AlzheimerRAG: Multimodal Retrieval Augmented Generation for Clinical Use Cases

Aritra Kumar Lahiri, *Member, IEEE*, Qinmin Vivian Hu, *Member, IEEE*

***Abstract*— Recent advancements in generative AI have fostered the development of highly adept Large Language Models (LLMs) that integrate diverse data types to empower decision-making. Among these, multimodal retrieval-augmented generation (RAG) applications are promising because they combine the strengths of information retrieval and generative models, enhancing their utility across various domains, including clinical use cases. This paper introduces AlzheimerRAG, a Multimodal RAG application for clinical use cases, primarily focusing on Alzheimer's Disease case studies from PubMed articles. This application incorporates cross-modal attention fusion techniques to integrate textual and visual data processing by efficiently indexing and accessing vast amounts of biomedical literature. Our experimental results, compared to benchmarks such as BioASQ and PubMedQA, have yielded improved performance in the retrieval and synthesis of domain-specific information. We also present a case study using our multimodal RAG in various Alzheimer's clinical scenarios. We infer that AlzheimerRAG can generate responses with accuracy non-inferior to humans and with low rates of hallucination.**

***Index Terms*— Alzheimer, Clinical, Context-Aware, Generative AI, Information Retrieval, LLMs, Multimodal, PubMed, RAG, Question-Answering.**

I. Introduction

The high volume and variety of data in medical research offer several opportunities and challenges. Of these, Alzheimer's Disease (AD) is a particularly compelling case study because it is multicausal, involving genetic, biochemical, and environmental factors, and also involves complex clinical presentations. Despite the tremendous progress, few effective methods exist for diagnosing, treating, and preventing Alzheimer's disease (AD). This knowledge gap is further exacerbated by the growing volume and fragmentation across various data modalities, including textual descriptions, clinical trial data, imaging studies, and molecular data. Traditional methods of synthesizing such a large volume of knowledge are ineffective; most have a single-modality approach, which may miss the insights obtained synergistically from integrated data. This gap in methodology underscores the need for a robust, unified framework that can leverage multiple modalities to enhance the retrieval process by making it more context-aware and reducing the retrieval of irrelevant or less pertinent information.

In this research, we describe a novel Multimodal Retrieval Augmented Generation (RAG) application, **AlzheimerRAG**[1], that integrates textual and visual modalities to improve contextual understanding and information synthesis from the biomedical literature. Our primary research objective in implementing multimodal RAG is to enhance context-aware retrieval capabilities by integrating heterogeneous data types, including textual data, images, and clinical trial information from PubMed articles. Existing methods [30], [34], [71], [58] typically focus on textual or visual data separately, leaving a gap for integrated multimodal solutions. The work by [28] laid the groundwork for RAG models by demonstrating how retrieval can enhance the generation capabilities of language models, particularly in knowledge-intensive tasks. Integrating RAG methodologies with multimodal inputs is a burgeoning area of research, as highlighted by Peng et al. [41], who proposed a multimodal RAG system that enhances data synthesis across text and image modalities. In light of these advancements, the novelty of our approach lies in the seamless integration and alignment of multimodal data during the cross-modal attention fusion process. The AlzheimerRAG framework combines rapid, accurate retrieval via object stores with specialized language models, enhancing its capability to address the nuances of multimodal information pertinent to Alzheimer's Disease. We utilize an optimized mechanism for fine-tuning by implementing Parameter-Efficient Fine-Tuning (PEFT) [19] and inducing cross-modal attention fusion to facilitate synergistic information flow between the text and image models. The fine-tuned models are then incorporated into a multimodal RAG workflow, developed as a Python-based web application with a user interface that allows end-users to retrieve context-aware answers from their queries. The target audience of this application includes biomedical researchers for synthesizing Alzheimer's literature and identifying disease trends, clinicians to support diagnosis and treatment planning for AD, and healthcare institutions for clinical trial design and support.

Benchmark datasets such as BioASQ [40] and PubMedQA [21] have been instrumental in measuring the effectiveness of multimodal RAG systems. BioASQ, a large-scale biomedical semantic indexing and Question-Answering (QA) dataset, provides a robust framework for assessing models' retrieval and QA capabilities. Similarly, PubMedQA offers insights into the

"This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) research grant RGPIN/6686-2019."

Aritra Kumar Lahiri is with the Toronto Metropolitan University, Toronto, ON, Canada (e-mail: aritra.lahiri@torontomu.ca).

Qinmin Vivian Hu is with the Toronto Metropolitan University, Toronto, ON, Canada (e-mail: vivian@torontomu.ca).

[1] Video demonstration - https://youtu.be/lR2pDjNSaRg

accuracy of models in handling biomedical queries, making it an essential tool for evaluating AlzheimerRAG's performance against existing benchmarks. In comparative studies, models that integrate multimodal data have been shown to outperform traditional single-modality systems. For instance, models like T5 [6] have been evaluated in the context of biomedical question answering, demonstrating significant gains when multimodal inputs are utilized. This trend reinforces the need for AlzheimerRAG's multimodal framework to enhance the understanding and treatment of AD.

In summary, our research contributions advance the Multimodal RAG domain in AD in the following aspects -

- *Context-Aware Retrieval-Augmented Generation* - Our framework enhances traditional RAG models with context-aware retrieval capabilities that prioritize the relevance of domain-specific information, thereby increasing accuracy and utility in biomedical applications.
- *Advanced Cross-Modal Attention Fusion* - AlzheimerRAG integrates multimodal data more effectively using transformer architectures and cross-modal attention mechanisms tailored to handle heterogeneous data types.
- *RAG user interface* - Our system implements the multimodal RAG as a web-based application using the latest state-of-the-art technologies like LangChain, FastAPI, Jinja2, and FaissDB to provide users with a robust interface for performing biomedical information retrieval tasks through the context-aware question-answering paradigm.
- *Comparable framework with state-of-the-art benchmarks* - We evaluate the capability of the Multimodal RAG application with benchmark datasets like BioASQ and PubMedQA, along with other comparable LLM RAG models. We also study the effectiveness of our AlzheimerRAG against human-generated responses for different clinical scenarios in Alzheimer's disease to gauge the accuracy and hallucination rates of the retrieved answers.

II. Related Work

The AlzheimerRAG framework is developed within the rapidly evolving landscape of multimodal data integration and retrieval-augmented generation techniques, which are becoming increasingly crucial in biomedical research. Recent studies have demonstrated the importance of leveraging multiple data modalities to enhance diagnosis, treatment, and understanding of complex diseases, such as Alzheimer's.

Existing research has highlighted the efficacy of attention mechanisms that span multiple modalities, which are instrumental in synthesizing heterogeneous information sources in medical contexts. For example, the effectiveness of multimodal token fusion for vision transformers [1] is demonstrated, which significantly improves the integration of visual and textual data in medical imaging. Similarly, cross-modal translation and alignment techniques [2] are showcased that facilitate survival analysis, emphasizing the benefits of integrating diverse data types to yield richer insights. Additionally, recent developments in knowledge distillation have further enhanced model efficiency in healthcare applications, as seen in the work [29] on knowledge distillation, which transfers knowledge from a larger model (teacher) to a smaller model (student), retaining performance while reducing computational costs. Various studies have adopted this methodology, notably the work that discovered integrating imaging and genetic data improves predictive outputs in Alzheimer's models [34].

The application of AI in Alzheimer's research has been underscored by studies [35], which leverage multimodal inputs to improve early diagnosis and patient stratification. Other research, such as [36], has focused on using AI to manage Alzheimer's symptoms, demonstrating that AI-driven solutions can provide valuable insights and recommendations for patient care. The BioBERT model [4] represents a significant advancement in biomedical text mining, emphasizing the utility of transformer models fine-tuned for biomedical applications. This model has been foundational in developing various biomedical applications, including those focused on Alzheimer's disease, where precision in information retrieval is critical. RAG methodologies have gained traction in biomedical research for efficiently synthesizing information from large datasets. The work [28] laid the groundwork for RAG models by demonstrating how retrieval can enhance the generation capabilities of language models, particularly in knowledge-intensive tasks. This has profound implications for healthcare, where accurate and timely information retrieval can guide clinical decisions.

Compared to these advancements, the AlzheimerRAG framework combines rapid, accurate retrieval via FaissDB with specialized language models, enhancing its capability to address the nuances of multimodal information pertinent to Alzheimer's Disease.

III. Methodology

The overall **AlzheimerRAG** architecture is described in Figure 1. In the subsequent sections, we describe each system design step, followed by a demonstration of the application with the technical components.

A. Data Collection and Preprocessing

The first step of our process involves collecting relevant articles from PubMed. We accomplish this by writing a Python script that calls the National Centre for Biotechnology Information (NCBI) Entrez Programming Utilities (E-utilities) API to fetch the top 2000 articles from the PubMed repository [37] related to the "Alzheimer's Disease" search term. The articles are fetched in batches per API request, adhering to NCBI API rate limits and sorted by relevance during the retrieval process. We parse each document, collecting the full texts, abstracts, tables, and figures for textual and image retrieval. After that, we clean and normalize the data for the data preprocessing step to ensure consistency and usability. This involves removing hyperlinks, references, and footnotes. We also standardize the figures/diagrams format by converting them to a consistent image format for uniform processing.

B. Textual Data Retrieval

This step retrieves the clinical text data related to Alzheimer's disease (AD) for textual and tabular data processing. In our workflow, for generating the text embedding,

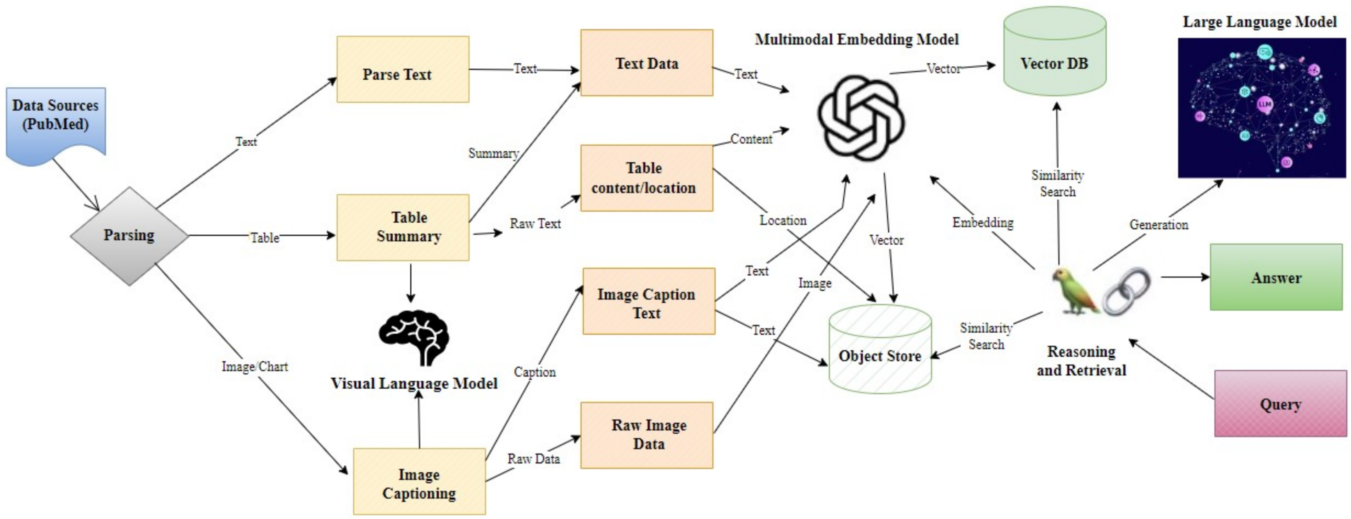


Fig. 1: AlzheimerRAG Architecture

we fine-tune the “Llama-2-7b-pubmed” [22] model by training with the PubMedQA [21] dataset from HuggingFace. The finetuning is employed using parameter-efficient finetuning (PEFT) techniques like QLoRA [23]. Table I outlines the QLoRA parameters and the training argument parameters used for finetuning.

