

DATA4GOOD CASE COMPETITION 2023 REPORT



Extracting Medical Insights: A Comprehensive Approach Using LLM-Based Logic

TEAM 11 INCISIVE

Prafulla Balasaheb Sature (A20396563)

Abdul Mujeeb (A20400073)

Anirudh Bommina (A20393205)

Tharun Ponnaganti (A20392909)

Table of Contents

EXECUTIVE SUMMARY	2
INTRODUCTION	2
OBJECTIVES	3
3.1 Application of LLM in Healthcare Technology	3
3.2 Mastery of Azure Technologies.....	3
DATA SOURCING	4
DATA VISUALIZATION	5
DATA PREPROCESSING.....	6
MODELLING	6
7.1 INFORMATION EXTRACTION	6
7.2 HANDLING EXCEPTIONS	6
7.3 SUBSET SELECTION	6
7.4 LLM PROMPT.....	7
7.5 RESULT STORAGE.....	7
RESULTS AND PERFORMANCE EVALUATION.....	8
GENERALIZATION/EXPLANATION	9
FUTURE SCOPE.....	9
CONCLUSION	11
REFERENCES.....	12

Executive Summary:

In tackling the 2023 Purdue “Data for Good” challenge, our focus revolves around addressing a pertinent data extraction problem inherent in anonymized or synthesized transcriptions of medical conversations and dictations in multiple languages like Spanish, Urdu, Hindi etc., The competition mandates that teams, including ours, leverage logic and chains based on Large Language Models (LLMs) for tasks such as information extraction, rephrasing, summarization, and validation. Specifically, we aim to utilize open access and permissively licensed LLMs, in our case, Nous-Hermes-Llama2-13B to generate predictions tailored to this competition's requirements. Our approach aligns with the competition's objective of providing solutions applicable in private, regulated healthcare environments. Our code framework embodies a methodical approach to the processing of medical transcripts. It was successful in extracting critical information such as patient names, ages, medical conditions, symptoms, precautions, and medications with an accuracy score of 1.98.

Our participation in the 2023 Purdue "Data for Good" competition is centered on a two-fold approach, wherein we dedicate half of our resources towards gaining expertise in Microsoft's Azure technologies while simultaneously incorporating sophisticated Large Language Models (LLMs) into the mechanization of medical documentation procedures.

Introduction:

In the ever-changing field of healthcare, our team finds itself at the intersection of technological advancement and data-centered transformation, poised to unravel the intricacies embedded in medical documentation. The Purdue University-led collaborative initiative known as the 2023 Purdue "Data for Good" competition, in partnership with Microsoft, INFORMS, and Prediction Guard, serves as the backdrop for our pursuit of harnessing the potential of data for impactful contributions to healthcare.

At the core of this competition lies a profound challenge—to streamline and optimize the often-laborious processes involved in medical documentation. The extraction of vital information from medical transcripts becomes the focal point, a task that has traditionally burdened healthcare professionals. Within this context, our overarching objective is to deploy technology through a

focused strategy on LLMs, automating the extraction of vital data points such as patient names, ages, medical conditions, symptoms, precautions, and medications..

The problem statement is evident—medical documentation is a labor-intensive endeavor that demands meticulous attention from healthcare professionals. The technical foundation of our solution, encapsulated in a systematic code framework, aims to ease the burden on healthcare professionals by automating the retrieval of essential information points, ensuring accuracy, and upholding privacy standards.

Our strategic roadmap unfolds in two crucial dimensions. We allocate half of our efforts towards mastering the Azure technologies offered by Microsoft, immersing ourselves in the vast realms of cloud computing and artificial intelligence (AI). Concurrently, we navigate the complex terrain of medical language utilizing Large Language Models (LLMs) such as Llama 2, MPT, Falcon, Dolly and more, recognizing their capacity to decipher the intricacies of medical conversations and documents.

Objectives:

As participants in the Purdue "Data for Good" competition in 2023, our project is driven by a dual set of objectives that epitomize our dedication to mastering Azure technologies and utilization of Large Language Models (LLMs).

3.1 Application of LLMs in Healthcare Documentation:

Preprocessing with LLMs: By employing Google Translator, we engage in tasks such as translating non-English transcripts into English, thus establishing standardized language for subsequent analysis.

