Enhancing Generative AI Output on Rare Diseases: Development of a Database for Retrieval-Augmented Generation System

1st Teppei Okazaki
Graduate School of
Integrative Science and Engineering
Tokyo City University
Tokyo, Japan
g2581412@tcu.ac.jp

2nd Toyofumi Fujiwara
Database Center for Life Sciences
Research Organization of
Information and Systems
Kashiwa, Japan
fujiwara@dbcls.rois.ac.jp

3rd Atsuko Yamaguchi
Department of Design and Data Science
Tokyo City University
Yokomaha, Japan
atsuko@tcu.ac.jp

***Abstract*—Rare diseases are a group of conditions that affect a small number of patients. Definitions vary by country and region; in the United States with diseases affecting fewer than 200,000 individuals, in Europe fewer than 5 cases per 10,000 people are considered rare diseases. With their limited patient populations, often lead to a scarcity of reliable information, leaving attending physicians with insufficient knowledge. While search engines can be used to find information, individuals without specialized expertise face challenges in assessing its accuracy and reliability. Generative AI has recently gained attention as a tool for individuals struggling to access specialized information. However, in fields requiring high accuracy, such as medical information, the limitations of generative AI, which refer to outdated or inaccurate data, can have critical consequences. The purpose of this study is to develop a Retrieval-Augmented Generation(RAG) system to enhance the output of the generative AI regarding rare diseases and to increase the transparency of the underlying information. Furthermore, this research aims to develop a comprehensive database by obtaining literature information on rare diseases from PubMed provided by the National Library of Medicine (NLM)(https://www.nlm.nih.gov/) that will serve as a source of information for the RAG system. Comparative experiments using a generative AI as the evaluator demonstrated that our system achieved ratings equal to or better than the GPT-4o engine for 31 out of 38 randomly selected rare diseases.**

***Index Terms*—Rare disease, Generative AI, Retrieval-Augmented Generation**

I. Introduction

Rare diseases are a group of conditions characterized by a small number of patients, defined as affecting fewer than 200,000 individuals in the United States and fewer than 5 in 10,000 individuals in Europe (1), resulting in a scarcity of available information. Patients, their families, and even their physicians often lack sufficient knowledge about these conditions (2). Although one can attempt to find information on rare diseases through search engines, individuals without specialized knowledge face significant challenges in determining the accuracy and reliability of the information obtained.

PubMed is a medical literature search database provided by the National Library of Medicine (NLM)(https://www.nlm.nih.gov/). With approximately 40 million articles indexed in PubMed, it serves as a vast repository of biomedical literature. While this wealth of information is invaluable, it presents a significant challenge when searching for specific topics such as rare diseases. Identifying and extracting literature related to rare diseases is particularly time-consuming because of the need to locate accurate and reliable sources of information from the vast amount of data. Therefore, the development of a database on rare diseases is essential to this research.

While medical literature serves as a critical source of information, the level of evidence it provides can vary significantly. Factors such as study design, sample size, methodology, and potential biases can influence the reliability and accuracy of findings (3). Rare diseases present challenges in recruiting the necessary number of subjects for cohort studies or clinical trials due to their rarity (4). As a result, the quality and rigor of studies and articles can vary significantly, making not all papers equally reliable. This variation poses difficulties for clinicians and researchers, particularly when making decisions in areas where robust or consistent evidence is limited, as is often the case with rare diseases (5).

Medical Subject Headings (MeSH) (6) is a comprehensive controlled vocabulary used by the NLM to index articles in PubMed and other biomedical literature databases. MeSH terms provide a standardized way to describe and categorize the content of medical articles, facilitating precise and efficient retrieval of relevant information. By organizing information hierarchically and including synonyms and related concepts, MeSH

This work was partially supported by JSPS KAKENHI grant number JP21K12148 and ROIS-DS-JOINT 029RP2024.

enables users to navigate the vast amount of biomedical literature more effectively. In this research, MeSH terms serve as an essential component for determining the format in which papers were researched, contributing to the development of a robust and reliable database.