Parameter	Value
QLoRA Parameters	
LoRA attention dimension	64
Alpha parameter for LoRA scaling	16
Dropout probability for LoRA layers	0.1
Training Arguments Parameters	
Number of training epochs	1
FP16/BF16 training	False (True for A100 GPU)
Training batch size per GPU	4
Evaluation batch size per GPU	4
Gradient accumulation steps	1
Enable gradient checkpointing	True
Max gradient norm (clipping)	0.3
Initial learning rate	2e-4
Weight decay	0.001
Optimizer	`paged_adamw_32bit`
Learning rate scheduler	`cosine`
Number of training steps	-1
Warmup ratio	0.03

TABLE I: QLoRA Hyperparameters: LLaMA

*1) **Textual and Tabular Data Processing**:* The extracted data is chunked into structured text and table summaries. Then, a layout model (for tables) and titles are used for candidate sub-sections of the document (e.g., Introduction, Methods, etc). Finally, post-processing is conducted to aggregate text under each title, and further chunking into text blocks is performed for downstream processing based on user-specific flags for each block. After this step, the text embeddings convert the smaller blocks into embedding vectors, which are used for cross-modal attention fusion.

C. Image Retrieval

For the generation of feature embeddings that capture image details from the PubMed articles, we fine-tune the “LlaVA” (Language and Vision Assistant Model, version 2) [17] model using the official LLaVA repo with the Llama-2 7B backbone language model [20]. LLaVA combines pre-trained language models (such as Vicuna or LLaMA) with visual models (such as CLIP’s [74] visual encoder) by converting visual features into embeddings that are compatible with the language model. Using the fine-tuned approach preserves the strengths of the large language model while lowering computational requirements, making it ideal for resource-limited environments and quick adaptation to new data. The hyperparameters are presented in Table II. QLoRA uses the 4-bit NormalFloat, which is explicitly designed for customarily distributed weights, thereby further reducing memory usage.

Parameter	Value
`lora_enable`	`True`
`lora_r`	`128`
`lora_alpha`	`256`
`mm_projector_lr`	`2e-5`
`bits`	`4`
`learning_rate`	`2e-4`
`weight_decay`	`0.001`
`warmup_ratio`	`0.03`

TABLE II: LlaVA Hyperparameter

D. Cross-Modal Attention Fusion

Cross-modal attention fusion is a mechanism that facilitates interaction between different modalities, within our current scope, specifically between text and images. It allows a model to selectively focus on relevant parts of both modalities by computing attention weights. These weights are used to modulate the embeddings from each modality, enabling a richer and more comprehensive representation. In our context, the cross-modal attention fusion ensures that the integrated textual and visual data contribute meaningfully to medical information retrieval. The process steps of Cross-Modal Attention Fusion are detailed in Figure 2 as a sequence diagram.

The three steps associated with this process are described below -

- Generate query, key, and value vectors from the text and image embeddings from Sections III-B and III-C, respectively.
- Compute the attention scores using the dot-product attention mechanism shown below:

$$\text{scores} = \frac{\text{queries} \cdot \text{keys}^\top}{\sqrt{d_k}} \tag{1}$$

  Where:
  - queries and keys are matrices of size $(n \times d_k)$, with n being the number of tokens and d_k the dimension of each key.
  - d_k is the dimensionality of the keys used for scaling.
  - $\sqrt{d_k}$ scales the dot-product, helping to stabilize gradients in deeper networks.
- Aggregate contributions from both modalities based on attention weights:

$$\text{aggr\_embeddings} = \text{attn\_wts} \cdot \text{values} \tag{2}$$

  Where:
  - attn_wts is a matrix representing the attention scores, with dimensions $(n \times m)$, where n is the number of tokens and m is the dimensionality of each value.
  - values is a matrix of values corresponding to tokens, typically with dimensions $(m \times d)$, where d is the embedding size.

  The resulting aggr_embeddings is a combination of the values weighted by attention.

$$\text{combined\_features} = \text{aggr\_attn}(\text{values}, \text{attn\_wts}) \tag{3}$$

Finally, the combined feature embeddings are indexed as vectors in an object store, which allows quicker retrieval of multimodal data.

E. AlzheimerRAG Demonstration

*1) **System Walkthrough**:* AlzheimerRAG is implemented as a Python Web Application utilizing FastAPI and Jinja2 Templates with LangChain integration. It provides a simple user interface [2] for leveraging efficient multimodal RAG capabilities related to AD. The application [3] is deployed in Heroku, a cloud-based Platform-as-a-Service (PaaS) solution that helps manage seamless continuous integration and deployment. It provides the functionality for information retrieval from user queries. The multimodal RAG component extracts context-aware relevant images as part of the output response. The demo video can be accessed from this link [4].

A sample response from the AlzheimerRAG pipeline user interface can be observed in Fig. 3, where relevant text and images are fetched for a particular user query related to Alzheimer's disease from the embedded PubMed articles.

*2) **Key Technical Components**:* The key technical components are summarized below -

FastAPI for API development: FastAPI is a high-performance web framework for API development that provides an intuitive interface for API development and integrates seamlessly with Python's async capabilities.

Jinja2 for Template Rendering: Jinja2 is a templating engine for Python that offers dynamic template rendering. It serves HTML content from backend data, enabling a seamless and interactive user experience.

FaissDB for embedding Multimodal data: FaissDB [24], a vector DB, is widely used for embedding multimodal data. Embedding is the process of converting content into a numerical representation (i.e., vectors) for large language learning models and is crucial for transforming preprocessed healthcare knowledge into individual vectors. The text and image embeddings are encoded into uniform, high-dimensional vectors and indexed for efficient similarity searches. When a query is made, the reasoning and retrieval component searches the vector space to extract relevant information. The benefit is that it uses an approximate nearest neighbor (ANN) search to quickly locate embeddings in high-dimensional space, which is essential for large-scale applications. The generation component uses the retrieved Multimodal representations to produce outputs in various formats, such as text or images.

LangChain as a Retrieval Agent for Multimodal RAG: The Retrieval Agent is a medium to pinpoint the most relevant knowledge in response to user queries. This process involves using the embedding model to convert the text from the user query into vectors, which are then searched through the vector storage to identify the closest matching vectors. The effectiveness of a Retrieval Agent is closely tied to the underlying framework upon which it is built. Therefore, we utilize the Langchain [18] framework, a premium existing open-source framework, along with LlamaIndex [20] because of its significant advantages in i) Preservation of table data integrity, ii) Streamlining the handling of Multimodal data, iii) Enhanced Semantic Embedding. Together, LlamaIndex and LangChain enhance the context-awareness of extracted content, enabling efficient retrieval and synthesis of information and producing nuanced outputs.

IV. Experimental Results

A. Comparative Evaluation

We compare our AlzheimerRAG application against state-of-the-art techniques in the biomedical domain and evaluate the performance of our methods. In our experiments, we select BioBERT [4], a transformer model fine-tuned on biomedical text, and MedPix [39], which utilizes deep learning for medical image classification. To compare the cross-modal attention fusion, we introduce a naive fusion of text and image modalities among two models, primarily by concatenating the embeddings without significant interaction between the modalities. Among the newer variants, we include PubMedBERT [73], LlaVA-Med [16] and BioRAG [58] in our evaluation.

Table III represents the performance, where it is observed that the AlzheimerRAG, with its multimodal RAG design,