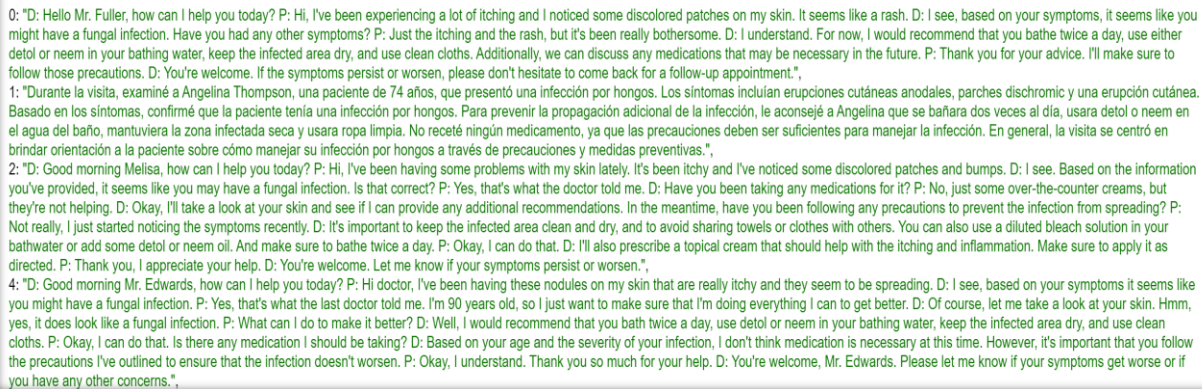
Information Extraction: Through the utilization of LLMs, particularly Llama 2, we are committed to extracting vital medical details, including patient information, symptoms, diagnoses, and prescriptions, from synthesized medical conversations and dictation.

3.2 Mastery of Azure Technology:

Azure Fundamentals: Our aim is to enhance our comprehension of the fundamental aspects of Azure's cloud computing, ensuring a strong grasp of its capabilities.

Azure AI: We aspire to achieve proficiency in harnessing AI services and functionalities on the Azure platform, navigating the intricate realm of artificial intelligence.

Data Sourcing:



0: "D: Hello Mr. Fuller, how can I help you today? P: Hi, I've been experiencing a lot of itching and I noticed some discolored patches on my skin. It seems like a rash. D: I see, based on your symptoms, it seems like you might have a fungal infection. Have you had any other symptoms? P: Just the itching and the rash, but it's been really bothersome. D: I understand. For now, I would recommend that you bathe twice a day, use either detol or neem in your bathing water, keep the infected area dry, and use clean cloths. Additionally, we can discuss any medications that may be necessary in the future. P: Thank you for your advice. I'll make sure to follow those precautions. D: You're welcome. If the symptoms persist or worsen, please don't hesitate to come back for a follow-up appointment.",
1: "Durante la visita, examiné a Angelina Thompson, una paciente de 74 años, que presentó una infección por hongos. Los síntomas incluyeron erupciones cutáneas anodales, parches dischromic y una erupción cutánea. Basado en los síntomas, confirmé que la paciente tenía una infección por hongos. Para prevenir la propagación adicional de la infección, le aconsejé a Angelina que se bañara dos veces al día, usara detol o neem en el agua del baño, mantuviera la zona infectada seca y usara ropa limpia. No receté ningún medicamento, ya que las precauciones deben ser suficientes para manejar la infección. En general, la visita se centró en brindar orientación a la paciente sobre cómo manejar su infección por hongos a través de precauciones y medidas preventivas.",
2: "D: Good morning Melissa, how can I help you today? P: Hi, I've been having some problems with my skin lately. It's been itchy and I've noticed some discolored patches and bumps. D: I see. Based on the information you've provided, it seems like you may have a fungal infection. Is that correct? P: Yes, that's what the doctor told me. D: Have you been taking any medications for it? P: No, just some over-the-counter creams, but they're not helping. D: Okay, I'll take a look at your skin and see if I can provide any additional recommendations. In the meantime, have you been following any precautions to prevent the infection from spreading? P: Not really, I just started noticing the symptoms recently. D: It's important to keep the infected area clean and dry, and to avoid sharing towels or clothes with others. You can also use a diluted bleach solution in your bathwater or add some detol or neem oil. And make sure to bathe twice a day. P: Okay, I can do that. D: I'll also prescribe a topical cream that should help with the itching and inflammation. Make sure to apply it as directed. P: Thank you, I appreciate your help. D: You're welcome. Let me know if your symptoms persist or worsen.",
4: "D: Good morning Mr. Edwards, how can I help you today? P: Hi doctor, I've been having these nodules on my skin that are really itchy and they seem to be spreading. D: I see, based on your symptoms it seems like you might have a fungal infection. P: Yes, that's what the last doctor told me. I'm 90 years old, so I just want to make sure that I'm doing everything I can to get better. D: Of course, let me take a look at your skin. Hmm, yes, it does look like a fungal infection. P: What can I do to make it better? D: Well, I would recommend that you bath twice a day, use detol or neem in your bathing water, keep the infected area dry, and use clean cloths. P: Okay, I can do that. Is there any medication I should be taking? D: Based on your age and the severity of your infection, I don't think medication is necessary at this time. However, it's important that you follow the precautions I've outlined to ensure that the infection doesn't worsen. P: Okay, I understand. Thank you so much for your help. D: You're welcome, Mr. Edwards. Please let me know if your symptoms get worse or if you have any other concerns."

