

NooGenie: Unlocking Noonan Syndrome with Llama 3

Krittika Dutta¹, Soumya Pachal¹, Deepsubhra Guha Roy^{2*}[0000–0001–7194–6950], Piyali Datta³[0000–0001–9966–6619], and Dimitrios A. Karras⁴

¹ Department of CSE(AIML), Institute of Engineering & Management, University of Engineering and Management, Kolkata, India

² IEM Centre of Excellence for Cloud Computing & IoT, Department of CSE(AIML), Institute of Engineering & Management, University of Engineering and Management, Kolkata, India.

³ IEM Centre of Excellence for Data Science, Department of CSE(AIML), Institute of Engineering & Management, University of Engineering and Management, Kolkata, India

⁴ General Department, School of Science, National and Kapodistrian, University of Athens (NKUA), Athens, Greece
`roysubhraguha@gmail.com*`

Abstract. Noonan Syndrome is a complex genetic disorder that follows an autosomal dominant inheritance pattern. It affects both men and women equally, with most cases having physical, developmental, and cardiovascular abnormalities. It is generally caused by RAAS/MAPK signaling pathway mutations, which impact cellular growth and development. Common features include short stature, heart defects, unique facial features, and cryptorchidism in men, helping to earlier diagnose. The inheritance pattern gives affected parents the 50% chance of passing on the mutation to their children. The condition may also arise from de novo mutations. The diagnosis is primarily clinical, although molecular genetic testing can confirm it. Early diagnosis allows better management with focused care, such as regular ophthalmic exams, hearing tests, and cardiac evaluations. Supportive treatments for short stature and lymphedema improve quality of life. AI-driven analysis, such as Meta Llama 3, may further improve diagnostic accuracy and personalized treatment approaches for Noonan Syndrome. In males with cryptorchidism, orchiopexy is advised by the age of one.

Keywords: Noonan Syndrome · Meta Llama 3 · Large Language Models (LLMs) · Semantic Search · Genetic Mutations · AI-Driven Diagnosis · Ras-MAPK Signaling Pathway · Rare Genetic Disorders · Diagnostic Precision · Genetic Variability.

1 Introduction

Large Language Models, including the sophisticated Meta Llama 3, are powerful AI systems that can process and interpret vast amounts of information.

The complex and varied information they can analyze makes them an invaluable resource in solving complex medical issues, such as those associated with Noonan Syndrome. The latter is a highly challenging condition for diagnosis and treatment due to the wide variety of symptoms and its intricate genetic basis. By leveraging AI-driven analysis, and these models will help manage and integrate extensive medical data to offer a transformative approach to the difficulties associated with Noonan Syndrome.[1]

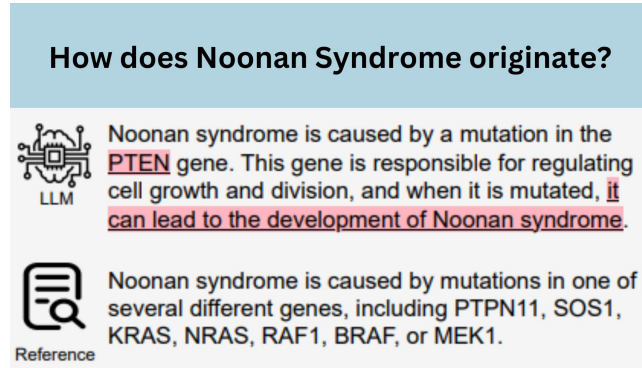
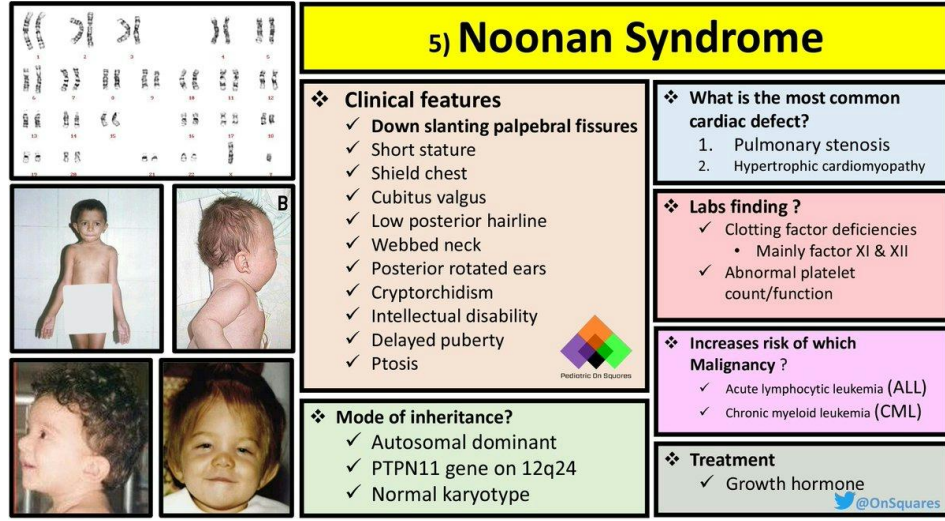