[2] App - https://pubmed-multimodal-rag-ae786f93140b.herokuapp.com/

[3] Source Code - https://tinyurl.com/AlzheimerRAG

[4] Video Demonstration - https://youtu.be/lR2pDjNSaRg

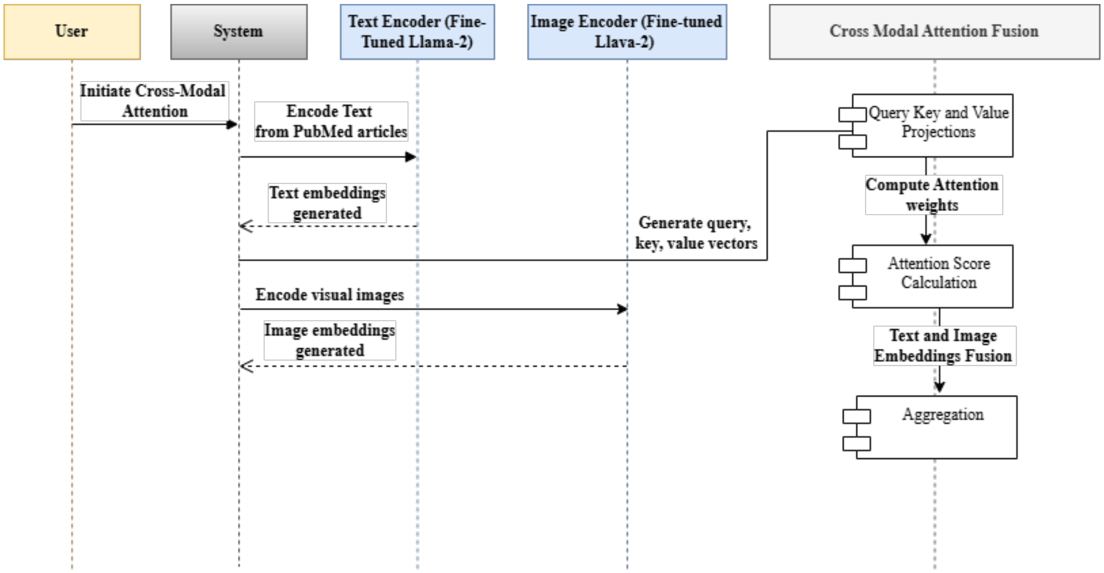
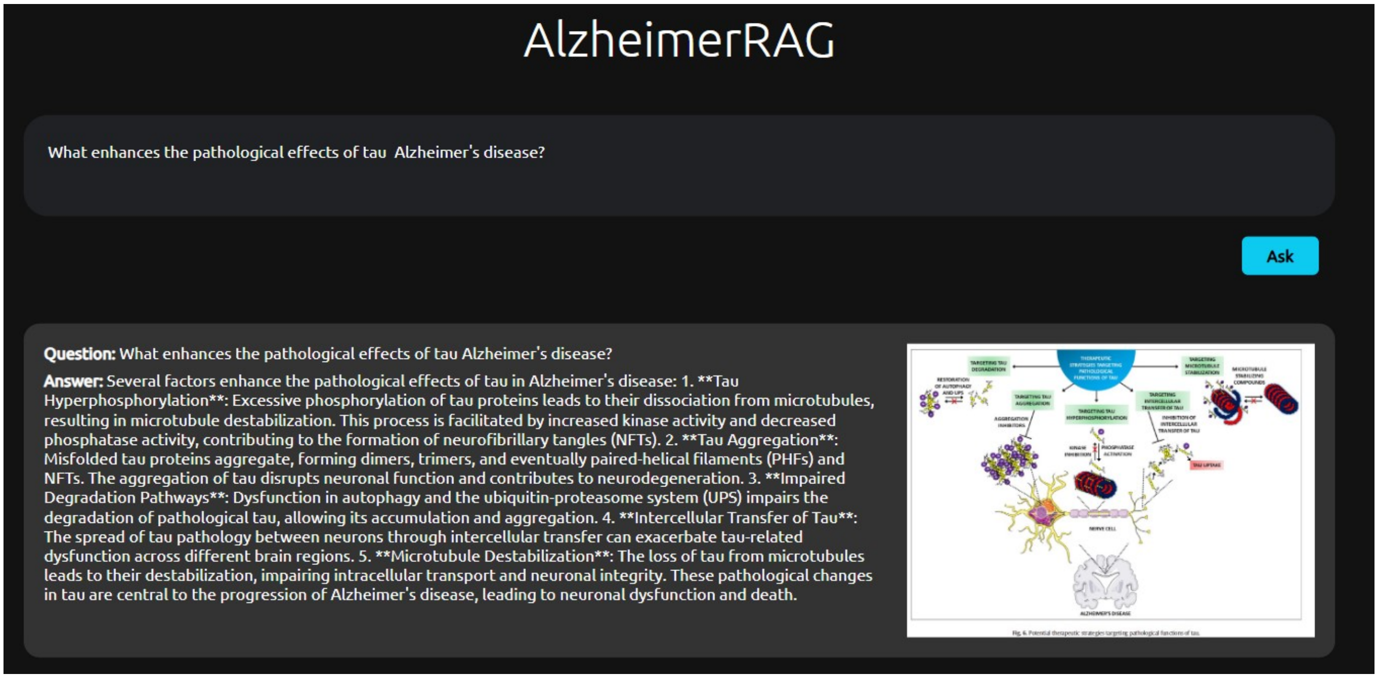

User
System
Text Encoder (Fine-Tuned Llama-2)
Image Encoder (Fine-tuned Llava-2)
Cross Modal Attention Fusion
Initiate Cross-Modal Attention
Encode Text from PubMed articles
Text embeddings generated
Generate query, key, value vectors
Encode visual images
Image embeddings generated
Query Key and Value Projections
Compute Attention weights
Attention Score Calculation
Text and Image Embeddings Fusion
Aggregation

Fig. 2: AlzheimerRAG: Cross-Modal Attention Sequence Diagram


AlzheimerRAG
Ask
THERAPEUTIC STRATEGIES TARGETING PATHOLOGICAL FUNCTIONS OF TAU
TARGETING TAU DEGRADATION
RESTORATION OF AUTOPHAGY AND UPS
TARGETING TAU AGGREGATION
AGGREGATION INHIBITORS
TARGETING TAU HYPERPHOSPHORYLATION
KINASE INHIBITION
PHOSPHATASE ACTIVATION
TARGETING INTERCELLULAR TRANSFER OF TAU
INHIBITION OF INTERCELLULAR TRANSFER OF TAU
TAU UPTAKE
TARGETING MICROTUBULE STABILIZATION
MICROTUBULE STABILIZING COMPOUNDS
NERVE CELL
ALZHEIMER'S DISEASE
Fig. 6. Potential therapeutic strategies targeting pathological functions of tau.

What enhances the pathological effects of tau Alzheimer's disease?

Question: What enhances the pathological effects of tau Alzheimer's disease?

Answer: Several factors enhance the pathological effects of tau in Alzheimer's disease: 1. **Tau Hyperphosphorylation**: Excessive phosphorylation of tau proteins leads to their dissociation from microtubules, resulting in microtubule destabilization. This process is facilitated by increased kinase activity and decreased phosphatase activity, contributing to the formation of neurofibrillary tangles (NFTs). 2. **Tau Aggregation**: Misfolded tau proteins aggregate, forming dimers, trimers, and eventually paired-helical filaments (PHFs) and NFTs. The aggregation of tau disrupts neuronal function and contributes to neurodegeneration. 3. **Impaired Degradation Pathways**: Dysfunction in autophagy and the ubiquitin-proteasome system (UPS) impairs the degradation of pathological tau, allowing its accumulation and aggregation. 4. **Intercellular Transfer of Tau**: The spread of tau pathology between neurons through intercellular transfer can exacerbate tau-related dysfunction across different brain regions. 5. **Microtubule Destabilization**: The loss of tau from microtubules leads to their destabilization, impairing intracellular transport and neuronal integrity. These pathological changes in tau are central to the progression of Alzheimer's disease, leading to neuronal dysfunction and death.

Fig. 3: AlzheimerRAG: Pipeline User Interface Response

retains the lead over LlaVA-Med, a multimodal model for biomedicine, but lacks retrieval capabilities, and BioRAG, a text-only RAG model with PubMed integration.

Against benchmark datasets like BioASQ [40], a large-scale biomedical semantic indexing and question-answering dataset, and PubMedQA [21], developed for QA tasks using a PubMed corpus, we assess the capability of our multimodal RAG by evaluating the document retrieval from given queries and generating accurate answers to Alzheimer related questions from the data against GPT-4. The results are highlighted in Table IV.

The metrics used in our evaluation involve - i) *Precision@k* - which measures the relevance of the top-k(10) retrieved document; ii) *Recall* - which evaluates how many relevant documents are retrieved from the corpus; iii) *Mean Average Precision(MAP)* - which provides the mean average precision values for all queries. In terms of Question-Answering tasks, our evaluation metrics include Accuracy (percentage of correctly answered questions), Exact Match (EM) (percentage of questions that are responded to with exact word matches to the ground truth), and F1-score (considers both precision and Recall for evaluating answer span quality).

We further conducted a comparative qualitative evaluation with other models adaptable for the biomedical domain, focusing on retrieval and question-answering capabilities, as depicted in Table V. The comparison results are presented

Model	Recall	Precision@10	F1
BioBERT	0.72	0.69	0.71
MedPix	0.65	0.62	0.63
BioBERT + MedPix	0.78	0.75	0.76
PubMedBERT	0.80	0.77	0.78
LlaVA-Med	0.82	0.79	0.80
BioRAG	0.87	0.84	0.85
AlzheimerRAG	**0.88**	**0.85**	**0.86**

TABLE III: AlzheimerRAG evaluation with comparative benchmark models

Benchmark	Metrics	AlzheimerRAG	GPT-4
BioASQ	**Precision@10**	0.71	0.70
	Recall	0.80	0.78
	MAP	0.78	0.74
	QA Accuracy	0.72	0.76
	F1 Score	0.75	0.77
PubMedQA	**Accuracy**	0.74	0.78
	Exact Match	0.71	0.73
	F1 Score	0.76	0.79

TABLE IV: Benchmark Dataset Evaluation: AlzheimerRAG vs GPT-4

by considering the GLUE (General Language Understanding Evaluation) [80] and SuperGLUE (Super General Language Understanding Evaluation) [79] benchmarking leaderboards, which serve as metrics for evaluating how well NLP models handle a wide range of complex and straightforward natural language understanding tasks. It can be observed that BioBERT [4] stands out in biomedical applications due to its PubMed pre-training, achieving high precision in retrieval. SciBERT [8], with its broader scientific text pre-training, is more versatile but may need fine-tuning for top biomedical QA tasks. BM25 [12], as a traditional keyword-based model, sets a baseline but lacks deep semantic understanding. ColBERT [5] combines efficient retrieval with semantic depth, though it performs moderately without specific domain adjustments. The BERT+TF-IDF [7] hybrid model strikes a balance between deep learning and traditional retrieval, yielding reasonable results but limited contextual depth. Lastly, T5 [6] excels in QA, especially when fine-tuned for biomedical contexts, leveraging its generative capabilities to achieve high accuracy. In comparison to these, AlzheimerRAG combines fast, accurate retrieval via FaissDB with specialized language models, making it a powerful tool for biomedical retrieval and QA. Its ability to handle text and images offers a significant advantage in contexts where visual data is essential.

B. Ablation Studies

The primary objective of our ablation studies is to assess the significance of critical components in our mechanism. We conducted multiple combinations for our experiments by removing the cross-modal attention mechanism, QLoRA fine-tuning techniques, and multimodal integration. Each of these simulations was designed to isolate and evaluate the impact of the specific component.

By removing **Cross-Modal Attention**, we assess that the model's ability to integrate and leverage text and image data effectively will degrade. We replaced the cross-modal attention mechanism with a simple text and image embedding concatenation. Similarly, we fine-tuned the techniques without **QLoRA** to observe the computation costs and performance. Lastly, we removed the **Multimodal Integration** to check whether the model's overall performance was downgraded.

Each variation's performance metrics were recorded and consolidated in Table VI.

As observed, **Cross-Modal Attention** enables effective interaction between text and image data, with its removal leading to considerable metric degradation. **QLoRA Fine-Tuning** improves precision and clinical relevance with lower computational costs than traditional methods. Lastly, **Multimodal Integration** is essential to the framework's overall effectiveness, as isolating text and image processing substantially reduces recall, precision, and practical application.

V. Clinical Case Study Analysis

We design a case study to evaluate AlzheimerRAG in clinical scenarios related to AD using five primary clinical scenarios - 1) **Early Diagnosis and Monitoring**, 2) **Medication Management**, 3) **Non-Pharmacological Interventions**, 4) **Caregiver Support and Education**, 5) **Behavioral Symptom Management**. The clinical scenario descriptions are provided in Box V.1.

Box V.1: Clinical Scenarios

- **Early Diagnosis and Monitoring**: Assess the system's ability to recommend diagnostic tools and interpret results for early detection.
- **Medication Management**: Determine the ability to guide current medications, potential side effects, and interactions specific to Alzheimer's treatments.
- **Non-Pharmacological Interventions**: Evaluate recommendations for cognitive therapies, physical activities, and lifestyle modifications to slow disease progression.
- **Caregiver Support and Education**: Assess the capability to generate materials for educating caregivers about disease progression and management strategies.
- **Behavioral Symptom Management**: Evaluate the effectiveness of offering strategies to manage common symptoms like agitation, depression, and anxiety.