In recent years, generative AI such as large language models (LLMs) technologies have made remarkable advancements (7; 8). By the mid-2020s, tools and systems once limited to specialists and corporations became widely accessible as personal applications and web services. This ease of use—requiring only a smartphone and an Internet connection—combined with growing social awareness, has contributed to the widespread adoption of generative AI among the general population. Despite their advancements, LLMs are prone to referencing outdated or incorrect information, often leading to hallucinations (9). In the medical domain, such inaccuracies can have serious consequences for patient care and outcomes. If language models are to play a role in medical decision-making, ensuring the accuracy and reliability of the information they provide becomes critical, as errors could directly impact diagnoses, treatment plans, and overall patient safety (10). Retrieval-Augmented Generation (RAG) (11) is known to be an effective technique that offers an intuitively promising solution for dealing with hallucinations (12).

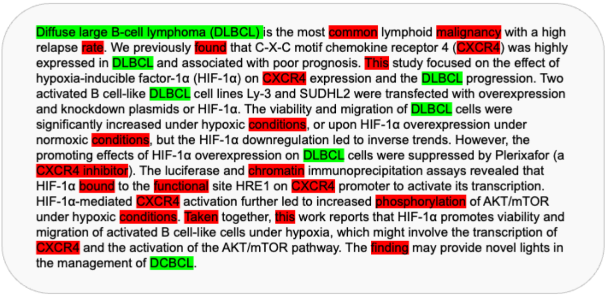
This study addresses the challenge of developing a database specific to rare diseases and combining it with the Open AI's GPT-4o (13) engine to implement RAG to retrieve accurate and reliable rare disease information. The database is constructed by annotating the titles and abstracts of articles registered in PubMed, extracting the relevant disease names using the Mondo ontology (14) ID (Mondo ID), and storing each article with its bibliographic and annotation information, which is also used to implement the RAG system. The database is stored in JSON files for each article, with bibliographic information including PubMed ID (PMID), title, abstract and MeSH, and annotation information including annotated words, their location, annotated disease names, their Mondo ID, disease definition, synonyms and cross-references, subsets is stored. This approach aims to improve the accuracy and reliability of information retrieval for rare disease research and medical applications.

II. Method

The methodology of this study is divided into two main sections: database construction and RAG implementation.

A. Data Construction

In this study, annotation tools are used to extract and annotate titles and abstracts of PubMed literature related to rare diseases. A JSON file is created for each literature, in which the literature information and annotated terms are stored. The literature information contains the literature title, abstract, PubMed ID (PMID), MeSH term and MeSH ID, while the annotated information contains the annotated word, its location, annotated disease name, Mondo ID of the disease name, disease definition, synonym, cross references and subsets are included.



Diffuse large B-cell lymphoma (DLBCL) is the most common lymphoid malignancy with a high relapse rate. We previously found that C-X-C motif chemokine receptor 4 (CXCR4) was highly expressed in DLBCL and associated with poor prognosis. This study focused on the effect of hypoxia-inducible factor-1α (HIF-1α) on CXCR4 expression and the DLBCL progression. Two activated B cell-like DLBCL cell lines Ly-3 and SUDHL2 were transfected with overexpression and knockdown plasmids or HIF-1α. The viability and migration of DLBCL cells were significantly increased under hypoxic conditions, or upon HIF-1α overexpression under normoxic conditions, but the HIF-1α downregulation led to inverse trends. However, the promoting effects of HIF-1α overexpression on DLBCL cells were suppressed by Plerixafor (a CXCR4 inhibitor). The luciferase and chromatin immunoprecipitation assays revealed that HIF-1α bound to the functional site HRE1 on CXCR4 promoter to activate its transcription. HIF-1α-mediated CXCR4 activation further led to increased phosphorylation of AKT/mTOR under hypoxic conditions. Taken together, this work reports that HIF-1α promotes viability and migration of activated B cell-like cells under hypoxia, which might involve the transcription of CXCR4 and the activation of the AKT/mTOR pathway. The finding may provide novel lights in the management of DCBCL.

Fig. 1. Example of Abstract Annotation

1) Data Retrieval: Given the impracticality of creating a comprehensive database encompassing all literature published to date before confirming the effectiveness of this method, the scope of this study was limited to literature registered in PubMed in 2023.