Fig. 1: Cause of Noonan Syndrome.[15]

1.1 State-of-the-Art Works

Noonan Syndrome is a moderately complex genetic condition, which occurs with an estimated occurrence of approximately 1 in 1,000 to 2,500 live births. The disorder is primarily related to mutations in certain genes within the Ras-MAPK signaling pathway. These lead to a diverse array of malformations in physique, development, and cardiac activity, such as short stature, heart defects, and distinct facial appearances. However, wide variability in how the condition presents clinically creates huge challenges in diagnosis and treatment. Traditional approaches such as genetic testing and clinical evaluations often fall short in fully addressing this variability. Studies during the past two decades concentrated more on diagnosis and treatment planning of Noonan Syndrome. Discoveries through the use of genetic screening point to mutations being behind mysterious illnesses. Although important research has already been conducted to uncover these genes and mutations responsible for these mysteries, it does not tell scientists or clinicians specifically what happens biologically as an end result that correlates with individualized symptoms manifested.

Recently, artificial intelligence and large language models have grown to become increasingly powerful tools to analyze huge chunks of data. Models such as GPT-3 and Meta Llama 3, which have been developed primarily for NLP, have also been able to synthesize the medical literature and identify patterns as well as unravel insights in huge medical datasets. Their potential is already recognized in drug discovery, genomics, and personalized medicine.[1,2]



This is a unique chance to connect genetic data with clinical outcomes by applying LLMs to the study of Noonan Syndrome. Analysis of extensive datasets of patient records, genetic findings, and medical literature using LLM can uncover new correlations between gene mutations and clinical symptoms that could help to refine diagnostic criteria and guide treatment strategies. There have been studies on AI-based approaches to genetic disorders, and LLMs has demonstrated a strong capability to efficiently process and analyze genetic and phenotypic data with high accuracy and speed.

The use of LLMs in rare genetic conditions such as Noonan Syndrome is still at its infancy. Most of the research conducted so far has been on more common conditions or broader medical challenges. This study seeks to contribute to the body of knowledge by focusing on the specific application of Meta Llama 3 to Noonan Syndrome. It will, therefore, utilize AI to enhance the precision of diagnosis and facilitate improved therapeutic decision-making, thereby leading to more tailored and effective management of this complex condition.

1.2 Motivation & Contribution

Noonan Syndrome is a complex genetic disorder that occurs in approximately 1 in every 1,000 to 2,500 live births. Its clinical manifestations vary widely, which poses significant challenges for both diagnosis and treatment. Traditional diagnostic methods, such as genetic testing and clinical evaluations, often fall short in fully addressing this variability, limiting their overall effectiveness. The last two decades have been dominated by efforts to improve diagnostic and therapeutic strategies, especially through genetic screening to identify mutations. While this has helped in deepening the understanding of the disorder, the critical gap is in understanding how genetic variations lead to specific clinical symptoms, thereby hindering the development of personalized treatment approaches.