A. System Evaluation

The clinical scenarios are identified from the medical literature [44], [45] due to their recognized importance in Alzheimer's treatment. 350 responses were evaluated, comprising 50 human-generated, 150 LLM-generated, and 150 LLM-RAG-generated responses. The correctness of the responses was determined based on established guidelines [10] and expert reviews. The validation criteria were factual correctness, absence of hallucinations, and clinical applicability.

Selection of Domain Experts. The domain experts selected in the study were senior researchers from Vector Institute, specializing in the biomedical domain and with a strong familiarity with PubMed literature.

Human-generated answers, provided by domain experts described in Section V-C, were used as a comparison. Figure 4 represents the LLM-RAG performances regarding correct answer percentages.

Evaluation criteria concerned accuracy and safety. Responses with at least 75% accuracy in instructions were

Model	**BioASQ (Retrieval)**	**PubMedQA (QA)**	**Domain**	**Multimodal**
AlzheimerRAG	High Precision & Recall	High Accuracy & F1	Biomedical	Yes
BioBERT	High for Text	Good Accuracy	Biomedical	No
SciBERT	High for Scientific Texts	Moderate, versatile	Scientific	No
BM25 (Baseline)	Fair, keyword-based	Basic QA	N/A	No
ColBERT	Efficient	Moderate	General-purpose	No
BERT+TF-IDF (for QA)	Fair	Moderate	General-purpose	No
T5 (fine-tuned)	Good, versatile	High in QA when fine-tuned	General-purpose	Emerging

TABLE V: Benchmark Models Comparison Across GLUE and SuperGLUE Metrics

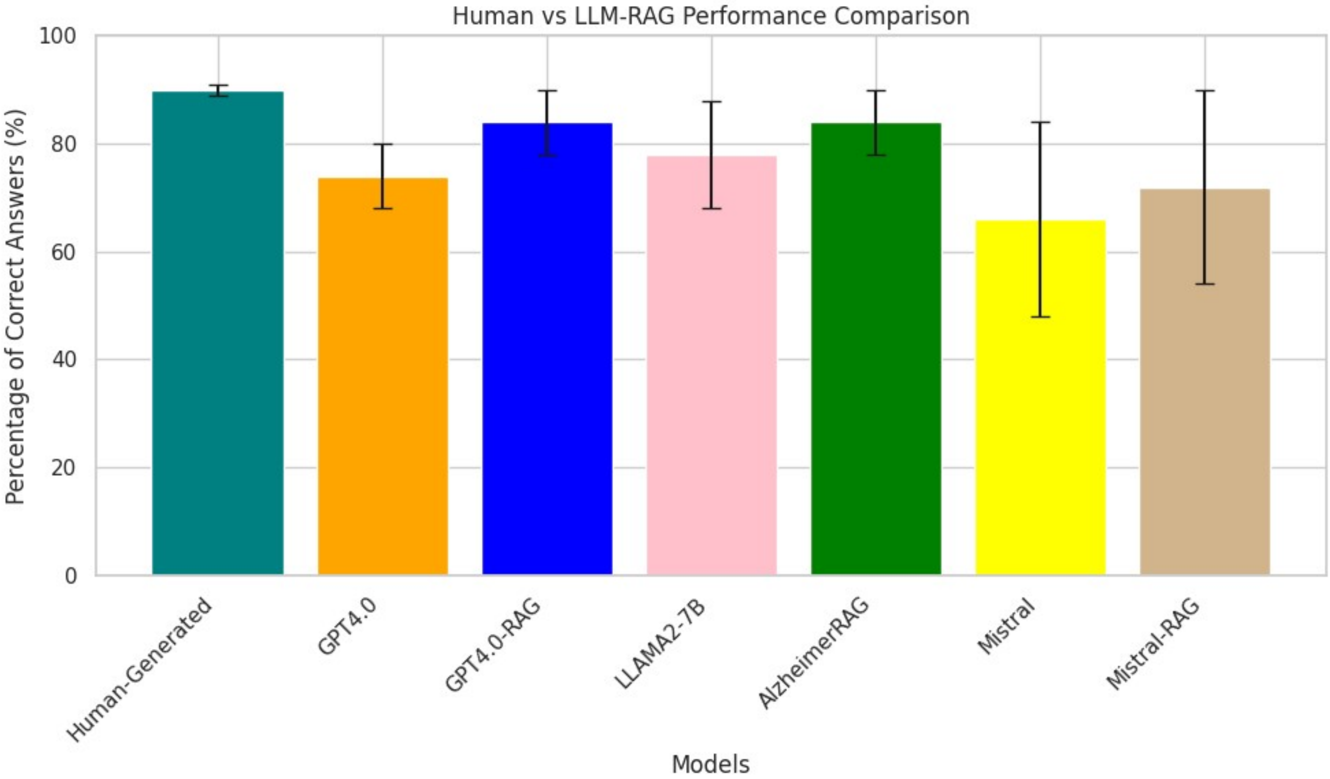


Fig. 4: Percentage of Correct Answers: LLM and LLM-RAG groups

Experiment	**Recall**	**Precision**	**F1 Score**
Baseline (AlzheimerRAG)	0.88	0.85	0.86
Without Cross-Modal Attention	0.75	0.72	0.74
Without QLora Fine-Tuning	0.80	0.77	0.78
Without Multimodal Integration	0.70	0.68	0.69

TABLE VI: Ablation Studies across Multiple Components

deemed “correct.” However, any response containing a significant medical error was categorized as “wrong (hallucination).” Table VII indicates the accuracy and the hallucination rate results, AlzheimerRAG (84%) and GPT4.0-RAG (84%) are the best-performing RAGs compared to the human-generated answers.

B. Statistical Evaluation

We also use statistical tools, such as Cohen’s H and the chi-square test, to evaluate and compare the performance of human-generated responses against the AlzheimerRAG responses.

Cohen’s H [62] is a measure for evaluating the effect size of differences between two proportions. Since the number of answers obtained in our experimental evaluation differs for human-generated and Alzheimer’s responses, this metric can provide us with context on the accuracy of the responses between them.

The chi-square test [63] can assess whether there is a significant association or difference in the responses generated between the two categories. It is helpful to test the differences in the distribution of responses across multiple clinical scenarios.

Table VIII indicates the statistical evaluation results of the different clinical scenarios used in our analysis. In the case of Early Diagnosis and Monitoring, both proportions are the same, inferring there is no difference; hence, the chi-square value is zero. For other clinical scenarios, the results indicate small effect sizes, with notable differences observed only in the Medication Management category. Thus, from the overall results, it can be concluded that there are no major statistically significant differences between the human-generated and the

Models	Early Diagnosis and Monitoring	Medication Management	Non-Pharma Interventions	Caregiver-Support and Education	Behavioral Symptom Management	Total Correct	Hallucinations Present
Human-Generated	10/10 (100.0%)	8/10 (80.0%)	9/10 (90.0%)	9/10 (90.0%)	9/10 (90.0%)	45/50 (90.0%)	-
GPT4.0	9/10 (90.0%)	9/10 (90.0%)	6/10 (60.0%)	7/10 (70.0%)	6/10 (60.0%)	37/50 (74.0%)	(3/50) 6%
GPT4.0-RAG	10/10 (100.0%)	10/10 (100.0%)	8/10 (80.0%)	8/10 (80.0%)	6/10 (60.0%)	42/50 (84.0%)	(3/50) 6%
LLAMA2-7B	9/10 (90.0%)	9/10 (90.0%)	7/10 (70.0%)	7/10 (70.0%)	7/10 (70.0%)	39/50 (78.0%)	(5/50) 10%
Alzheimer RAG	10/10 (100.0%)	9/10 (90.0%)	7/10 (70.0%)	8/10 (80.0%)	8/10 (80.0%)	42/50 (84.0%)	(3/50) 6%
Mistral	8/10 (80.0%)	8/10 (80.0%)	5/10 (50.0%)	6/10 (60.0%)	6/10 (60.0%)	33/50 (66.0%)	(9/50) 18%
Mistral-RAG	8/10 (80.0%)	8/10 (80.0%)	6/10 (60.0%)	7/10 (70.0%)	7/10 (70.0%)	36/50 (72.0%)	(9/50) 18%

TABLE VII: Accuracy & hallucination response: human-generated vs. LLM & LLM-RAG

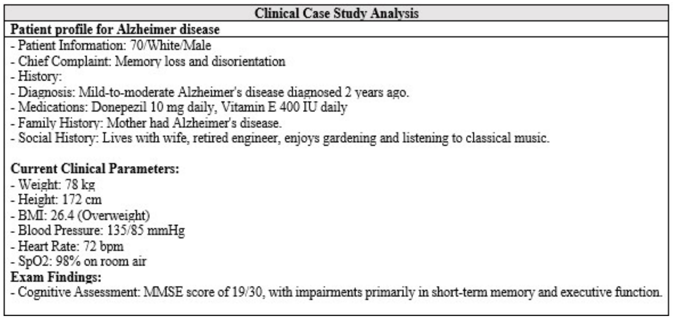
AlzheimerRAG answers.

Clinical Scenarios	**Cohen's h**	**Chi-square**
Early Diagnosis and Monitoring	0	0
Medication Management	-0.234	0.3922
Non-Pharma Interventions	0.404	1.25
Caregiver-Support and Education	0.234	0.3922
Behavioral Symptom Management	0.234	0.3922

TABLE VIII: Comparison between Human and AlzheimerRAG answers

C. AlzheimerRAG Responses vs. Human-Generated Clinical Scenario Responses

To illustrate the AlzheimerRAG outputs with sample human-generated responses from domain experts, we provide two detailed patient profiles for Alzheimer's Disease (AD), as shown in Figure 5 and Figure 6. We input queries with tailored questions answered by the domain experts, in the AlzheimerRAG application for each clinical scenario we designed, and retrieved the responses. The human-generated responses for queries curated for Patient Profile 1 to different clinical scenarios are presented in the Box V.2, V.3, V.4, and V.5. The human-generated responses curated for queries for Patient Profile 2 to different clinical scenarios are presented in the Box V.6, V.7, and V.8. The corresponding AlzheimerRAG responses for Patient Profile 1 are provided in Figures 7, 8, 9, 10. Similarly, for Patient Profile 2, AlzheimerRAG responses are depicted in Figures 11, 12, 13.



Clinical Case Study Analysis

Patient profile for Alzheimer disease

- Patient Information: 70/White/Male
- Chief Complaint: Memory loss and disorientation
- History:
- Diagnosis: Mild-to-moderate Alzheimer's disease diagnosed 2 years ago.
- Medications: Donepezil 10 mg daily, Vitamin E 400 IU daily
- Family History: Mother had Alzheimer's disease.
- Social History: Lives with wife, retired engineer, enjoys gardening and listening to classical music.