SPARQL is a query language designed for querying and processing RDF data published as Linked Open Data. It enables the retrieval of publicly available data from PubMed.

2) Annotation: In this research, titles and abstracts were annotated separately using an annotation tool to identify relevant rare disease terms. We used Concept Mapper Annotator (15) as the annotation tool. The annotation ontology used is Mondo (14), a comprehensive ontology that integrates the definitions of diseases and extracts the annotated terms and their corresponding Mondo IDs from the annotation results. Each disease registered in the Mondo ontology has a subset, and those that are regarded as rare diseases have a subset ”rare”, so the database in this study only handles literature annotated with disease names that have ”rare” in their subset. **Fig. 1** illustrates how the abstract of an article is annotated. The green highlighted areas represent disease names annotated in the Mondo ontology with the subset ’rare,’ while the red highlighted areas indicate names that are either not annotated in Mondo or lacked the ’rare’ subset designation.

3) MeSH Data Retrieval: MeSH is a hierarchical thesaurus that serves to unify the expression of diseases and other terms to avoid variation, each with its own unique ID and term. It exists in PubMed RDF as a URI with a unique ID at the end. For example, Gastric Cancer is unified in ”Stomach Neoplasms” and is managed with a unique ID of ”D013274” and a URI of ”http://id.nlm.nih.gov/mesh/D013274”. Since the MeSH terms retrieved via SPARQL are provided as URIs, it is challenging to immediately identify the associated MeSH concepts without additional processing or searching. To address this, the MeSH RDF is utilized to extract and store the unique IDs and labels of the MeSH terms, making the data more accessible and easier to handle.

B. Retrieval-Augmented Generation Implementation

This study uses OpenAI's Assistants API in the GPT-4o (13) engine as a LLM and implements RAG.

1) GPT-4o Integration: Up to 10 literature, categorized by evidence level, are selected in ascending order of evidence level and formatted as JSON files for input into the Assistants API to generate the final results. Each JSON file includes the PMID, title, abstract, and study characteristics of the respective literature . **Listing 1** shows the example of the JSON file, the name of the disease treated in this example is ”inborn error of immunity”. The name of disease is contained in the first key ”name”, and under the ”Literatures” key, there are key of tiers, which categorizes literature into tiers based on their level of evidence. The structure of these tiers is detailed in **Listing 2** and is derived from https://openmd.com/guide/levels-of-evidence (16). Tiers are determined by the Publication Type associated with the MeSH terms. This classification follows a tree structure, with Tier 2 encompassing all MeSH terms under the 'Clinical Trial' MeSH node. The Assistants API instructions include a description of how the contents of a given JSON file are structured. And the prompt specifies that the entered disease name should be explained using the literature information provided in the given JSON file. Additionally, it requires that the title and Publication Type of the referenced literature are always included in the explanation.