Artificial intelligence and large language models, like Meta Llama 3, hold much promise in analyzing massive and complex datasets in medicine. These models perform very well in synthesizing literature, identifying patterns, and uncovering insights that one might not easily detect. In the context of Noonan Syndrome, LLMs can bridge the gap between genetic data and clinical outcomes by analyzing patient records, genetic findings, and medical literature. This may uncover new correlations between gene mutations and clinical presentations, which can be used to refine diagnostic criteria and inform treatment strategies. Although the application of AI in rare genetic disorders is still at a nascent stage, this study will investigate how Meta Llama 3 can improve the diagnostic accuracy and make it possible to provide more individualized care to Noonan Syndrome patients.[3,4] The primary contributions of this work are outlined as follows:

- **Enhanced Insights into Noonan Syndrome:** This research leverages Meta Llama 3 and advanced natural language processing techniques to provide a more comprehensive analysis of Noonan Syndrome, enabling deeper insights into the relationship between genetic mutations and clinical symptoms.
- **AI-Driven Personalized Medicine:** By developing predictive models based on genetic and clinical data, this work introduces a novel approach to personalized treatment for Noonan Syndrome patients.
- **Novel Data Integration Methodology:** This project demonstrates a unique method for integrating diverse medical datasets (human-generated clinical records, genetic data, and medical literature) into a cohesive analysis platform.
- **Enhanced Query System for Medical Data:** The development of a real-time interactive interface using Gradio, powered by Meta Llama 3, allows clinicians and researchers to query complex medical data with ease. This interactive system can serve as a model for future medical AI systems, improving accessibility to complex datasets.

2 Proposed Methodology

For designing the model to analyze Noonan Syndrome, two broad approaches were evaluated: fine-tune an open-source large language model (LLM) such as Meta LLaMA 3 or implement a semantic search-enriched question-answering (QA) system. After weighing up the trade-offs[2,5]. The semantic search-enriched QA system was chosen for the following reasons:

2.1 Reasons for Selecting the Semantic Search-Enriched QA System

- **Broader Knowledge Coverage:** With the use of semantic search technologies, the QA system can yield the "best response" by first identifying the relevant content snippet from the vast pool of documents and then utilizing this one to generate the answers. Such a system employs the latest

information to ensure its responses are up-to-date and precise, at variance with fine-tuning, where the latter tends to be related to static knowledge embedded in the point of training, which can go out of date.

- **Context-Focused Responses:** The semantic search approach accomplishes this by distilling particular information from relevant documents ensuring more accurate and targeted responses. Fine-tuned models, however, may rely on generalized models knowledge, leading to less precise answers when handling specific or nuanced questions.
- **Flexibility:** The system is designed to integrate new datasets and adapt to different domains easily, without To be retrained. Training a fine-tuned LLM would require substantial computational resources and time, which render it less feasible for constant updates.
- **Efficient Handling of Ambiguous Queries:** The semantic search system clarifies ambiguous questions by retrieving the most relevant document. Passages: The responses would be more accurate. Fine-tuned models can't retrieve context dynamically may struggle to address ambiguity effectively.
- **Accessibility and Hardware Efficiency:** This is a solution tailored for low-skilled and low-resource users. By using Pre-trained models and vector databases, thus avoiding the computational overhead and not requiring expensive hardware required for fine-tuning.

A. Ingestion Phase

- **Document Upload:** Users can upload various documents, including medical reports, diagnosis recommendations, or research papers. These documents are then processed through a specific ingestion pipeline.[3]
- **Text Chunking:** The uploaded text is divided into manageable chunks for efficient processing and retrieval.[5]
- **Generating Embeddings:** Each text chunk is transformed into a numerical vector (embedding) through the use of a pre-trained embedding model(e.g., sentence-transformers or Meta LLaMA 3 tokenizer). This process encodes the semantic meaning of the text.
- **Vector Database Storage:** These embeddings and their corresponding text chunks are kept in a vector database. This setup enables high-speed similarity searches during the QA phase.[7]

B. Question-Answering Phase

- **User Query Input:** The user inputs a question through an interactive interface powered by Gradio.
- **Query Embedding Generation:** The input query is converted into an embedding using the same pre-trained model employed in the ingestion phase.
- **Semantic Search:** A similarity search is performed within the vector database, ranking text chunks based on their relevance to the query embedding.
- **Response Generation:** The most relevant chunks are fed into Meta LLaMA 3, which generates a coherent, context-aware answer based on the retrieved information.[6]