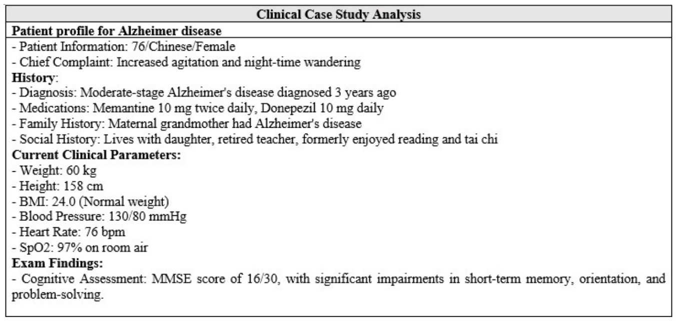
Current Clinical Parameters:

- Weight: 78 kg
- Height: 172 cm
- BMI: 26.4 (Overweight)
- Blood Pressure: 135/85 mmHg
- Heart Rate: 72 bpm
- SpO2: 98% on room air

Exam Findings:

- Cognitive Assessment: MMSE score of 19/30, with impairments primarily in short-term memory and executive function.

Fig. 5: Patient Profile 1: Alzheimer's disease



Clinical Case Study Analysis

Patient profile for Alzheimer disease

- Patient Information: 76/Chinese/Female
- Chief Complaint: Increased agitation and night-time wandering

History:

- Diagnosis: Moderate-stage Alzheimer's disease diagnosed 3 years ago
- Medications: Memantine 10 mg twice daily, Donepezil 10 mg daily
- Family History: Maternal grandmother had Alzheimer's disease
- Social History: Lives with daughter, retired teacher, formerly enjoyed reading and tai chi

Current Clinical Parameters:

- Weight: 60 kg
- Height: 158 cm
- BMI: 24.0 (Normal weight)
- Blood Pressure: 130/80 mmHg
- Heart Rate: 76 bpm
- SpO2: 97% on room air

Exam Findings:

- Cognitive Assessment: MMSE score of 16/30, with significant impairments in short-term memory, orientation, and problem-solving.

Fig. 6: Patient Profile 2: Alzheimer's disease

Box V.2: Patient 1 - Clinical Scenario: Early Diagnosis and Monitoring

Clinical Scenario Query: The patient is a 70-year-old male diagnosed with mild-to-moderate Alzheimer's disease, with his recent MMSE score of 19/30, indicating progressive memory loss. What are the recommended diagnostic or monitoring strategies to assess his disease progression?

Human-Generated response:

- **Cognitive Testing**: Continue regular cognitive assessments such as the Montreal Cognitive Assessment (MoCA) every 6-12 months to monitor changes.
- **Neuroimaging**: Consider an MRI to evaluate brain atrophy and rule out other neurological conditions.
- **Biomarker Analysis**: Use cerebrospinal fluid (CSF) biomarkers or PET imaging, if available, to track amyloid-beta and tau protein levels, particularly in research settings.

Box V.3: Patient 1 - Clinical Scenario: Non-pharmacological interventions

Clinical Scenario Query: How can caregivers support cognitive function in this person with early-stage Alzheimer's disease without medication?

Human-Generated response:
- **Cognitive Stimulation**: Engage the patient in activities like puzzles, memory games, or learning a new hobby.
- **Physical Exercise**: Encourage regular physical exercise, such as walking, to improve brain health.
- **Social Interaction**: Promote social engagement through group activities or community programs.
- **Balance Nutrition**: Ensure a balanced diet rich in omega-3 fatty acids, antioxidants, and vitamins.

Box V.4: Patient 1 - Clinical Scenario: Caregiver Support and Education

Clinical Scenario Query: The patient's wife reports increased difficulty managing his daily routines and is seeking advice to help support his condition better. What are the recommendations for caregiver support and education?

Human-Generated response:
- **Caregiver Education**: Provide educational resources on Alzheimer's disease progression, specific behavioral strategies, and effective communication techniques.
- **Support Groups**: Recommend participation in Alzheimer's support groups to share experiences and strategies with other caregivers.
- **Daily Routine Planning**: Develop structured routines to reduce anxiety and disorientation for both patient and caregiver, integrating familiar activities.
- **Respite Care**: Ensure a balanced diet rich in omega-3 fatty acids, antioxidants, and vitamins.

Box V.5: Patient 1 - Clinical Scenario: Medication Management

Clinical Scenario Query: The patient has been on Donepezil for two years for Alzheimer's management. He is facing increased difficulty remembering recent events despite stable medication. What adjustments or additional treatments should be considered?

Human-Generated response:
- **Medication Review**: Ensure adherence to the prescribed Donepezil regimen and evaluate liver function to monitor for side effects.
- **Add Memantine**: Consider introducing Memantine, especially since symptoms are progressing, offering additional cognitive support in moderate stages.
- **Vitamin E and Antioxidants**: Continue Vitamin E supplementation as a supportive measure, respecting current evidence and guidelines.

Box V.6: Patient 2 - Clinical Scenario: Early Diagnosis and Monitoring

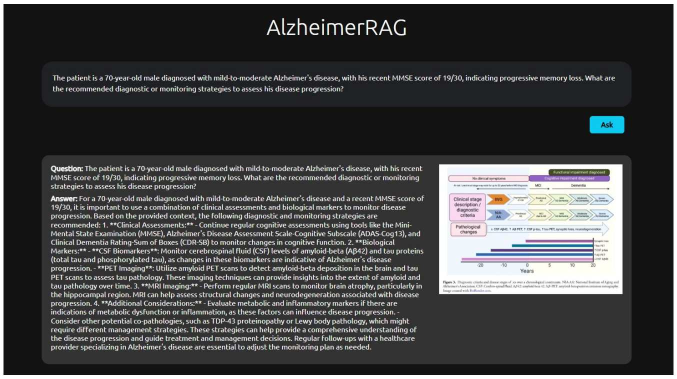
Clinical Scenario Query: What are the early diagnostic criteria for Alzheimer's disease, and how can regular monitoring be implemented for a 76-year-old female presenting with increased agitation and nighttime wandering?

Human-Generated response:
- Use comprehensive clinical evaluations, including patient history and physical examination.
- Conduct cognitive testing using tools such as the Mini-Mental State Examination (MMSE) or Montreal Cognitive Assessment (MoCA).
- Consider neuroimaging techniques, like MRI, to assess brain structural changes.
- Implement regular monitoring every 6 to 12 months to track cognitive decline.
- Engage caregivers to note behavioral changes and symptoms.
- Utilize wearable technology to track daily activities and monitor wandering behavior.

Box V.7: Patient 2 - Clinical Scenario: Medication Management

Clinical Scenario Query: What considerations should be taken into account when managing medications for this 76-year-old female with moderate Alzheimer's experiencing agitation and nighttime wandering?

Human-Generated response:
- Continue cognitive enhancers like Memantine and Donepezil, adjusting doses as needed.
- Evaluate the need for antipsychotic medications for managing severe agitation, considering side effects.
- Regularly review and adjust medications based on patient response and tolerability.
- Monitor drug interactions, especially if the patient is on multiple medications.
- Explore adding new treatments only if current management is ineffective.

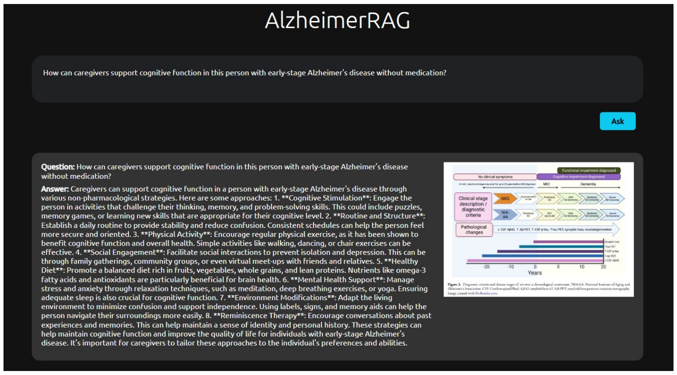


The patient is a 70-year-old male diagnosed with mild-to-moderate Alzheimer's disease, with his recent MMSE score of 19/30, indicating progressive memory loss. What are the recommended diagnostic or monitoring strategies to assess his disease progression?

Question: The patient is a 70-year-old male diagnosed with mild-to-moderate Alzheimer's disease, with his recent MMSE score of 19/30, indicating progressive memory loss. What are the recommended diagnostic or monitoring strategies to assess his disease progression?

Answer: For a 70-year-old male diagnosed with mild-to-moderate Alzheimer's disease and a recent MMSE score of 19/30, it is important to use a combination of clinical assessments and biological markers to monitor disease progression. Based on the provided context, the following diagnostic and monitoring strategies are recommended: 1. **Clinical Assessments:** - Continue regular cognitive assessments using tools like the Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog13), and Clinical Dementia Rating-Sum of Boxes (CDR-SB) to monitor changes in cognitive function. 2. **Biological Markers:** - **CSF Biomarkers**: Monitor cerebrospinal fluid (CSF) levels of amyloid-beta (Aβ42) and tau proteins (total tau and phosphorylated tau), as changes in these biomarkers are indicative of Alzheimer's disease progression. - **PET Imaging**: Utilize amyloid PET scans to detect amyloid-beta deposition in the brain and tau PET scans to assess tau pathology. These imaging techniques can provide insights into the extent of amyloid and tau pathology over time. 3. **MRI Imaging:** - Perform regular MRI scans to monitor brain atrophy, particularly in the hippocampal region. MRI can help assess structural changes and neurodegeneration associated with disease progression. 4. **Additional Considerations:** - Evaluate metabolic and inflammatory markers if there are indications of metabolic dysfunction or inflammation, as these factors can influence disease progression. - Consider other potential co-pathologies, such as TDP-43 proteinopathy or Lewy body pathology, which might require different management strategies. These strategies can help provide a comprehensive understanding of the disease progression and guide treatment and management decisions. Regular follow-ups with a healthcare provider specializing in Alzheimer's disease are essential to adjust the monitoring plan as needed.

Fig. 7: AlzheimerRAG response: Patient 1 - Early Diagnosis and Monitoring



How can caregivers support cognitive function in this person with early-stage Alzheimer's disease without medication?

Question: How can caregivers support cognitive function in this person with early-stage Alzheimer's disease without medication?

Answer: Caregivers can support cognitive function in a person with early-stage Alzheimer's disease through various non-pharmacological strategies. Here are some approaches: 1. **Cognitive Stimulation**: Engage the person in activities that challenge their thinking, memory, and problem-solving skills. This could include puzzles, memory games, or learning new skills that are appropriate for their cognitive level. 2. **Routine and Structure**: Establish a daily routine to provide stability and reduce confusion. Consistent schedules can help the person feel more secure and oriented. 3. **Physical Activity**: Encourage regular physical exercise, as it has been shown to benefit cognitive function and overall health. Simple activities like walking, dancing, or chair exercises can be effective. 4. **Social Engagement**: Facilitate social interactions to prevent isolation and depression. This can be through family gatherings, community groups, or even virtual meet-ups with friends and relatives. 5. **Healthy Diet**: Promote a balanced diet rich in fruits, vegetables, whole grains, and lean proteins. Nutrients like omega-3 fatty acids and antioxidants are particularly beneficial for brain health. 6. **Mental Health Support**: Manage stress and anxiety through relaxation techniques, such as meditation, deep breathing exercises, or yoga. Ensuring adequate sleep is also crucial for cognitive function. 7. **Environment Modifications**: Adapt the living environment to minimize confusion and support independence. Using labels, signs, and memory aids can help the person navigate their surroundings more easily. 8. **Reminiscence Therapy**: Encourage conversations about past experiences and memories. This can help maintain a sense of identity and personal history. These strategies can help maintain cognitive function and improve the quality of life for individuals with early-stage Alzheimer's disease. It's important for caregivers to tailor these approaches to the individual's preferences and abilities.

