt	What is (are) Polycystic ovary	Polycystic ovary syndrome is	https://www	Polycystic ovary syndro	0	3/3	Polycystic ovary syndro	0	3/3
1	ADAM_0003147_Sec2.txt	What causes Polycystic ovary	PCOS is linked to changes in	https://www	Polycystic ovary syndro	0	4/8	The exact cause of Polyc	1	1/8
2	ADAM_0002818_Sec7.txt	What are the medical, develo	Buildup of fluid in tissues of	https://www	Noonan syndrome is a	1	1/4	Noonan syndrome is ass	0	2/4
3	ADAM_0002818_Sec9.txt	How to prevent Noonan syndr	Couples with a family history	https://www	Noonan syndrome is a	0	1/1	There is currently no kno	1	0/1
4	GARD_0004375_Sec1.txt	What are the detailed signs a	The Human Phenotype Onto	https://rare	Neurofibromatosis-No	0	3/8	Neurofibromatosis-Noor	0	3/8
5	GARD_0004450_Sec3.txt	How is Noonan syndrome inh	The Human Phenotype Onto	https://rare	Noonan syndrome is in	0	1/8	Noonan syndrome is inh	1	0/8
6	GARD_0004450_Sec4.txt	What are the treatment optio	Management generally focus	https://rare	Noonan syndrome is a	0	3/5	The treatment options a	1	1/5
7	GHR_0000804_Sec2.txt	How many people are affecte	Polycystic kidney disease is a	https://ghr.n	There are an estimated	0	4/5	The prevalence rate of p	1	1/5
8	ADAM_0002818_Sec1.txt	What is (are) Noonan syndro	Noonan syndrome is a disea	https://www	Noonan syndrome (NS)	0	2/6	Noonan syndrome is a ge	0	4/6
9	GARD_0004450_Sec1.txt	What is (are) Noonan syndro	Noonan syndrome is a genet	https://rare	Noonan syndrome is a	0	5/5	Noonan syndrome is a ge	0	3/5
10	GHR_0000804_Sec1.txt	What is polycystic kidney dis	Polycystic kidney disease is a	https://ghr.n	Polycystic kidney disea	0	6/8	Polycystic kidney disea	0	6/8
11	GHR_0000804_Sec3.txt	What are the genetic change	Mutations in the PKD1, PKD2	https://ghr.n	Polycystic kidney disea	0	2/8	The genetic changes rela	1	3/8
12	GHR_0000804_Sec4.txt	Is polycystic kidney disease li	Most cases of polycystic kidn	https://ghr.n	Polycystic kidney disea	0	5/8	Yes, polycystic kidney dis	0	3/8
13	ADAM_0000719_Sec1.txt	What is (are) Celiac disea -	Celiac disease is an immune	https://www	Celiac disease is an imm	0	5/5	Celiac disease is a digest	0	3/5
14	ADAM_0000721_Sec1.txt	What is (are) Celiac disea -	Celiac disease is a condition	https://www	Celiac disease is a cond	0	3/3	Celiac disease - sprue is c	0	3/3
15	ADAM_0000721_Sec3.txt	What are the symptoms of Ce	The symptoms of celiac dise	https://www.r	Celiac disease is a gen	1	1/8	The symptoms of Celiac	0	1/8
16	ADAM_0000721_Sec7.txt	What are the complications o	You must carefully contin	https://www.r	Celiac disease is an im	0	3/6	The complications of Cel	0	4/6
17	ADAM_0000721_Sec8.txt	When should I see a doctor f	Call your health care provide	https://medlin	* When should I see a	0	1/4	You should see a doctor	0	2/4
18	ADAM_0000721_Sec9.txt	How to prevent Celiac disea	Because the exact cause is	https://www.r	Celiac disease is an au	0	2/2	The best way to prevent	1	1/2
19	ADAM_0000218_Sec1.txt	What is (are) Anesthesia - w	You are scheduled to have a	https://www.r	Anesthesia is a drug or	0	3/5	Anesthesia is a medical	0	3/5
20	ADAM_0001986_Sec1.txt	What information do you hav	You had a hip joint replacem	https://www.r	<[end_of_text]>	1	0/4	Based on the given disc	0	2/4

Appendix Table 4: Llama 3.1 and Asclepius Manual Evaluation Results

Sources

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

- Ben Abacha, A., & Demner-Fushman, D. (2019). A question-entailment approach to question answering. *BMC bioinformatics*, 20, 1-23.
- Cascella, M., Montomoli, J., Bellini, V. *et al.* Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst* 47, 33 (2023). <https://doi.org/10.1007/s10916-023-01925-4>
- Garon, C. (2024). *The revolutionary impact of BioBERT in biomedical natural language processing*. Retrieved December 9, 2024, from <https://christophegaron.com/articles/research/the-revolutionary-impact-of-biobert-in-bio-medical-natural-language-processing/>
- Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet*. <https://doi.org/10.13026/C2XW26>
- Kweon, S., Kim, J., Kim, J., Im, S., Cho, E., Bae, S., ... & Choi, E. (2023). Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes. *arXiv preprint arXiv:2309.00237*.
- Li, J., Consul, S., Zhou, E., Wong, J., Farooqui, N., Ye, Y., ... & Damos, G. (2024). Banishing LLM hallucinations requires rethinking generalization. *arXiv preprint arXiv:2406.17642*.
- Liu, S. Y., Liu, Z., Huang, X., Dong, P., & Cheng, K. T. (2023). Llm-fp4: 4-bit floating-point quantized transformers. *arXiv preprint arXiv:2310.16836*.
- Li, R., Wang, X., & Yu, H. (2024, May). LlamaCare: An Instruction Fine-Tuned Large Language Model for Clinical NLP. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 10632-10641).
- Meta AI. (2024). *Llama 3.1: Advancing conversational AI*. Retrieved from <https://ai.meta.com/blog/meta-llama-3-1/>
- Nazi, Z. A., & Peng, W. (2024, August). Large language models in healthcare and medical domain: A review. In *Informatics* (Vol. 11, No. 3, p. 57). MDPI.
- Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., & Belinkov, Y. (2024). LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. *arXiv preprint arXiv:2410.02707*.

- OpenAI. (2024). *Research*. OpenAI. Retrieved December 9, 2024, from <https://openai.com/research/>
- Unsloth AI. (2024). *Unsloth AI documentation*. Retrieved December 9, 2024, from <https://docs.unsloth.ai/>
- Wang, J., Yang, Z., Yao, Z., & Yu, H. (2024). Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.
- Xu, C., Guo, D., Duan, N., & McAuley, J. (2023). Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.
- Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., & Liu, N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4), 255-263.
- Yang, Y., Mishra, S., Chiang, J. N., & Mirzasoleiman, B. (2024). SmallToLarge (S2L): Scalable Data Selection for Fine-tuning Large Language Models by Summarizing Training Trajectories of Small Models. *arXiv preprint arXiv:2403.07384*.