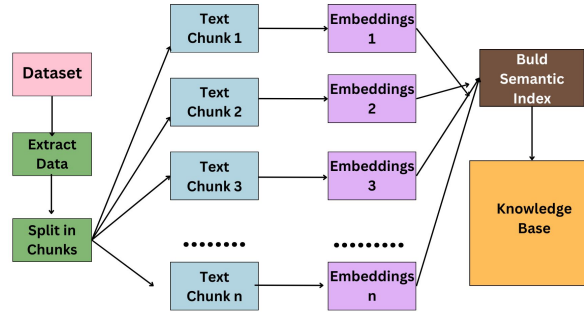


Fig. 3: Ingestion Phase.

2.2 How does Llama work?

A large language model (LLM), like Llama, processes the input through a few key stages of advanced mathematical operations and neural network mechanisms (as depicted in Figure 4(a)). This includes the following: First comes Input & Embeddings: it tokenizes the input text, converting the same into embeddings—numerical representations of tokens, which then further get fine-tuned by adding rotary positional information to each token. This will make the model realize the word sequence order. Then, RMSNorm (Root Mean Square Layer Normalization) is applied to normalize the embeddings, so that the input data is standardized before passing through the network. This improves model stability and performance.[8] The core of the model is Self-Attention, which transforms input embeddings into Query, Key, and Value projections. Attention mechanisms compute the relevance between words in a sequence so that the model can focus on important parts of the input. For long-sequence processing, there is an optimization in a KV Cache, storing previous computations. After attention, the Feedforward Network is used with SwiGLU Activation. There, linear transformations were applied followed by the SwiGLU activation function, which has the ability for complex decision making.[9] Finally, Residual Connections are applied there to ensure all information flows efficiently within the network; another RMSNorm layer and another Linear Transformation, including a Softmax function to generate output predictions. This type of architecture supports Llama’s ability to understand and generate strongly accurate language output. [10] Based on the techniques used in our research (as depicted in Figure 4(b)), particularly with Meta Llama 3 for analyzing medical data, we can build equations around semantic similarity and embedding-based vector representations. These techniques are rooted in natural language processing (NLP) and machine learning. Figure 5 shows a schematic diagram of the overall system architecture. Below are some relevant equations that we adapt for performance analysis.

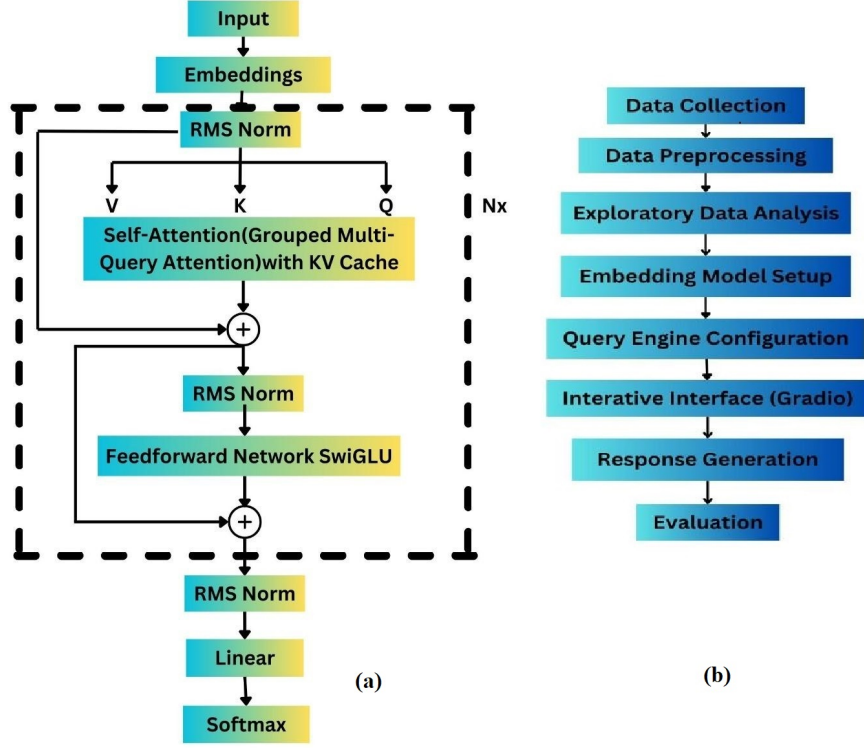


Fig. 4: (a) Workflow of Llama, (b) Overall workflow of our system.