Fig. 8: AlzheimerRAG response: Patient 1 - Non-pharmacological Interventions

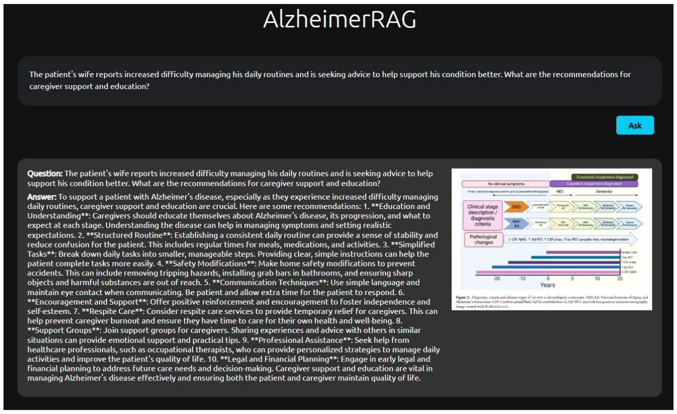
Box V.8: Patient 2 - Clinical Scenario: Behavioral Symptom Management

Clinical Scenario Query: How can behavioral symptoms like increased agitation and wandering in Alzheimer's be effectively managed without heavy reliance on medications?

Human-Generated response:
- Identify triggers for agitation and wandering by tracking behavioral patterns.
- Create a peaceful, structured environment with consistent routines
- Engage patients in soothing activities, such as pet or music therapy.
- Redirect attention when agitation occurs, employing distraction techniques rather than confrontation.
- Schedule engaging activities in the late afternoon or early evening to prevent nighttime wandering.
- Train caregivers in behavioral management techniques to ensure uniform care strategies.

VI. Conclusion

The AlzheimerRAG application represents a significant advancement in biomedical research, particularly in understanding and managing Alzheimer's disease. By integrating multimodal data—including textual information from PubMed articles, imaging studies, and clinical trial scenarios — this innovative Retrieval-Augmented Generation (RAG) tool provides a comprehensive platform for analyzing complex biomedical data. The use of cross-modal attention fusion enhances the alignment and processing of diverse data types, leading to improved accuracy in generating insights relevant to diagnosis, treatment planning, and understanding the pathophysiology of Alzheimer's disease. The experimental results indicate that AlzheimerRAG outperforms existing methodologies in terms of accuracy and robustness, demonstrating the value of a

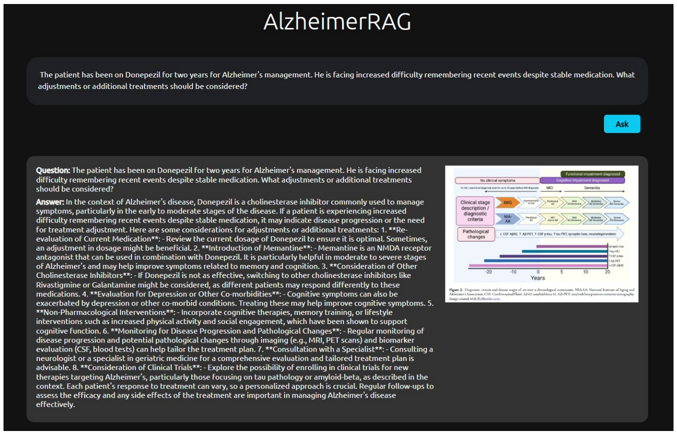


The patient's wife reports increased difficulty managing his daily routines and is seeking advice to help support his condition better. What are the recommendations for caregiver support and education?

Question: The patient's wife reports increased difficulty managing his daily routines and is seeking advice to help support his condition better. What are the recommendations for caregiver support and education?

Answer: To support a patient with Alzheimer's disease, especially as they experience increased difficulty managing daily routines, caregiver support and education are crucial. Here are some recommendations: 1. **Education and Understanding**: Caregivers should educate themselves about Alzheimer's disease, its progression, and what to expect at each stage. Understanding the disease can help in managing symptoms and setting realistic expectations. 2. **Structured Routine**: Establishing a consistent daily routine can provide a sense of stability and reduce confusion for the patient. This includes regular times for meals, medications, and activities. 3. **Simplified Tasks**: Break down daily tasks into smaller, manageable steps. Providing clear, simple instructions can help the patient complete tasks more easily. 4. **Safety Modifications**: Make home safety modifications to prevent accidents. This can include removing tripping hazards, installing grab bars in bathrooms, and ensuring sharp objects and harmful substances are out of reach. 5. **Communication Techniques**: Use simple language and maintain eye contact when communicating. Be patient and allow extra time for the patient to respond. 6. **Encouragement and Support**: Offer positive reinforcement and encouragement to foster independence and self-esteem. 7. **Respite Care**: Consider respite care services to provide temporary relief for caregivers. This can help prevent caregiver burnout and ensure they have time to care for their own health and well-being. 8. **Support Groups**: Join support groups for caregivers. Sharing experiences and advice with others in similar situations can provide emotional support and practical tips. 9. **Professional Assistance**: Seek help from healthcare professionals, such as occupational therapists, who can provide personalized strategies to manage daily activities and improve the patient's quality of life. 10. **Legal and Financial Planning**: Engage in early legal and financial planning to address future care needs and decision-making. Caregiver support and education are vital in managing Alzheimer's disease effectively and ensuring both the patient and caregiver maintain quality of life.

Fig. 9: AlzheimerRAG response: Patient 1 - Caregiver Support and Education

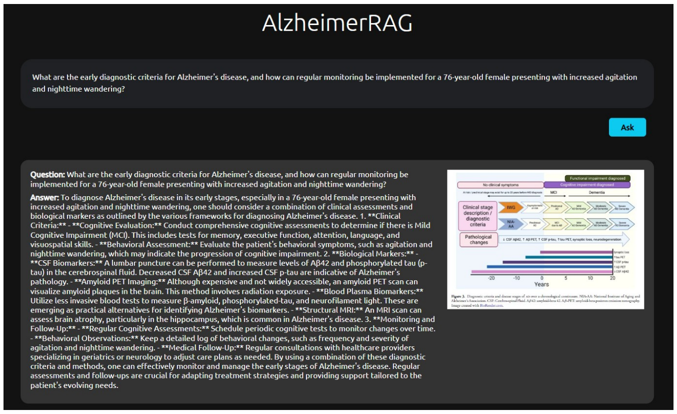


The patient has been on Donepezil for two years for Alzheimer's management. He is facing increased difficulty remembering recent events despite stable medication. What adjustments or additional treatments should be considered?

Question: The patient has been on Donepezil for two years for Alzheimer's management. He is facing increased difficulty remembering recent events despite stable medication. What adjustments or additional treatments should be considered?

Answer: In the context of Alzheimer's disease, Donepezil is a cholinesterase inhibitor commonly used to manage symptoms, particularly in the early to moderate stages of the disease. If a patient is experiencing increased difficulty remembering recent events despite stable medication, it may indicate disease progression or the need for treatment adjustment. Here are some considerations for adjustments or additional treatments: 1. **Re-evaluation of Current Medication**: - Review the current dosage of Donepezil to ensure it is optimal. Sometimes, an adjustment in dosage might be beneficial. 2. **Introduction of Memantine**: - Memantine is an NMDA receptor antagonist that can be used in combination with Donepezil. It is particularly helpful in moderate to severe stages of Alzheimer's and may help improve symptoms related to memory and cognition. 3. **Consideration of Other Cholinesterase Inhibitors**: - If Donepezil is not as effective, switching to other cholinesterase inhibitors like Rivastigmine or Galantamine might be considered, as different patients may respond differently to these medications. 4. **Evaluation for Depression or Other Co-morbidities**: - Cognitive symptoms can also be exacerbated by depression or other co-morbid conditions. Treating these may help improve cognitive symptoms. 5. **Non-Pharmacological Interventions**: - Incorporate cognitive therapies, memory training, or lifestyle interventions such as increased physical activity and social engagement, which have been shown to support cognitive function. 6. **Monitoring for Disease Progression and Pathological Changes**: - Regular monitoring of disease progression and potential pathological changes through imaging (e.g., MRI, PET scans) and biomarker evaluation (CSF, blood tests) can help tailor the treatment plan. 7. **Consultation with a Specialist**: - Consulting a neurologist or a specialist in geriatric medicine for a comprehensive evaluation and tailored treatment plan is advisable. 8. **Consideration of Clinical Trials**: - Explore the possibility of enrolling in clinical trials for new therapies targeting Alzheimer's, particularly those focusing on tau pathology or amyloid-beta, as described in the context. Each patient's response to treatment can vary, so a personalized approach is crucial. Regular follow-ups to assess the efficacy and any side effects of the treatment are important in managing Alzheimer's disease effectively.

Fig. 10: AlzheimerRAG response: Patient 1 - Medication Management

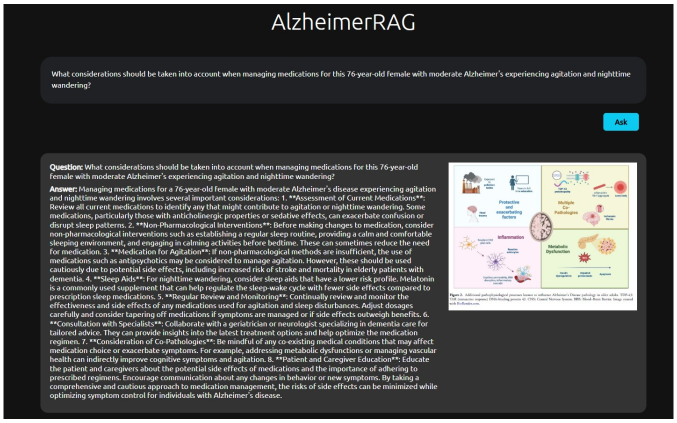


What are the early diagnostic criteria for Alzheimer's disease, and how can regular monitoring be implemented for a 76-year-old female presenting with increased agitation and nighttime wandering?

Question: What are the early diagnostic criteria for Alzheimer's disease, and how can regular monitoring be implemented for a 76-year-old female presenting with increased agitation and nighttime wandering?