Listing 1. Example of a JSON file used to implement a RAG (disease name: ”inborn error of immunity”)

```
{
  "name": "inborn error of immunity",
  "Literatures": {
    "Tier1": [
      {
        "pmid": "37328647",
        "title": "Purine Nucleoside Phosphorylase Deficient...",
        "abstract": "Purine nucleoside phosphorylase deficient...
               hypogammaglobulinaemia...",
        "Publication_Type": "Systematic Review"
      },
      {
        "pmid": "37924455",
        "title": "Systematic review: malignancy in hyper-IgE...",
        "abstract": "The hyper-IgE syndrome (HIES)... malignancy.",
        "Publication_Type": "Systematic Review"
      },
      {
       "pmid": "38051840",
        "title": "Th17 Pathway and Candida...",
        "abstract": "Candida albicans... IL-17A/F in susceptibility.",
        "Publication_Type": "Systematic Review"
      }
    ],
    "Tier2": [
      {
        "pmid": "36399712",
        "title": "P I 3K  inhibitor trial",
        "abstract": "Activated PI3K pathway... deficiency in APDS.",
        "Publication_Type": "Randomized Controlled Trial"
      }
    ],
    "Tier3": [
      {
        "pmid": "36209879",
        "title": "RF-EMF exposure and hypersensitivity",
        "abstract": "Some individuals attribute symptoms... transient
               condition.",
        "Publication_Type": "Cohort Studies"
      },
      {
        "pmid": "38025336",
        "title": "BCG-related SCID cases",
        "abstract": "SCID... BCG vaccine at birth.",
        "Publication_Type": "Cohort Studies"
      },
      {
        "pmid": "37240507",
        "title": "Warning signs in IEI",
        "abstract": "IEI refers to genetic disorders... severe PIDs.",
        "Publication_Type": "Cohort Studies"
      },
      {
        "pmid": "37588055",
        "title": "MIS-C and genetic variants",
        "abstract": "Multisystem inflammation... disease dynamics.",
        "Publication_Type": "Cohort Studies"
      },
      {
        "pmid": "37404832",
        "title": "Corrigendum: IEI exome diagnosis",
        "abstract": "Correction to IEI article...",
        "Publication_Type": "Cohort Studies"
      },
      {
        "pmid": "37544429",
        "title": "COVID-19 in congenital immunodeficiency",
        "abstract": "SARS-CoV-2 risk... immune error factors.",
        "Publication_Type": "Cohort Studies"
      }
    ],
    "Tier4": [
      {
        "pmid": "36873640",
        "title": "HLH induced by SARS-CoV-2 and EBV in APECED",
        "abstract": "APECED... HLH triggered by viral infections... fatal
               outcome.",
        "Publication_Type": "Case Reports"
      }
    ]
  }
}
```

Listing 2. Structure of Tier List

```
{
    "Tier1": [
        "Systemtic Review",
        "Meta-Analysis"
    ],
    "Tier2":{
        "Clinical Trial":[
            "Adaptive Clinical Trial",
            "Clinical Trial, Phase I",
            "Clinical Trial, Phase II",
            "Clinical Trial, Phase III",
            "Clinical Trial, Phase IV",
            {
                "Controlled Clinical Trial":[
                    "Equivalence Trial",
                    "Pragmatic Clinical Trial",
                    "Randomized Controlled Trial"
                ]
            }
        ]
    },
    "Tier3":[
        "Cohort Studies",
        "Case-Control Studies"
    ],
    "Tier4": "Case Reports"
}
```

III. Computational Experiment

We compared the normal ChatGPT-4o with our method incorporating our RAG implementation based on their respective outputs, focusing on the propensity for hallucinations.

A. Experimental Setup

The first step in the output generation process involves using the GPT-4o engine to produce an overview of the disease. The prompt instructs the model to describe the entered disease across five key categories: (1) demographics, (2) symptoms and signs, (3) causes, (4) diagnosis, and (5) treatment. These categories are selected to align closely with the structure of the Rare

Disease Database provided by the National Organization for Rare Disorders (NORD) (https://rarediseases.org/rare-diseases/), which serves as the reference set for correct answers. This generated overview serves as a baseline for comparison.

Next, the RAG-implemented GPT-4o engine is used to generate output for comparison. The prompt instructed the model to take the baseline overview as input and supplement it with additional information derived from the provided article information, as shown in **Listing. 1**.

B. Evaluation Method

LLMs configured as evaluators have been proposed as an effective method for comparing and assessing generated outputs (17). In this experiment, the GPT-4o model was utilized as an evaluator to rate three aspects of the generated outputs—correctness, comprehensiveness, and currentness—using the NORD Rare Diseases Database (https://rarediseases.org/rare-diseases/) as the reference set for correct answers. These dimensions were selected to provide a well-rounded and practical assessment of language model performance, particularly in real-world applications that involve retrieving or generating factual information.

Correctness evaluates whether the model produces factually accurate and logically consistent responses. This criterion is fundamental to ensure that the output does not contain hallucinations or wrong information.

Comprehensiveness measures the degree to which the response covers all relevant aspects of the question or prompt. This dimension reflects the model's ability to provide complete and informative answers rather than partial or shallow ones.

Currentness assesses whether the information presented is up to date, especially in response to time-sensitive or evolving topics. One of the known causes of hallucinations in LLM is their reliance on outdated information.