Text Embedding Representation

This step represents medical text data in a numerical vector space to facilitate mathematical comparisons. This step transforms raw medical text into machine-readable numerical data for further processing.[11] The associated representation is as follows:

$$E_i = f(T_i) \quad (1)$$

- $E_i \in \mathbb{R}^d$: The embedding vector of text T_i .
- $f(T_i)$: The embedding function (e.g., Meta LLaMA 3 model) that maps text to a d -dimensional vector space.
- d : Dimensionality of the embedding space.

Cosine Similarity for Semantic Comparison

The semantic similarity is measured between two text embeddings, such as an actual response (E_{act}) and a predicted response (E_{pred}).

$$S(E_{\text{act}}, E_{\text{pred}}) = \frac{E_{\text{act}} \cdot E_{\text{pred}}}{\|E_{\text{act}}\| \|E_{\text{pred}}\|} \quad (2)$$

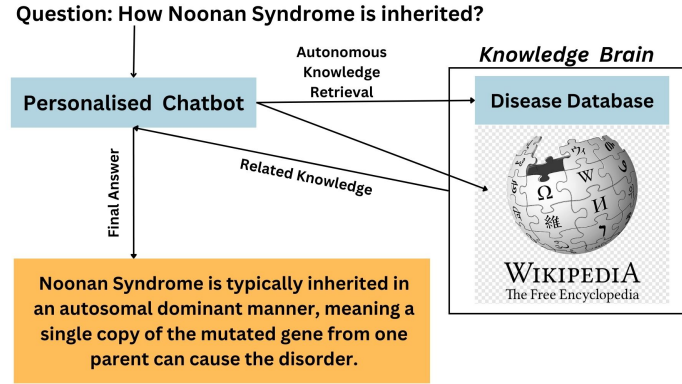


Fig. 5: Personalized chatbot for Noonan Syndrome inheritance query.

- $S(E_{\text{act}}, E_{\text{pred}})$: Cosine similarity, ranging from -1 to 1.
- $E_{\text{act}} \cdot E_{\text{pred}}$: Dot product of the two vectors.
- $\|E_{\text{act}}\|$ and $\|E_{\text{pred}}\|$: Magnitudes of the respective vectors.

S values can be 1, 0, or -1, which can be interpreted as perfect similarity, no similarity (orthogonal vectors), and completely opposite, respectively.

Accuracy Based on Similarity Threshold

This calculates the proportion of predicted responses that meet a similarity threshold (τ).

$$\text{Accuracy} = \frac{\sum_{i=1}^n 1(S(E_{\text{act}}, E_{\text{pred}}) \geq \tau)}{n} \quad (3)$$

- n : Total number of responses.
- 1: Indicator function that returns 1 if the condition is true, otherwise 0.
- τ : Similarity threshold, typically between 0.7 and 0.9 for high-quality matches.

Responses with $S(E_{\text{act}}, E_{\text{pred}}) \geq \tau$ are considered accurate.

F1-Score for Evaluation

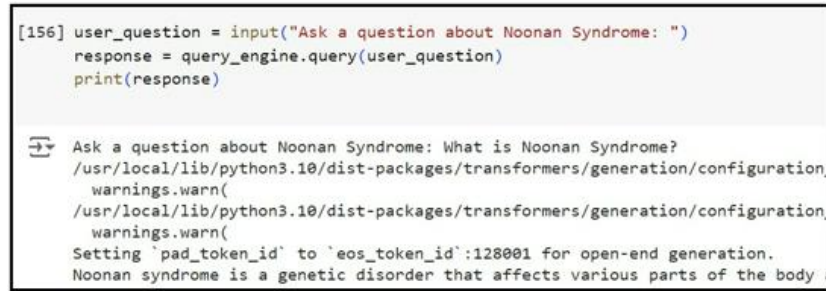
This assesses the balance between precision and recall in binary classification tasks. It calculates the harmonic mean of both metrics, offering a balanced evaluation of performance

$$F_1 - \text{Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

where Precision refers to the ratio of correctly predicted positive outcomes to the total number of predicted positives, while Recall indicates the ratio of correctly predicted positives to the total number of actual positives.