Answer: To diagnose Alzheimer's disease in its early stages, especially in a 76-year-old female presenting with increased agitation and nighttime wandering, one should consider a combination of clinical assessments and biological markers as outlined by the various frameworks for diagnosing Alzheimer's disease. 1. **Clinical Criteria:** - **Cognitive Evaluation:** Conduct comprehensive cognitive assessments to determine if there is Mild Cognitive Impairment (MCI). This includes tests for memory, executive function, attention, language, and visuospatial skills. - **Behavioral Assessment:** Evaluate the patient's behavioral symptoms, such as agitation and nighttime wandering, which may indicate the progression of cognitive impairment. 2. **Biological Markers:** - **CSF Biomarkers:** A lumbar puncture can be performed to measure levels of Aβ42 and phosphorylated tau (p-tau) in the cerebrospinal fluid. Decreased CSF Aβ42 and increased CSF p-tau are indicative of Alzheimer's pathology. - **Amyloid PET Imaging:** Although expensive and not widely accessible, an amyloid PET scan can visualize amyloid plaques in the brain. This method involves radiation exposure. - **Blood Plasma Biomarkers:** Utilize less invasive blood tests to measure β-amyloid, phosphorylated-tau, and neurofilament light. These are emerging as practical alternatives for identifying Alzheimer's biomarkers. - **Structural MRI:** An MRI scan can assess brain atrophy, particularly in the hippocampus, which is common in Alzheimer's disease. 3. **Monitoring and Follow-Up:** - **Regular Cognitive Assessments:** Schedule periodic cognitive tests to monitor changes over time. - **Behavioral Observations:** Keep a detailed log of behavioral changes, such as frequency and severity of agitation and nighttime wandering. - **Medical Follow-Up:** Regular consultations with healthcare providers specializing in geriatrics or neurology to adjust care plans as needed. By using a combination of these diagnostic criteria and methods, one can effectively monitor and manage the early stages of Alzheimer's disease. Regular assessments and follow-ups are crucial for adapting treatment strategies and providing support tailored to the patient's evolving needs.

Fig. 11: AlzheimerRAG response: Patient 2 - Early Diagnosis and Monitoring

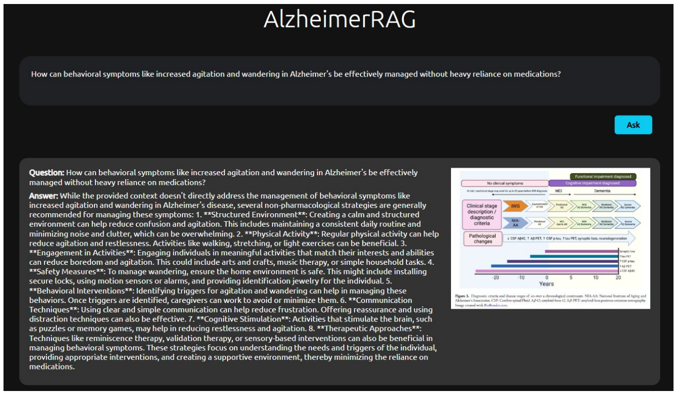


What considerations should be taken into account when managing medications for this 76-year-old female with moderate Alzheimer's experiencing agitation and nighttime wandering?

Question: What considerations should be taken into account when managing medications for this 76-year-old female with moderate Alzheimer's experiencing agitation and nighttime wandering?

Answer: Managing medications for a 76-year-old female with moderate Alzheimer's disease experiencing agitation and nighttime wandering involves several important considerations: 1. **Assessment of Current Medications**: Review all current medications to identify any that might contribute to agitation or nighttime wandering. Some medications, particularly those with anticholinergic properties or sedative effects, can exacerbate confusion or disrupt sleep patterns. 2. **Non-Pharmacological Interventions**: Before making changes to medication, consider non-pharmacological interventions such as establishing a regular sleep routine, providing a calm and comfortable sleeping environment, and engaging in calming activities before bedtime. These can sometimes reduce the need for medication. 3. **Medication for Agitation**: If non-pharmacological methods are insufficient, the use of medications such as antipsychotics may be considered to manage agitation. However, these should be used cautiously due to potential side effects, including increased risk of stroke and mortality in elderly patients with dementia. 4. **Sleep Aids**: For nighttime wandering, consider sleep aids that have a lower risk profile. Melatonin is a commonly used supplement that can help regulate the sleep-wake cycle with fewer side effects compared to prescription sleep medications. 5. **Regular Review and Monitoring**: Continually review and monitor the effectiveness and side effects of any medications used for agitation and sleep disturbances. Adjust dosages carefully and consider tapering off medications if symptoms are managed or if side effects outweigh benefits. 6. **Consultation with Specialists**: Collaborate with a geriatrician or neurologist specializing in dementia care for tailored advice. They can provide insights into the latest treatment options and help optimize the medication regimen. 7. **Consideration of Co-Pathologies**: Be mindful of any co-existing medical conditions that may affect medication choice or exacerbate symptoms. For example, addressing metabolic dysfunctions or managing vascular health can indirectly improve cognitive symptoms and agitation. 8. **Patient and Caregiver Education**: Educate the patient and caregivers about the potential side effects of medications and the importance of adhering to prescribed regimens. Encourage communication about any changes in behavior or new symptoms. By taking a comprehensive and cautious approach to medication management, the risks of side effects can be minimized while optimizing symptom control for individuals with Alzheimer's disease.

Fig. 12: AlzheimerRAG response: Patient 2 - Medication Management



How can behavioral symptoms like increased agitation and wandering in Alzheimer's be effectively managed without heavy reliance on medications?

Question: How can behavioral symptoms like increased agitation and wandering in Alzheimer's be effectively managed without heavy reliance on medications?

Answer: While the provided context doesn't directly address the management of behavioral symptoms like increased agitation and wandering in Alzheimer's disease, several non-pharmacological strategies are generally recommended for managing these symptoms: 1. **Structured Environment**: Creating a calm and structured environment can help reduce confusion and agitation. This includes maintaining a consistent daily routine and minimizing noise and clutter, which can be overwhelming. 2. **Physical Activity**: Regular physical activity can help reduce agitation and restlessness. Activities like walking, stretching, or light exercises can be beneficial. 3. **Engagement in Activities**: Engaging individuals in meaningful activities that match their interests and abilities can reduce boredom and agitation. This could include arts and crafts, music therapy, or simple household tasks. 4. **Safety Measures**: To manage wandering, ensure the home environment is safe. This might include installing secure locks, using motion sensors or alarms, and providing identification jewelry for the individual. 5. **Behavioral Interventions**: Identifying triggers for agitation and wandering can help in managing these behaviors. Once triggers are identified, caregivers can work to avoid or minimize them. 6. **Communication Techniques**: Using clear and simple communication can help reduce frustration. Offering reassurance and using distraction techniques can also be effective. 7. **Cognitive Stimulation**: Activities that stimulate the brain, such as puzzles or memory games, may help in reducing restlessness and agitation. 8. **Therapeutic Approaches**: Techniques like reminiscence therapy, validation therapy, or sensory-based interventions can also be beneficial in managing behavioral symptoms. These strategies focus on understanding the needs and triggers of the individual, providing appropriate interventions, and creating a supportive environment, thereby minimizing the reliance on medications.

Fig. 13: AlzheimerRAG response: Patient 2 - Behavioral Symptom Management

multimodal approach in addressing the complexities inherent in Alzheimer's research. While it exhibits low hallucination rates, the risks of generating misleading information in nuanced clinical scenarios remain, necessitating further research and clinical validation for real-world safety and applicability. Future enhancements could expand its scope to other neurodegenerative disorders, such as Parkinson's, incorporate additional data sources (e.g., wearable devices, electronic health records (EHR)), refine the user interface for improved interpretability, and optimize clinical trial support through enhanced patient recruitment and monitoring. Continuous improvements informed by user feedback will further enhance its utility and functionality for researchers and clinicians. In summary, while the AlzheimerRAG shows great promise in enhancing Alzheimer's research, pursuing these outlined future directions will be essential for maximizing its impact in clinical settings.

Appendix I
Ethical Consideration Statement

AlzheimerRAG prioritizes ethical integrity by using only publicly available PubMed data, avoiding private patient information, and adhering to ethical standards without requiring institutional review board approval. While rigorous filtering and cross-modal attention mitigate biases from historical PubMed imbalances, users must interpret results cautiously due to potential underrepresentation. Outputs are traceable to sources, though cross-modal complexity limits full transparency. Despite low hallucination rates (6%), clinical oversight remains critical to validate recommendations and address outdated or conflicting data.

The system aims to accelerate Alzheimer's research and aid underserved regions, yet risks of bias perpetuation or misinterpretation necessitate clear disclaimers and user education. Future commitments include continuous updates with the latest data, partnerships with clinicians and ethicists, and enhancements to align with real-world needs, ensuring the responsible integration of biomedical workflows while balancing societal benefits with ethical safeguards.

Acknowledgment

We would like to thank the researchers from the Vector Institute in Toronto, Canada, for their invaluable contributions, including expert guidance on the clinical study and participation in curating the human-generated responses for the tool evaluation. This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) research grant RGPIN/6686-2019.

References

[1] Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F. and Wang, Y.: Multimodal token fusion for vision transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 12186–12195 (2022)

[2] Zhou, F. and Chen, H.: Cross-modal translation and alignment for survival analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 21485–21494 (2023)

[3] Liang, P.P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L., Wu, P., Lee, M.A., Zhu, Y., and others: Multibench: Multiscale benchmarks for multimodal representation learning. Advances in neural information processing systems 2021(DB1), 1 (2021)

[4] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4), 1234–1240 (2020)

[5] Khattab, O. and Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over Bert. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 39–48 (2020)

[6] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research 21(140), 1–67 (2020)

[7] Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[8] Beltagy, I., Lo, K., and Cohan, A.: SciBERT: A pre-trained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019)

[9] Madan, S., Lentzen, M., Brandt, J., Rueckert, D., Hofmann-Apitius, M. and Fröhlich, H.: Transformer models in biomedicine. BMC Medical Informatics and Decision Making 24(1), 214 (2024)

[10] Ke, Y., Jin, L., Elangovan, K., Abdullah, H.R., Liu, N., Sia, A.T.H., Soh, C.R., Tung, J.Y.M., Ong, J.C.L., and Ting, D.S.W.: Development and Testing of Retrieval Augmented Generation in Large Language Models–A Case Study Report. arXiv preprint arXiv:2402.01733 (2024)

[11] Jiang, X., Hu, Z., Wang, S. and Zhang, Y.: Deep learning for medical image-based cancer diagnosis. Cancers 15(14), 3608 (2023)

[12] Robertson, S., Zaragoza, H., and others: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval 3(4), 333–389 (2009)

[13] Zhang, X., Zhang, W., Sun, W., Sun, X. and Jha, S.K.: A Robust 3-D Medical Watermarking Based on Wavelet Transform for Data Protection. Computer Systems Science & Engineering 41(3) (2022)

[14] Campillo-Sánchez, P. and Gómez-Sanz, J.: Modelling and Simulation of Alzheimer's disease Scenarios. Procedia Computer Science 83, 353–360 (2016)

[15] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., and others: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

[16] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H. and Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems 36 (2024)

[17] Zhu, Y., Zhu, M., Liu, N., Xu, Z., and Peng, Y.: Llava-phi: Efficient multi-modal assistant with small language model. In: Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited, 18–22 (2024)

[18] LangChain: Question Answering. Docs, 2021. https://python.langchain.com/v0.1/docs/use_cases/question_answering/