IV. Results

Table I shows the results of the evaluation of each disease overview generated by the regular GPT-4o model and the RAG-implemented GPT-4o model. Normal Total and RAG Total each implies that the total score of ratings which given on a 5-point scale of 3 items by the GPT-4o model set up as the evaluator. RAG Higher is a flag that sets the higher RAG score to 1 and the lower score to 0. 31 of the 38 diseases scored equal or higher by the RAG-implemented model. In many of the cases that scored lower than the GPT-4o, the prompts were not recognized correctly and the results did not follow the format given in the instructions. **Table II** presents the average scores of the normal GPT-4o and the RAG-implemented GPT-4o models across three evaluation criteria: correctness, comprehensiveness, and currentness. For each model and each evaluation criterion, we computed the average score over all cases. This can be formally expressed as follows. Let M be the set f models, K the set of evaluation criteria, and N the number of test cases. Let $s_i^{(m,k)}$ denote the score given to model $m \in M$ for criterion $k \in K$ on the i-th case. The average score $\bar{s}^{(m,k)}$ for model m on criterion k is computed as:

$$\bar{s}^{(m,k)} = \frac{1}{N} \sum_{i=1}^{N} s_i^{(m,k)}$$

The normal GPT-4o model demonstrated slightly higher performance in correctness, while the RAG-implemented GPT-4o outperformed in comprehensiveness and currentness.

TABLE I
COMPARISON OF NORMAL GPT-4O AND RAG-IMPLEMENTED GPT-4O SCORES ACROSS DISEASES

Disease Name	Normal Total	RAG Total	Difference	RAG Higher
neutropenia	15	15	0	1
amyloidosis	14	15	+1	1
burning_mouth_syndrome	13	15	+2	1
Reunion_island_Larsen_syndrome	12	13	+1	1
adrenoleukodystrophy	12	15	+3	1
autism_spectrum_disorder	11	13	+2	1
vasculitis	12	15	+3	1
graft_versus_host_disease	14	12	-2	0
neuroblastoma	12	15	+3	1
lymphoid_leukemia	13	15	+2	1
neuralgia	12	11	-1	0
glioma	13	15	+2	1
Wiskott-Aldrich_syndrome	14	15	+1	1
patent_ductus_arteriosus	13	15	+2	1
pulmonary_fibrosis	13	15	+2	1
His_bundle_tachycardia	15	15	0	1
mast_cell_activation_syndrome	15	11	-4	0
neurodegeneration_with_brain_iron_accumulation_2A	12	14	+2	1
granulomatosis_with_polyangiitis	11	13	+2	1
Sjogren-Larsson_syndrome	15	15	0	1
pneumocystosis	15	13	-2	0
neuroendocrine_neoplasm	12	15	+3	1
interstitial_lung_disease_2	12	15	+3	1
SATB2_associated_disorder	15	15	0	1
central_precocious_puberty	15	15	0	1
hypotonia-cystinuria_syndrome	15	15	0	1
spondyloepiphyseal_dysplasia	15	12	-3	0
catastrophic_antiphospholipid_syndrome	15	13	-2	0
myeloid_leukemia	15	15	0	1
hemorrhagic_disease	12	12	0	1
Down_syndrome	14	15	+1	1
transposition_of_the_great_arteries	15	10	-5	0
iron_poisoning	11	15	+4	1
well-differentiated_liposarcoma	14	14	0	1
osteochondritis_dissecans	15	15	0	1
severe_acute_respiratory_syndrome	15	15	0	1
lymphopenia	10	13	+3	1
classic_familial_adenomatous_polyposis	11	14	+3	1

V. Discussion

When LLMs are used as evaluators, there tends to be a bias toward LLM-generated summaries, with a preference for these over human-generated summaries (17). **Table I** highlights several instances where the standard GPT-4o received a perfect score of 15 points. This outcome may be attributed to GPT-4o sharing the same evaluation framework during both generation and evaluation, leading to outputs that align closely with the evaluation criteria and thus receive higher ratings. This bias could potentially disadvantage summaries generated using RAGs, which incorporate and utilize external knowledge. However, the fact that the RAG-implemented LLM scored equally well or better in many cases indicates that its performance may surpass what is reflected in the evaluation scores alone.