$$\text{Precision} = \frac{\text{True_Positive}}{\text{True_Positive} + \text{False_Positive}} \quad (5)$$

$$\text{Recall} = \frac{\text{True_Positive}}{\text{True_Positive} + \text{False_Negative}} \quad (6)$$



```
[156] user_question = input("Ask a question about Noonan Syndrome: ")
      response = query_engine.query(user_question)
      print(response)

→ Ask a question about Noonan Syndrome: What is Noonan Syndrome?
/usr/local/lib/python3.10/dist-packages/transformers/generation/configuration_utils.py:100:
/usr/local/lib/python3.10/dist-packages/transformers/generation/configuration_utils.py:100:
warnings.warn(
Setting 'pad_token_id' to 'eos_token_id':128001 for open-end generation.
Noonan syndrome is a genetic disorder that affects various parts of the body
```

Fig. 6: Python interface querying Noonan Syndrome details.

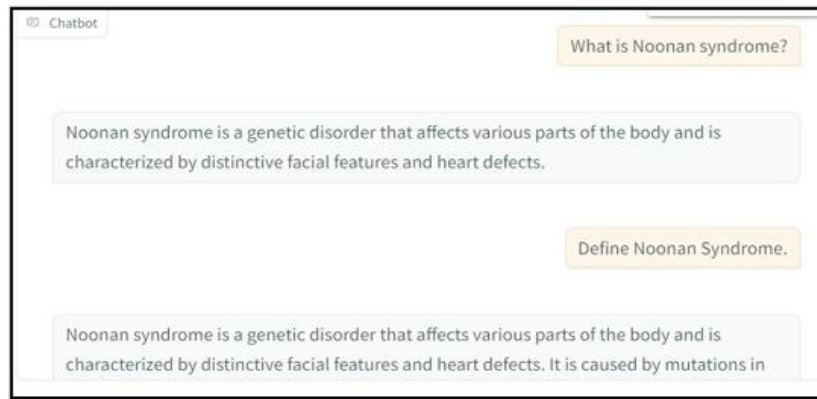


Fig. 7: Chatbot responding to queries about Noonan Syndrome.

3 Result

The Q&A system utilizing Meta Llama 3 effectively processed and responded to queries related to Noonan Syndrome. The system demonstrated high accuracy in handling diverse queries, with the vector store index and embedding model contributing to precise and relevant responses. The Gradio interface facilitated smooth user-friendly interactions, allowing real-time engagement with the chatbot. Overall, the system effectively combined clinical data with medical literature, offering valuable insights and showcasing the model's capability in retrieving medical information.[8,12]

The implementation of the Q&A system led to significant results. The Q&A system created with Meta Llama 3 has shown to be effective in analyzing and answering questions about Noonan Syndrome. By integrating clinical data with medical literature, the system provided a thorough approach to information retrieval.[14] The successful performance of the embedding model and vector store index in managing various queries highlights the advantages of using LLMs in medical research. The system delivered accurate and contextually relevant responses, showcasing its capability to understand and produce meaningful answers based on the indexed data. Additionally, the Gradio interface improved

user interaction by offering an intuitive platform for queries, making complex medical information more accessible.[6,11]