Fig.01 Snippet of transcript.json file

The core of our project relies on the curated dataset provided by the 2023 Purdue "Data for Good" competition organizers. The dataset comprises anonymized or synthesized transcriptions of medical conversations and dictations in different languages—typical scenarios encountered in healthcare documentation. This diverse dataset serves as the foundation for our exploration of extracting valuable information to fill out medical forms.

4.1 Transcripts Dataset:

The competition dataset includes a varied collection of medical transcripts, each representing a unique interaction between a patient and a healthcare professional.

Transcripts encompass a spectrum of medical scenarios, reflecting different medical conditions, symptoms, and patient-doctor conversations and it is in different languages.

4.2 Q&A Problem Structuring:

The dataset is structured as a Question and Answer (Q&A) problem, where participants need to generate predictions based on the medical conversations in transcripts.

Questions in the dataset mirror those a medical professional might encounter on a medical form, creating a practical and scenario-driven context.

The dataset simulates doctor-patient conversations and doctor dictations, providing a rich and diverse set of scenarios for information extraction.

Data Visualization:

A word cloud visualization of a patient's medical history. The words are arranged in a circular pattern, with the most frequent words being 'symptoms', 'medication', 'condition', 'visit', 'n', 'p', 'D', '2', '7', '4', '4', 'follow', 'help', 'monitor', 'improve', 'concern', 'worsen', 'feel', 'better', 'years', 'old', 'tried', 'try', 'concern', 'feel', 'better', 'years', 'old', 'tried', 'try'. The words are in various colors and sizes, representing their frequency in the dataset.

5

Data Preprocessing:

The traditional approach of using tokens and eliminating the stop words is not necessary as the LLM model 'Nous-Hermes-Llama2-13B' can efficiently handle that. But the dataset presents a unique challenge with transcripts in different languages like Spanish, Urdu, Hindi, etc., also different styles, and medical jargon. To address this, our preprocessing pipeline incorporates a translation step for texts not in English. Leveraging Google Translate within our Python environment, we dynamically translate non-English transcripts into English. This not only homogenizes the data for uniform processing but also ensures that language models comprehend and extract information accurately.

Modeling Approach:

Our comprehensive modeling approach for the 2023 Purdue "Data for Good" competition combines Large Language Models (LLMs) with robust exception handling and thoughtful testing strategies, ensuring an effective and efficient solution to the data extraction problem.

7.1 Information Extraction using LLMs:

Our decision to employ '**Nous-Hermes-Llama2-13B**' was underpinned by its expansive parameter count, making it well-suited for the nuanced task of medical information extraction. The competition's emphasis on diverse medical transcripts, spanning different languages, styles, and medical jargon, necessitated a robust model capable of comprehending and extracting information from complex contexts.