Table II demonstrates that the RAG-implemented GPT-4o outperforms in comprehensiveness and currentness but falls short in correctness. As noted earlier, many instances where the RAG-implemented GPT-4o

TABLE II
Comparison of Normal GPT-4o and RAG-Implemented GPT-4o Scores on Each Criterion

Category	**Score**		
	Correctness	**Comprehensiveness**	**Currentness**
Normal GPT-4o	4.6579	4.3684	4.3421
RAG-Implemented GPT-4o	4.5263	4.6579	4.7632

scored low in **Table I** were due to prompts not being correctly recognized, resulting in unexpected outputs. This issue significantly affected the correctness score, leading to a lower average correctness score for the RAG-implemented GPT-4o model. On the other hand, the RAG-implemented GPT-4o model achieved higher scores for Currentness, likely because the content of the articles incorporated the latest treatments and recent discoveries that the normal GPT-4o models were unaware of. This indicates that developing a system specifically optimized for providing up-to-date information could be both possibly more effective and meaningful.

Notably, the RAG-implemented model exhibited distinct advantages in handling diseases with limited or complex descriptions in existing literature. For example, diseases such as ”neuroblastoma” and ”amyloidosis” saw substantial improvements in scores due to the integration of high-quality evidence from the constructed database. These results underscore the potential of retrieval-augmented generation in improving the quality of AI outputs in domains requiring high accuracy and reliability.

Cases of low scores for models implementing RAGs may be attributed to incomplete or ambiguous annotations in the database. These findings underscore the critical importance of high-quality annotations and highlight opportunities for further refinement in the data construction process to enhance model performance.

In addition to quantitative scores, qualitative observations highlighted the improved ability of the RAG-implemented model to contextualize rare disease information, providing richer and more actionable outputs for users. This qualitative aspect aligns with the study's objective of enhancing the utility of generative AI for rare disease research and clinical applications.

VI. Conclusion

This study aimed to address the challenges of accessing accurate and reliable information on rare diseases by developing a RAG-implemented system and constructing a specialized database. The RAG-implemented GPT-4o model demonstrated superior performance in many cases, achieving equal or higher scores for 31 out of 38 diseases compared to the standard GPT-4o model. This improvement highlights the potential of retrieval-augmented generation in enhancing the quality of generative AI outputs in medical domains. By annotating Pubmed articles and using the Mondo ontology to summarise what diseases the article is about, the system enabled users to obtain more accurate and useful output. We hope this research demonstrates the potential of combining advanced AI techniques with domain-specific resources to revolutionize research and clinical decision-making in the field of rare diseases.

VII. Limitations and Future Work

Although the results are positive, several limitation that remain for future work.

1) Bias in LLM-based assessments and the reliance on accurate annotation tools are important considerations that warrant careful discussion in future research. While these factors were not directly measured in this experiment, their potential impact on evaluation outcomes and system performance cannot be overlooked.
2) There is concern that the system's effectiveness is limited to cases where the exact name of the disease is known by the patient, their family, or the attending physician, or where the disease has already been narrowed down to a small set of potential names.To address this limitation, it would be desirable to develop a system that incorporates multiple input steps, such as identifying the disease name through a simulated medical interview or by entering symptoms. This enhancement would enable users who are unfamiliar with the disease name to accurately identify it.
3) In this study, the database was experimentally constructed using only literature published in 2023. As the experimental results of this study were positive, it is clear that better results can be achieved by referring to more comprehensive and up-to-date information, so further results are expected by applying the same method to store all available article information in the database. Additionally, the need for semi-automatic updates to incorporate newly published papers into the database will be addressed in subsequent research.
4) Currently, the system developed in this study operates as a Python program and is therefore limited to terminal-based use. However, the target users of this system are patients with rare diseases and their families, who require reliable information about

these conditions, as well as doctors who are not specialists in rare diseases. To better serve these users, future work should focus on transforming the system into a web application or similar platform, making it more accessible and user-friendly for a broader audience.