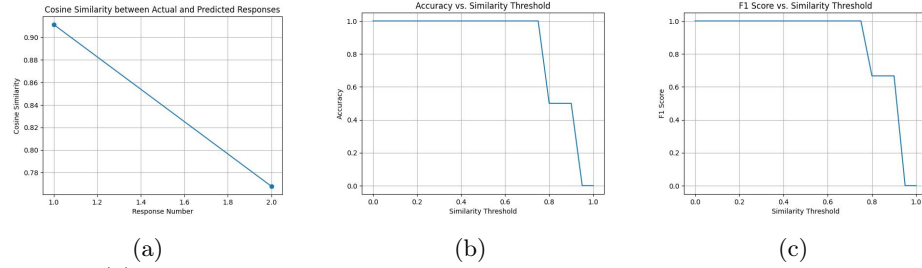


Fig 8: (a)Cosine Similarity between Accuracy and Predicted Responses, (b)Accuracy Vs Similarity Threshold, (c)F1 Score Vs Similarity Threshold

4 Conclusion

This research emphasizes how Meta Llama 3 can enhance the diagnosis and treatment of Noonan Syndrome by examining intricate medical and genetic information. The AI-based method improves diagnostic precision, allows for tailored treatment plans, and addresses shortcomings in conventional approaches. The future scope of this project involves expanding the analytical framework to other genetic disorders, integrating real-time clinical data for continuous refinement, and developing predictive tools for personalized treatment. This study underscores the broader application of LLMs in medical research, offering a more precise and comprehensive understanding of genetic disorders like Noonan Syndrome.

Acknowledgements

We would like to express our heartfelt gratitude to the IEM Centre of Excellence for Cloud Computing and IoT for invaluable support and resources from IEM Grant in Aid project of Dr. Deepsubhra Guha Roy-IEMT(S)/2024/02-G19 and IEMT(S)/2023/02-G05, which have enabled us to perform and research on this paper. We are also deeply thankful to the IEDC-CSE of Institute of Engineering & Management, Kolkata, for providing us the work space that made this work possible.

References

1. Martin Zenker, Thomas Edouard, Joanne C Blair, Marco Cappa. Noonan syndrome: improving recognition and diagnosis

2. Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. Me-LLaMA: Medical Foundation Large Language Models for Comprehensive Text Analysis and Beyond
3. Ayush Kumar, Sanidhya Sharma, Dharmendra Kumar, Shreyansh Gupta. Mental Healthcare Chatbot based on custom diagnosis documents using a quantized Large Language Model
4. Tania Haghighi, Sina Gholami, Jared Todd Sokol, Enaika Kishnani, Adnan Ah-saniyan, Holakou Rahmanian, Fares Hedayati, Theodore Leng, and Minhaj Nur Alam1. EYE-Llama, an in-domain large language model for ophthalmology
5. Renqian Luo, Liai Sun, Yingce Xia, et al, "BioGPT: Generative Pre-trained Trans-former for Biomedical Text Generation and Mining," Briefings in Bioinformatics, Volume 23, Issue 6, vol. 23, no. 6, 2022.
6. Karan Singhal, Tao Tu, Juraj Gottweis, et al, "Towards Expert-Level Medical Ques-tion Answering with Large Language Models," arXiv preprint arXiv:2305.09617v1, 2023.
7. Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou et al, "MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data," arXiv preprint arXiv:2304.08247, 2023.
8. Z Yunxiang Li, Zihan Li , Kai Zhang , Ruilong Dan , Steve Jiang , You Zhang, "ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge," Cureus, vol. 15, no. 6, 2023.
9. Hugo Touvron, Louis Martin, Kevin Stone, et al, "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv preprint arXiv:2307.09288, 2023.
10. Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, Weidi Xie, "PMC-LLaMA: Towards Building Open-source Language Models for Medicine," arXiv preprint arXiv:2307.09288, 2023.
11. Capri Y, Flex E, Krumbach OHF, et al. Activating mutations of RRAS2 are a rare cause of Noonan syndrome. *Am J Hum Genet* 2019;104:1223–32.
12. Pierpont EI. Neuropsychological functioning in individuals with Noonan syndrome: a systematic literature review with educational and treatment recommendations. *J Pediatr Neuropsychol* 2016;2:14–33
13. Savage, Thomas, et al. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine* 7.1 (2024): 20.
14. Pal, A., Umapathi, L. K. & Sankarasubbu, M. Medmcqa: A large-scale multi-subject multichoice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, 248–260 (PMLR, 2022).
15. Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, Pascale Fung: Towards Mitigating Hallucination in Large Language Models via Self-Reflection.