[19] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.M., Chen, W., and others: Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence 5(3), 220–235 (2023)

[20] LLamaIndex: Documentation. 2022. https://www.llamaindex.ai/

[21] Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W. and Lu, X.: Pubmedqa: A dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146 (2019)

[22] HuggingFace: Llama-2-7b-pubmed. 2022. https://huggingface.co/botch/Llama-2-7b-pubmed

[23] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L.: Qlora: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems 36 (2024)

[24] Meta AI: Faiss. 2022. https://ai.meta.com/tools/faiss/

[25] Huang, L.K., Chao, S.P. and Hu, C.J.: Clinical trials of new drugs for Alzheimer's disease. Journal of Biomedical Science 27, 1–13 (2020)

[26] Chen, W., Hu, H., Chen, X., Verga, P., and Cohen, W.W.: Murag: Multimodal retrieval-augmented generator for open question answering over images and text. arXiv preprint arXiv:2210.02928 (2022)

[27] Lahiri, A.K., Hasan, E., Hu, Q.V. and Ding, C.: TMU at TREC Clinical Trials Track 2023. arXiv preprint arXiv:2403.12088 (2024)

[28] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küt tler, H., Lewis, M., Yih, W., Rocktäschel, T. and others: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33, 9459–9474 (2020)

[29] Hinton, G.: Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531 (2015)

[30] Chen, J., Lin, H., Han, X., and Sun, L.: Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 17754–17762 (2024)

[31] Gupta, N., Zhang, P., Kannan, R. and Prasanna, V.: PaCKD: Pattern-Clustered Knowledge Distillation for Compressing Memory Access Prediction Models. In: 2023 IEEE High-Performance Extreme Computing Conference (HPEC), 1–7 (2023)

[32] Fang, Z., Zhu, S., Chen, Y., Zou, B., Jia, F., Qiu, L., Liu, C., Huang, Y., Feng, X., Qin, F., and others: GFE-Mamba: Mamba-based AD Multi-modal Progression Assessment via Generative Feature Extraction from MCI. arXiv preprint arXiv:2407.15719 (2024)

[33] Yao, Z., Wang, H., Yan, W., Wang, Z., Zhang, W., Wang, Z. and Zhang, G.: Artificial intelligence-based diagnosis of Alzheimer's disease with brain MRI images. European Journal of Radiology 165, 110934 (2023)

[34] Zhao, R., Chen, H., Wang, W., Jiao, F., Do, X.L., Qin, C., Ding, B., Guo, X., Li, M., Li, X. and others: Retrieving multimodal information for an augmented generation: A survey. arXiv preprint arXiv:2303.10868 (2023)

[35] Liu, X., Chen, K., Wu, T., Weidman, D., Lure, F. and Li, J.: Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. Translational Research 194, 56–67 (2018)

[36] Li, J., Li, X., Chen, F., Li, W., Chen, J. and Zhang, B.: Studying the Alzheimer's disease continuum using EEG and fMRI in single-modality and multi-modality settings. Reviews in the Neurosciences (2024)

[37] National Library of Medicine: PubMed. 1996. https://pubmed.ncbi.nlm.nih.gov/

[38] Marino, K., Rastegari, M., Farhadi, A. and Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, 3195–3204 (2019)

[39] Data Discovery: NIH. 2008. https://datadiscovery.nlm.nih.gov/

[40] Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D. and others: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics 16, 1–28 (2015)

[41] Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., Zhang, L., Zou, J. and Yao, H.: Mmed-rag: Versatile multimodal rag system for medical vision language models. arXiv preprint arXiv:2410.13085 (2024)

[42] Amugongo, L.M., Mascheroni, P., Brooks, S.G., Doering, S. and Seidel, J.: Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review. arXiv (2024)

[43] Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, C.R., Jagust, W.J., Shaw, L.M., Toga, A.W. and others: Alzheimer's disease Neuroimaging Initiative (ADNI) clinical characterization. Neurology 74(3), 201–209 (2010)

[44] Atri, A.: The Alzheimer's disease clinical spectrum: diagnosis and management. Medical Clinics 103(2), 263–293 (2019)

[45] Murray, M.E., Graff-Radford, N.R., Ross, O.A., Petersen, R.C., Duara, R. and Dickson, D.W.: Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. The Lancet Neurology 10(9), 785–796 (2011)

[46] Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chét elat, G., Teunissen, C.E., Cummings, J. and van der Flier, W.M.: Alzheimer's disease. The Lancet 397(10284), 1577–1590 (2021)

[47] Weller, J. and Budson, A.: Current understanding of Alzheimer's disease diagnosis and treatment. F1000Research 7 (2018)

[48] Lane, C.A., Hardy, J. and Schott, J.M.: Alzheimer's disease. European Journal of Neurology 25(1), 59–70 (2018)

[49] Twarowski, B. and Herbet, M.: Inflammatory processes in Alzheimer's disease—pathomechanism, diagnosis, and treatment: a review. International journal of molecular sciences 24(7), 6518 (2023)

[50] Rostagno, A.A.: Pathogenesis of Alzheimer's disease. International Journal of Molecular Sciences 24(1), 107 (2022)

[51] Knapskog, A.B., Engedal, K., Selbæk, G. and Øksengård, A.R.: Alzheimers sykdom–diagnostikk og behandling. Tidsskrift for Den norske legeforening (2021)

[52] Mantzavinos, V. and Alexiou, A.: Biomarkers for Alzheimer's disease diagnosis. Current Alzheimer Research 14(11), 1149–1154 (2017)

[53] Eratne, D., Loi, S.M., Farrand, S., Kelso, W., Velakoulis, D. and Looi, J.C.L.: Alzheimer's disease: clinical update on epidemiology, pathophysiology, and diagnosis. Australasian psychiatry 26(4), 347–357 (2018)

[54] Ogbodo, J.O., Agbo, C.P., Njoku, U.O., Ogugofor, M.O., Egba, S.I., Ihim, S.A., Echezona, A.C., Brendan, K.C., Upaganlawar, A.B. and Upasani, C.D.: Alzheimer's disease: pathogenesis and therapeutic interventions. Current aging science 15(1), 2–25 (2022)

[55] Oboudiyat, C., Glazer, H., Seifan, A., Greer, C. and Isaacson, R.S.: Alzheimer's disease. In: Seminars in neurology, 313–329 (2013)

[56] Thomo, A.: PubMed Retrieval with RAG Techniques. In: Digital Health and Informatics Innovations for Sustainable Health Care Systems, 652–653 (2024)

[57] Xiong, G., Jin, Q., Lu, Z. and Zhang, A.: Benchmarking retrieval-augmented generation for medicine. arXiv preprint arXiv:2402.13178 (2024)

[58] Wang, C., Long, Q., Xiao, M., Cai, X., Wu, C., Meng, Z., Wang, X. and Zhou, Y.: BioRAG: A RAG-LLM Framework for Biological Question Reasoning. arXiv preprint arXiv:2408.01107 (2024)

[59] Aisen, P.S., Cummings, J., Jack, C.R., Morris, J.C., Sperling, R., Frölich, L., Jones, R.W., Dowsett, S.A., Matthews, B.R., Raskin, J. and others: On the path to 2025: understanding the Alzheimer's disease continuum. Alzheimer's research & therapy 9, 1–10 (2017)

[60] Mangialasche, F., Solomon, A., Winblad, B., Mecocci, P. and Kivipelto, M.: Alzheimer's disease: clinical trials and drug development. The Lancet Neurology 9(7), 702–716 (2010)

[61] Yang, R., Tan, T.F., Lu, W., Thirunavukarasu, A.J., Ting, D.S.W. and Liu, N.: Large language models in health care: Development, applications, and challenges. Health Care Science 2(4), 255–263 (2023)

[62] Cohen, J.: Statistical power analysis for the behavioral sciences. (2013)

[63] Pearson, K.: X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50(302), 157–175 (1900)

[64] Yang, R., Marrese-Taylor, E., Ke, Y., Cheng, L., Chen, Q. and Li, I.: Integrating umls knowledge into large language models for medical question answering. arXiv e-prints arXiv:2310 (2023)

[65] Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., Daneshjou, R., Frankle, J., Liang, P., Carbin, M., and others: Biomedlm: A 2.7 b parameter language model trained on biomedical text. arXiv preprint arXiv:2403.18421 (2024)

[66] Treder, M.S., Lee, S., and Tsvetanov, K.A.: Introduction to Large Language Models (LLMs) for dementia care and research. Frontiers in Dementia 3, 1385303 (2024)

[67] Meta: Introducing Llama: A foundational, 65-billion-parameter language model. (2023) https://ai.meta.com/blog/large-language-model-llama-meta-ai/

[68] Amazon: Amazon Mechanical Turk. 2005. https://www.mturk.com/

[69] Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X.E., Adi, Y., Liu, J., Sauvestre, R., Remez, T. and others: Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950 (2023)

[70] Topsakal, O. and Akinci, T.C.: Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In: International Conference on Applied Engineering and Natural Sciences, 1050–1056 (2023)

[71] Braunschweiler, N., Doddipatla, R., Keizer, S. and Stoyanchev, S.: Evaluating Large Language Models for Document-grounded Response Generation in Information-Seeking Dialogues. arXiv preprint arXiv:2309.11838 (2023)

[72] Sticha, A.: Utilizing large language models for question answering in task-oriented dialogues. (2023)

[73] Gu, Y., Tinn, R. and Cheng, H.: PubMedBERT-2: Advanced Biomedical Language Understanding. arXiv preprint arXiv:2402.15785 (2024)

[74] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML) (2021)

[75] Xia, P., Zhu, K., Li, H., Zhu, H., Li, Y., Li, G., Zhang, L. and Yao, H.: RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models. arXiv preprint arXiv:2407.05131 (2024)

[76] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. CRC Press, 1994.

[77] Student, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.

[78] J. Neyman and E. S. Pearson, "IX. On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694–706, pp. 289–337, 1933.

[79] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "SuperGLUE: A stickier benchmark for general-purpose language understanding systems," arXiv preprint arXiv:1905.00537, 2020. [Online]. Available: https://arxiv.org/abs/1905.00537

[80] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. ICLR*, 2019.

[81] A. K. Lahiri and Q. V. Hu, "Descriptor: Open-Domain Long-Form Context-Aware Question-Answering Dataset (DragonVerseQA)," in IEEE Data Descriptions, vol. 2, pp. 141-150, 2025, doi: 10.1109/IEEEDATA.2025.3562173

[82] Lahiri, A. K., & Hu, Q. V. (2024, December). HouseOfTheDragonQA: Open-Domain Long-form Context-Aware QA Pairs for TV Series. In 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (pp. 150-157). IEEE.

[83] Lahiri, Aritra Kumar, and Qinmin Vivian Hu. "Gameofthronesqa: Answer-aware question-answer pairs for TV series." In European Conference on Information Retrieval, pp. 180-189. Cham: Springer International Publishing, 2022.

[84] Y. Seonwoo, J.-H. Kim, J.-W. Ha, and A. Oh, "Context-aware answer extraction in question answering," *arXiv preprint arXiv:2011.02687*, 2020.