Acknowledgment

We would like to thank Mr Nobutaka Mitsuhashi, researcher at the Database Center for Life Science, for allowing us to use the SPARQL endpoints in PubMed. We are also grateful to the members of my lab for their support and feedback on this research.

References

[1] S. C. Groft, M. Posada, and D. Taruscio, “Progress, challenges and global approaches to rare diseases,” *Acta Paediatrica*, vol. 110, no. 10, pp. 2711–2716, Oct 2021, ©2021 Foundation Acta Paediatrica. Published by John Wiley & Sons Ltd. [Online]. Available: https://doi.org/10.1111/apa.15974

[2] A. Schieppati, J.-I. Henter, E. Daina, and A. Aperia, “Why rare diseases are an important medical and social issue,” *The Lancet*, vol. 371, no. 9629, pp. 2039–2041, Jun 2008. [Online]. Available: https://doi.org/10.1016/S0140-6736(08)60872-7

[3] S. Kamath and G. Guyatt, “Importance of evidence-based medicine on research and practice,” *Indian Journal of Anaesthesia*, vol. 60, no. 9, pp. 622–625, Sep 2016. [Online]. Available: https://doi.org/10.4103/0019-5049.190615

[4] R. C. Griggs, M. Batshaw, M. Dunkle, R. Gopal-Srivastava, E. Kaye, J. Krischer, T. Nguyen, K. Paulus, and P. A. Merkel, “Clinical research for rare disease: opportunities, challenges, and solutions,” *Molecular Genetics and Metabolism*, vol. 96, no. 1, pp. 20–26, Jan 2009. [Online]. Available: https://doi.org/10.1016/j.ymgme.2008.10.003

[5] M. Pai, C. H. T. Yeung, E. A. Akl, A. Darzi, C. Hillis, K. Legault, J. J. Meerpohl, N. Santesso, D. Taruscio, M. Verhovsek, H. J. Schünemann, and A. Iorio, “Strategies for eliciting and synthesizing evidence for guidelines in rare diseases,” *BMC Medical Research Methodology*, vol. 19, no. 1, p. 67, Mar 2019. [Online]. Available: https://doi.org/10.1186/s12874-019-0713-0

[6] C. E. Lipscomb, “Medical subject headings (mesh),” *Bulletin of the Medical Library Association*, vol. 88, no. 3, pp. 265–266, Jul 2000.

[7] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, “A comprehensive overview of large language models,” *arXiv preprint arXiv:2307.06435*, 2023.

[8] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, “A survey on evaluation of large language models,” *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[9] A. Pal, L. K. Umapathi, and M. Sankarasubbu, “Med-halt: Medical domain hallucination test for large language models,” *arXiv preprint arXiv:2307.15343*, 2023.

[10] E. Ullah, A. Parwani, M. M. Baig, and R. Singh, “Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology – a recent scoping review,” *Diagnostic Pathology*, vol. 19, no. 1, p. 43, Feb 2024. [Online]. Available: https://doi.org/10.1186/s13000-024-01464-7

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[12] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, “Retrieval augmentation reduces hallucination in conversation,” *arXiv preprint arXiv:2104.07567*, 2021.

[13] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.

[14] N. A. Vasilevsky, N. A. Matentzoglu, S. Toro, J. E. Flack IV, H. Hegde, D. R. Unni, G. F. Alyea, J. S. Amberger, L. Babb, J. P. Balhoff *et al.*, “Mondo: Unifying diseases for the world, by the world,” *MedRxiv*, pp. 2022–04, 2022.

[15] C. Funk, W. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. B. Cohen, L. E. Hunter, and K. Verspoor, “Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters,” *BMC bioinformatics*, vol. 15, pp. 1–29, 2014.

[16] B. S. Moira Tannenbaum, Stacy Sebastian, “Level of Evidence,” https://openmd.com/guide/levels-of-evidence.

[17] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “G-eval: NLG evaluation using gpt-4 with better human alignment,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2511–2522. [Online]. Available: https://aclanthology.org/2023.emnlp-main.153/