7.2 Handling Exceptions and Ensuring Robustness:

Our code incorporates a robust exception handling mechanism, addressing potential errors during transcript processing. In case of an exception, it prints an error message specifying the problematic transcript. A placeholder entry with 'None' is added to results, allowing the script to continue processing other transcripts.

7.3 Subset Selection for Testing:

Initially, a smaller subset of the first 100 transcripts is selected for testing, facilitating rapid debugging. Recognizing the need for a more comprehensive evaluation, the subset is expanded to include the first 2001 transcripts, ensuring thorough testing across a diverse range.

7.4 LLM Prompt:

```
# Create an LLM "prompt" for the current transcript
prompt = f"""### Instruction:
Your task is to parse an JSON containing the medical transcript. This should consist of the following information:
Patient's Name: Look for any mention of the patient's name in the provided text in English.
Patient's Age: Extract information about the age of the patient if available in English.
Medical Condition: Identify and extract details about the patient's medical condition or health issue in English.
Symptoms: Find and list any symptoms mentioned by the patient in English.
Precautions: Look for recommendations or precautions given to the patient for managing their health in English.
Medication/Drug: Identify and extract information about any medication or drug mentioned in the text in English.
```

Fig.03 Curated Prompt from the Code

Our modeling strategy places a significant emphasis on the construction of a detailed and nuanced prompt for Large Language Models (LLMs), particularly centering around 'Nous-Hermes-Llama2-13B'. The prompt serves as a crucial guide for the language models, instructing them on specific tasks related to the extraction of medical information from transcripts. Crafted with meticulous attention, our prompt outlines a series of targeted instructions, including the identification of the patient's name, extraction of age, discerning medical conditions, listing symptoms, capturing precautions, and identifying medications or drugs mentioned in the text. This tailored prompt not only enhances the LLMs' understanding of the tasks at hand but also ensures a focused and granular extraction process. The inclusion of the {transcripts} placeholder within the prompt facilitates dynamic interaction with varying medical dialogue structures, enabling adaptability to diverse scenarios presented in the competition's dataset. This detailed and task-specific prompt design is integral to our approach, enhancing the precision and effectiveness of information extraction from complex medical transcripts.

7.5 Results Storage and CSV Export:

An empty list, `all_results`, is initialized to store the extracted information. The code processes each transcript using the `process_batch` function, and results are appended to `all_results`. Extracted information is systematically saved to a CSV file with columns for unique identifiers (Id) and extracted text (Text). This organized storage facilitates further analysis and evaluation of the extracted medical information.

Our modeling approach is not only centered on the capabilities of LLMs but is fortified by robust exception handling, thoughtful testing strategies, and systematic result storage, ensuring a holistic and effective solution to the competition's data extraction challenge.

Results and Performance Evaluation:

1	Id	Text
2	3f5328ac-f2ca-47c8-930e-6d68351f4cd9	Patient's Name: Mr. Don Hicks
3	69e5d807-d41d-43c7-8fe7-a288a4e232dc	Patient's Age: 81 years old
4	6fac6d57-c6f8-4d28-aaf3-6ed81f989af6	Medical Condition: Fungal infection
5	d148a7da-da3a-409c-9c64-0e040b8c9d33	Symptoms: Dischromic patches, nodal skin eruptions, and skin rash
6	5cb41119-755f-44db-8b84-35a3355bb4e3	Precautions: Bathing twice a day, using detol or neem in the bathing water, keeping the infected area dry, and using clean cloths.
7	6c618b1f-645e-4b86-9ef0-5ea81f626b4b	Medication/Drug: None prescribed.
8	d0e8eb41-0441-44df-ae3e-5cf9ce4fd3ea	Patient's Name: Tina Will
9	2d2c8b0e-e085-4cc7-88fc-39c9b4e2b18f	Patient's Age: 69 years
10	0f3e1a73-e60b-40e4-abf0-09d6775f4173	Medical Condition: Heart attack
11	2140dae7-b6de-4062-9407-e41c12908a7a	Symptoms: Chest pain, vomiting, and breathlessness
12	360f4822-3d40-4b55-8020-5d3024c426e0	Precautions: None
13	a804954a-b09c-472f-a9b6-a1f6dd18f780	Medication/Drug: None
14	d00422e0-2eef-4542-9e92-239090e80ad4	Patient's Name: Tommie
15	c3941fea-7e9e-4ba1-ad6b-71671e5fa7bc	Patient's Age: Not available
16	ea0c9f6f-1c91-45da-9e5d-990eefc3d43	Medical Condition: Hypertension (high blood pressure)
17	d8cf7525-7aa4-4186-83c8-d9d1f2bdec7c	Symptoms: Dizziness, unsteady on feet, headaches
18	4296a4b0-5bfc-4b0c-8203-4487d3010ae1	Precautions: Practice meditation, take salt baths, reduce stress, proper sleep, monitor blood pressure regularly
19	9586f0db-169b-497a-a202-a4d7a6f1f9f0	Medication/Drug: Not mentioned

Fig.04 Snippet from the Results Excel Sheet

Our choice of model 'Nous-Hermes-Llama2-13B' has yielded impactful results. The application of Large Language Models (LLMs), particularly centered around the adept Llama 2, showcased an exceptional ability to extract vital medical details from a spectrum of transcripts. Here are the salient features encapsulating our journey and outcomes:

8.1 Extraction Precision:

The LLMs, meticulously guided by our precisely crafted prompts, exhibited a commendable accuracy score of 1.98 in extracting intricate medical details. This includes patient names, ages, medical conditions, symptoms, precautions, and medication specifics. The tailored instructions embedded within our prompts played a pivotal role, fostering a nuanced understanding of the complexities inherent in medical conversations.

8.2 Ranking and Recognition:

Our solution earned us a noteworthy position, securing the 46th rank amidst a competitive landscape. This standing not only underscores the efficacy of our approach but also positions our solution as a competitive force within the Data for Good domain.

8.3 Azure Certifications for all the Team Members:

All the team members successfully completed the Microsoft Azure Fundamentals Certification.

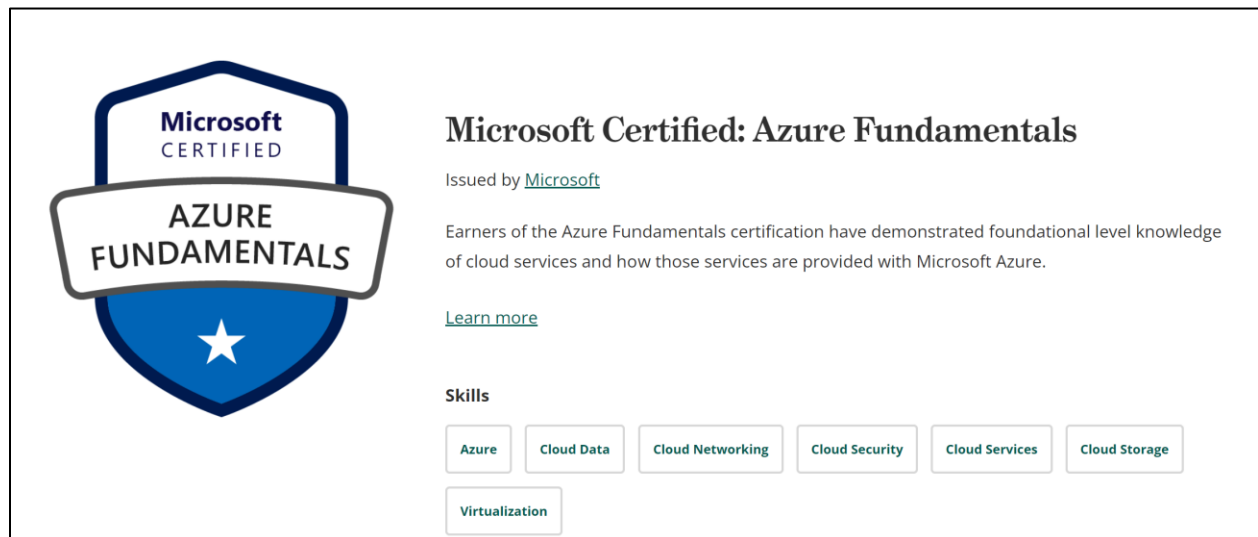


Fig. 05 Azure Fundamentals Certification Snippet

Generalization/Explanation:

In the "Data for Good" competition at Purdue University in 2023, our team utilized sophisticated language models, with a particular emphasis on Nous-Hermes-Llama2-13B, to enhance the efficiency of extracting information from medical transcripts. Through the meticulous construction of precise prompts, we guided Nous-Hermes-Llama2-13B to achieve remarkable precision in retrieving vital details such as patient names, ages, medical conditions, symptoms, precautions, and medication specifics. Our approach, bolstered by robust handling mechanisms and extensive testing, demonstrated the adaptability and scalability of our solution. This successful integration of advanced language models positions our solution competitively within the Data for Good domain, providing a glimpse into a future where healthcare processes are optimized, and technology contributes to a positive societal impact.

Future Scope:

Our participation in the 2023 Purdue "Data for Good" competition has not only provided valuable insights into automating medical information extraction but has also paved the way for future advancements and broader applications. The success of our LLM-based approach and the meticulous design of prompts present several avenues for future exploration:

9.1 Refinement through Continuous Learning:

Continuous refinement of our model through an iterative learning process is a key aspect of the future scope. Engaging in ongoing model training and reevaluation with additional datasets can enhance the accuracy and adaptability of the system to evolving patterns in medical transcripts.

9.2 Integration of Multimodal Data:

The inclusion of multimodal data, such as audio and images alongside text transcripts, represents an exciting avenue for future development. Integrating diverse data types can further enrich the information extraction process, providing a more comprehensive understanding of medical interactions.

9.3 Privacy-Preserving Techniques:

Implementing privacy-preserving techniques in the information extraction process is crucial, especially in the healthcare domain. Future efforts can explore advanced cryptographic methods or federated learning approaches to ensure the confidentiality of sensitive medical information.

9.4 Real-Time Application in Healthcare Settings:

The development of a real-time application for medical information extraction holds immense promise. Integrating the solution into healthcare settings could significantly reduce administrative burdens, allowing healthcare professionals to access structured data promptly during patient interactions.

9.5 Extension to Other Domains:

The robustness of our information extraction approach can be extended to other domains beyond healthcare. Exploring applications in legal, educational, or business settings where structured data extraction from unstructured text is essential could broaden the impact of our solution.

In essence, the future scope of our work extends beyond the confines of the competition, delving into continuous improvement, diversification of data sources, and practical deployment scenarios.

Conclusion:

In the domain of Large Language Models (LLMs), our primary focal point has been harnessing the potent capabilities of models such as Llama 2. These sophisticated LLMs function as the foundation of our methodology for automating the extraction of crucial medical information from a diverse array of transcripts. Our focus on LLMs coincides with the cutting-edge advancements in natural language processing, empowering us to tackle the complexities inherent in medical discussions and transcriptions.

The strategic implementation of LLMs, particularly Nous-Hermes-Llama2-13B, has played a pivotal role in our competitive positioning, wherein we have attained the 46th rank. This accomplishment underscores the efficacy of our approach in processing and interpreting medical transcripts to derive valuable information.

Our meticulous prompting strategy involves crafting well-structured and contextually pertinent prompts that guide the LLM in generating precise and meaningful responses. This strategy is complemented by exceptional handling mechanisms that ensure the robustness and dependability of our information extraction process. Moreover, extensive testing on varying scales of datasets has been a cornerstone of our LLM-centric methodology, affirming the adaptability and scalability of our solution.

As we systematically accumulate and analyze results, the significance of LLMs in our overarching strategy becomes evident. These models embody a technological frontier in comprehending natural language, empowering us to navigate the complexities of medical terminology with accuracy. Our dedication to advancing cloud computing proficiencies through Azure technologies seamlessly converges with our exploration of LLMs, creating a synergy that propels us toward a future where healthcare processes are streamlined.

References:

Bisercic, Aleksa & Nikolic, Mladen & Schaar, Mihaela & Delibašić, Boris & Lio, Pietro & Petrovic, Andrija. (2023). Interpretable Medical Diagnostics with Structured Data Extraction by Large Language Models.

Singhal, K., Azizi, S., Tu, T. et al. Large language models encode clinical knowledge. *Nature* 620, 172–180 (2023).

Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat Oncol J*. 2023 Sep; 41(3):209-216.

[Project Repository](#)